Next generation DNA sequencing based strategies; towards a new era for the traceability of endangered species and genetically modified organisms

Alfred Joseph Arulandhu

Next generation DNA sequencing based strategies; towards a new era for the traceability of endangered species and genetically modified organisms

Alfred Joseph Arulandhu

Thesis committee

Promotor
Prof. Dr Saskia M. van Ruth
Special professor Food Authenticity and Integrity
Wageningen University & Research

Co-promotor
Dr Esther J. Kok
Head Unit Novel Foods, programme leader Product Composition
RIKILT Wageningen University & Research

Other members
Prof. Dr Marcel H. Zwietering, Wageningen University & Research
Prof. Dr Ate D. Kloosterman, Netherlands Forensic Institute, Den Haag/University of Amsterdam
Dr Frédéric Debode, CRA-W, Gembloux, Belgium
Dr Bert Pöpping, FOCOS GbR, Alzenau, Germany

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences)

Next generation DNA sequencing based strategies; towards a new era for the traceability of endangered species and genetically modified organisms

Alfred Joseph Arulandhu

Thesis

submitted in fulfilment of the requirements for the degree of doctor

at Wageningen University

by the authority of the Rector Magnificus,

Prof. Dr A.P.J. Mol,

in the presence of the

Thesis Committee appointed by the Academic Board

to be defended in public

on Wednesday 19th December 2018

at 1:30 p.m. in the Aula.

Alfred Joseph Arulandhu

Next generation DNA sequencing based strategies; towards a new era for the traceability of endangered species and genetically modified organisms, 179 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2018)
With references, with summary in English

ISBN: 978-94-6343-368-6
DOI: https://doi.org/10.18174/463307

In memory of my mother Anthony Rajathi

Table of Contents

Chapter 1

General Introduction

1.1 Introduction

Food products are for the larger part composed of multiple ingredients that may be heavily processed. As a result, it is generally difficult, or even impossible, to determine the ingredient composition by visual inspection. At the same time, consumers demand detailed information on the composition of their food products before purchasing [1,2]. In the European Union (EU) food products and their ingredients are subjected to law. The basic law in this respect is the General Food Law that states that the labelling of food should not mislead consumers (European Commission (EC) regulation, 178/2002). More detailed legislation on labelling states that: 'labelling should allow consumers to make informed choices and to make safe use of food, while at the same time ensure the free movement of legally produced and marketed foods' (EC regulation, 1169/2011). Food authenticity studies have proven mislabelling for a wide range of food products, such as, meat, milk, honey, rice, edible oils, and spices [3-7]. Mislabelling might be caused by ignorance or lack of information regarding regulations, however, in many cases fraudulent intentions have shown to be the cause of mislabelling [3]. Fraud with foods and food ingredients raises concerns about food quality and safety, potentially poses health risks and, in addition, it may raise ethical issues when food products contain illegal ingredients.

An example of food fraud with illegal ingredients is the use of endangered species as an ingredient. Some endangered species may be considered to have medicinal properties, and the addition of such species, or parts thereof, increases the value of a product [8]. Multiple studies have shown that endangered species are used in a variety of products, including food products [8-10]. Globally an increasing number of species are overexploited, which poses the threat of extinction of certain species, i.e. elephants, rhino etc., but also many plant species. Around 35,000 species, belonging to various plant and animal taxa, are classified as endangered by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) www.cites.org. Worldwide, the legal trade in endangered species is regulated by a permit system linked to the CITES convention (1973). In the EU, enforcement of the CITES convention is based on EC regulation, 338/97 on the Protection of species of wild fauna and flora which are threatened by trade. This regulation defines necessary procedures and documents required for import, export and re-export (permits) for the species listed under CITES law. Besides the regulated legal trade in endangered species, a significant portion of the trade in endangered plant and animals is illegal (www.traffic.org), this involves billions of dollars per year [11,12]. Although international trade agreements are being implemented, illegal trading and the use of endangered species parts are still common practice, as can be seen from reports from, amongst others, European and other customs authorities [8].

In other food fraud categories, food ingredients may also relate to food safety regulations, such is the case for genetically modified organisms (GMOs). The use of GMOs in food products has been the subject of public debates, since the first GMO entered the world market in 1996 [13]. In the last two decades, the development and production of a wide range of different GM plants have been observed [14]. GM plants and their derived products have been commercialised for the use in food/feed in many countries [15]. The regulations for the use of genetically modified (GM) crops in food and feed products vary between countries. The consequence of different regulations may be asynchronous approval. For example, 195 GM varieties are allowed in the USA, while in the European Union (EU) 55 GM varieties (without stacks) are authorised for use in food or feed products in 2017 (http://www.isaaa.org/gmapprovaldatabase/). Within the EU the introduction of new GMOs in food and feed supply chains is regulated by, primarily, two regulations, EC regulation, 1829/2003 and EC, 1830/2003. EC regulation, 1829/2003 on genetically modified food and feed states that approved GMOs need to be assessed for their food/feed and environmental safety prior to market introduction, and that unauthorised GMOs (UGMOs) are not allowed on the European market. It is furthermore stipulated

that labelling of GMOs is mandatory in the EU for products that contain more than 0.9% of authorised GMO per ingredient. EC regulation, 1830/2003 is focused on the traceability and labelling requirement for GMOs entering EU market. In recent years, the number of incidents related to the presence of unauthorised GMOs (UGMOs) and UGMO-derived products (food and feed) have globally increased and this has also affected the EU markets [15].

1.2 Current analytical methods to address authenticity

In order to assess the authenticity of food/feed products entering the EU market, both traditional and modern analytical methods are used. Frequently applied analytical techniques to determine general authenticity aspects of food products are thin layer chromatography, gas chromatography, capillary electrophoresis, high-performance liquid chromatography and mass spectrometry-based techniques [16,17]. In the last three decades, molecular techniques have become more widely applied in routine set-ups to determine the composition of food products [18,19]. Currently, for the identification of endangered species polymerase chain reaction (PCR) with DNA barcode markers is performed and followed by Sanger sequencing, which is the most commonly applied molecular biological procedure to identify a single product [20,21]. For GMOs, crop-specific and GMO-specific (element, construct and event) TaqMan PCRs are usually performed to determine the presence of a GMO in a sample [22]. Unexplained targets in GMO screening may indicate the presence of UGMO. Generally for identifying the unknown region of a UGMO a genome walking approach, where known GM targets are used to 'read' into the unknown region, is applied in combination with Sanger sequencing [23].

In recent years the field of molecular biology is increasingly using Next Generation Sequencing (NGS) technologies to address issues in the identification of species, strains, varieties etc [24,25]. NGS allows for massive parallel sequencing of targeted DNA, potentially enabling an overview of the genetic composition of a product in a single analysis. The application of NGS-based approaches for the detection and identification of endangered species as well as of GMOs and UGMOs is currently, however, still limited. Nonetheless, the potential of such NGS approaches based on an initial selective amplification step is large for these two different areas of application related to food authenticity. In order to apply powerful NGS-based approaches in the field of food authenticity, it is necessary to gain more knowledge and insight in the available methodologies for initial amplification of sequences and the advantages and disadvantages of the different NGS methods, before being able to develop dedicated strategies for different food authenticity issues.

1.2.1 Endangered species

In the EU, enforcement of the CITES convention is mainly focused at the borders, where imported products suspected of containing endangered plants and/or animals will be seized by Customs and CITES authorities [26]. In cases where the morphological characteristics of the species are still present and can be used for visual identification (microscopy etc.), for instance, the coloured feathers of a bird, or the flowers of a plant, it is not very difficult to determine the species in a seized sample. However, identification becomes considerably more difficult when a product contains only parts of an animal or plant and the morphological characteristics are lost. The most difficult category is products with processed plant or animal parts that are pulverized and have become an ingredient of, for instance, food supplements or traditional medicines (TMs) [8,27]. In those cases, visual identification will no longer be feasible and DNA-based methods will be the method of choice to detect and identify species that may be present in the sample. DNA barcoding methods are often applied to identify species based on well-conserved yet variable DNA regions in the genome, so called barcode markers [28]. A DNA full-length

barcode marker is a short gene sequence, >600 bp in length, derived from a standardised part of the genome that can be used to identify the species in a particular sample. In addition, mini-barcode markers are shorter versions of the full-length barcoding marker, <400 bp, which are developed to identified species in processed samples, were the DNA is largely degraded. Many parts of the DNA across species are conserved and in specific cases these conserved regions are flanking species-specific regions in the DNA [28,29]. This allows the use of universal markers, located in two conserved flanking regions to amplify a species specific region in between, which can be used to identify the species. The use of universal markers makes it feasible to develop a universal barcode primer set to analyse and identify several species in a single PCR analysis.

For animal species identification, mitochondrial DNA (mtDNA) is an ideal choice to distinguish between the species. The copy number of the mtDNA is higher compared to the nuclear DNA (factor ~1000) per cell and mtDNA evolves in a rapid way, hence more often allowing for distinguishing close related species [30]. Predominantly, the mitochondrial Cytochrome *c* oxidase subunit 1 (*CO1*), 16S and Cytochrome b (cyt *b*) loci have been used as the DNA barcode locus for animal taxonomy [31-33]. Furthermore, several animal mini-barcode primer sets (CO1, 16S and cyt *b*) have been applied and efficient discrimination has been observed in several samples [34].

For terrestrial plants, the *CO1* gene and other mitochondrial regions are a poor choice for species identification, because of low substitution rates and intra-molecular recombination of mitochondrial DNA in plants [35]. The two "core" DNA barcode markers used to discriminate most plant species are a combination of RuBisCO large subunit (*rbcL)* and maturaseK (*matK)* [36]. The *rbcL* is easy to amplify in PCR, however, has less discriminatory power compared to *matK* [35]; vice versa, *matK* is a rapidly evolving coding sequence with a high discrimination power between plants, but with poor PCR amplification, specifically in non-angiosperms [35]. In a previous study it was suggested that the combination of *rbcL+matK* provides a better species discrimination compared to any other 2-marker or multi-marker plastid barcode markers [37]. Some studies indicate that the plastid intergenic spacer *trnH-psbA* region and the nuclear ribosomal ITS region could have more discrimination power across the land plants [36]. Designing universal mini-barcode primer sets for plants has been proven to be difficult, especially when DNA is degraded and only shorter DNA fragments are available that give less resolution to distinguish and identification is possible only at family level [38].

Currently, for species identification using a barcoding approach, a PCR amplification of the specific target is performed with the barcode marker and related primer sets. Subsequently, the obtained amplicon is Sanger sequenced to identify the sequence information. The obtained DNA barcoding sequence can be compared with the database to identify the species based on the nucleotides' variations observed between the DNA barcode sequences. The National Centre for Biotechnology Information (NCBI) and Barcode of Life Data System (BOLD) databases are collective barcode storage databases where sequence information can be derived for species identification and characterization. Nevertheless, the use of a DNA barcoding approach on complex samples can lead to several setbacks, such as, the use of universal primer sets will amplify all the specific targets in the sample, and Sanger sequencing cannot be applied to sequence products containing more than one species, unless additional time-consuming work (cloning of PCR products) is performed [18]. Furthermore, as mentioned before, the ingredients in the case of, for instance, TMs or food-supplements can be heavily processed. As a result, the DNA can be severely damaged and degraded [8].

1.2.2 Genetically modified organisms (GMOs)

In recent decades, the cultivation and production of an increasing range of different GMOs has been observed. GM plants and derived products have been commercialised in many countries in the last two decades [15]. GM plants that have not received market approval are not allowed on the European market. In the last few decades many methods for the detection, identification and quantification of individual GMOs have been developed [39]. Applicants that aim to market a new GM plant variety are required to provide a GMO-specific method as well as the related reference materials. These methods will be assessed and, if the basic requirements are met, further in an international validation study the EURL (the European Reference Laboratory) in collaboration with labs from the European Network of GMO Laboratories (ENGL). Usually a two-step GMO screening approach is performed to identify the presence of GMOs in a sample. Initially, the specific TaqMan PCR assays for GMO-related targets (endogenous, elements and constructs) are performed. The positive detection of these targets may indicate the presence of one or more GMOs. Secondly, based on the specific GMO elements as detected in the first step of the screening, GMO event-specific TaqMan PCRs are performed to verify the presence of authorised GMOs or known unauthorised GMOs for which event-specific methods are available [19,22]. When identified elements, or combinations of elements cannot be explained by the presence of an authorised GMO in the same sample, this may indicate the presence of an unauthorised GMO (UGMO) [22].

UGMOs are GMOs that have not yet been assessed for their food, feed and environmental safety in the country where these are marketed. In some countries, regulations for the low-level presence (LLP) of UGMOs in food or feed products have been established if the UGMO has already been approved in another country (EC regulation, 1829/2003). In the EU, only a limited number of UGMOs meet the specific set of requirements and these LLP varieties are allowed to be present in feed products up to a level of 0.1 % per ingredient (mass-based) (EC regulation, 1829/2003). The main bottleneck in relation to the identification of UGMOs is the generally limited available sequence information as a basis for the development of adequate methods for analysis [15]. Identification of UGMOs is only possible through the identification of the DNA sequence bridging the GMO construct and related endogenous plant DNA, as integration of the insert into the host genome has so far been a random process [23]. At the same time, most UGMOs will contain known GM elements that can serve as a starting point to 'read' into unknown regions.

1.3 Limitations of currently available detection methods for endangered species, GMO and UGMO identification

In recent years, a gradual shift can be observed in the field of molecular biology, where increasingly NGS strategies have been developed to use in protocols aimed at the identification of species, breeds, varieties, strains, etc. This development can be observed in the field of genomics but increasingly also in, for instance, the medical area and other areas, such as agricultural and environmental sciences [40-42]. Often these NGS strategies complement or replace more traditional PCR or Sanger sequencing based methods for identification [43]. Recently, studies have applied DNA metabarcoding (combining DNA barcoding and NGS) to identify species in complex food samples, such as, traditional medicines (TMs), or herbal supplements etc. In some studies the presence of endangered species (*Ursus thibetanus*, *Panax ginsenga*, *etc*) in this type of products has been confirmed. These endangered species were either declared on the ingredient list or found to be undeclared ingredients [9,18,44]. In these studies either barcoding or mini-barcoding have been used to identify either animal or plant species present in a simple to complex mixture. However, in the scientific literature so far no methodology has been

described combining informative plant and animal barcoding and mini-barcoding in a single analytical NGS-based strategy [23].

With respect to GMO detection and identification, current approaches focus on sequence specific amplification (TaqMan PCR) of GMO-related sequences and subsequent identification of specific GMOs present in the sample using event-specific methods. Due to the increased numbers of GMOs and its related targets multiple TaqMan PCR assays are necessary to have an informative screening. The addition of a new GM element to the screening requires the development and validation of a new GMO-specific method, making the screening less flexible and a time and budget -consuming process. For UGMO detection, a number of enrichment approaches (LAM PCR, LT-RADE, SiteFinding-PCR, A-T linker and LF PCR) have been developed to identify the unknown adjacent sequence of known GM elements. Using Sanger sequencing, sequence information of the unknown sequence can be obtained and this sequence will be analysed to verify whether it belongs to an EU authorised GMO or a UGMO [22,40,45,46]. However, none of these strategies have so far shown to be adapted to the specific demands of GMO detection, i.e. a 0.1% detection limit and enrichment of UGMO targets in a background of GMOs. Furthermore, Sanger sequencing is not compatible with sequencing multiple DNA amplicons in a single analysis [18].

1.4 Research aim

The objective of this thesis was to use detailed genetic differences to identify species/varieties in feed/food products based on advanced analytical NGS based strategies. The study focused on the identification of two target groups: (a) endangered species and (b) GMOs. Elucidating genetic composition was sub-divided into three main topics: 1) the development of efficient enrichment strategies, 2) selection of the optimal NGS strategy for the purpose and 3) actual identification of the species/GMOs of interest. With respect to endangered species, the aim was to explore whether NGS-based strategies allow the simultaneous identification of all species, including endangered species, present in a sample, even in complex samples that may be heavily processed (*chapter 2, 3 and 4*). With respect to GMOs, the aim was to reliable identify all GMOs and UGMOs present in a given sample, regardless of their relative abundance, based on enrichment of known or, in the case of UGMOs, additional adjacent unknown sequences (*chapter 5, 6 and 7*).

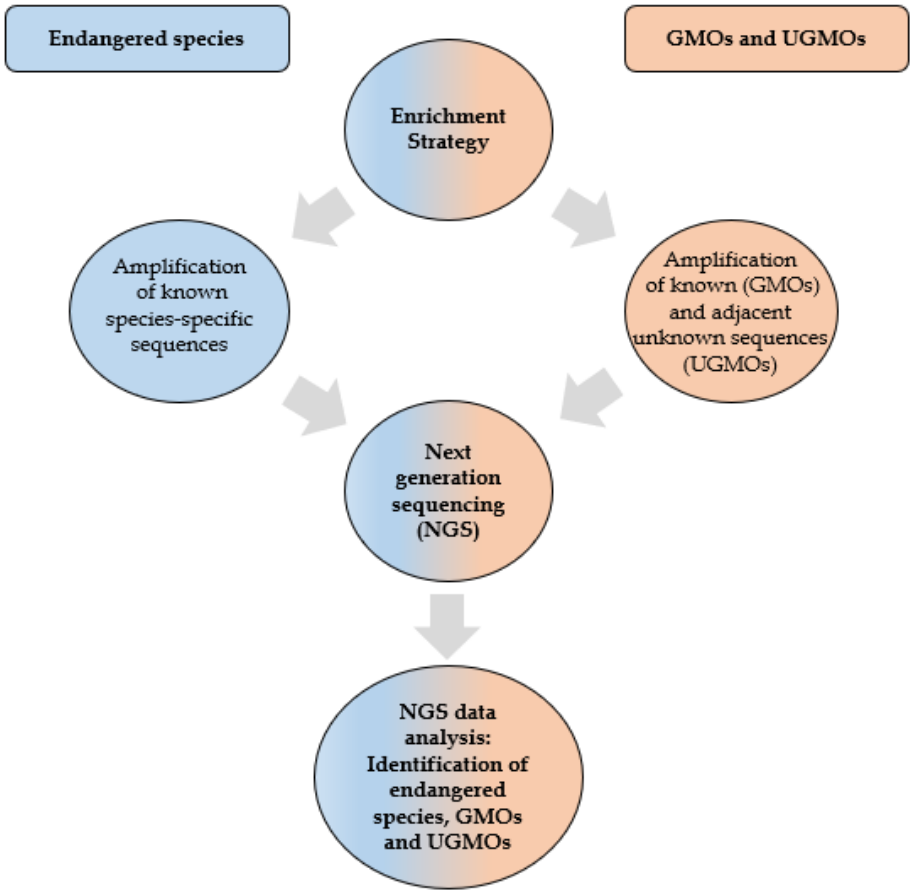


Figure 1.1: Schematic overview of the development of adequate enrichment strategies for 1) endangered species identification 2) GMOs and UGMOs detection and identification using NGS based approaches.

1.5 Outline of this thesis

Chapter 2 reviews existing literature to select the best available universal DNA full-length barcode and mini-barcode markers for both plants and animals to enable enrichment of species specific sequences and subsequent identification of any species, with an emphasis on the detection and identification of endangered species. More specifically, this chapter addresses the current challenges in obtaining good quality DNA from wild life forensic samples, gives an overview of the available plant and animal barcode and mini-barcode markers and discusses available NGS technologies and their suitability for a DNA metabarcoding approach for the screening of wild life forensic samples.

Chapter 3 addresses the development and validation of a multi-locus DNA metabarcoding approach for identification of endangered species in complex samples. The inventory study of the available markers (*chapter 2*) was used to select 12 markers (barcode and mini-barcode primer sets) for a metabarcoding platform and a single optimal PCR condition for amplification of these markers was defined. The efficiency of the multi-locus DNA metabarcoding approach was evaluated on the basis of 15 well-defined complex mixtures, including materials of endangered species. The repeatability and reproducibility of the approach was evaluated with a validation study across 16 laboratories using 10 samples, including two wild life forensic samples. The main goal of this chapter was to evaluate a multi-

locus DNA metabarcoding strategy for species identification in complex mixtures and ensure high resolution and quality, even in heavily processed samples.

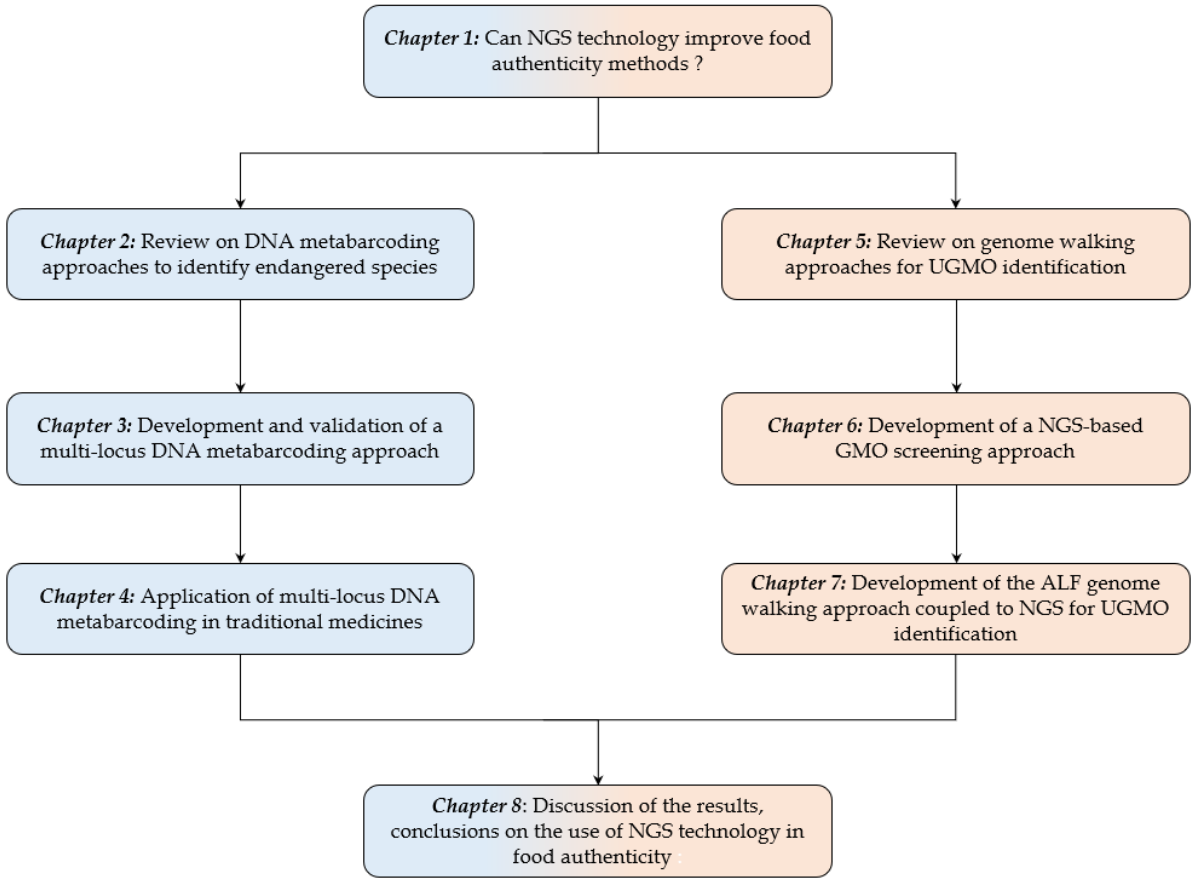


Figure 1.2 Coherence of the different chapters in this thesis.

Chapter 4 assesses the applicability of the multi-locus DNA metabarcoding described in *chapter 3* on real-life samples containing heavily degraded DNA and consisting of different matrices. Furthermore, the chapter addresses food authenticity aspects of TMs. An optimal DNA isolation method for highly processed DNA from TMs was identified by comparing 8 different DNA isolation methods. Here, 18 TMs, with varying matrices, were analysed and the identified species were compared with the respective ingredient lists. It was assessed whether the metabarcoding strategy could be applied effectively in the different types of TMs, and, secondly, whether the TMs might actually contain any endangered species.

Chapter 5 reviews the available enrichment strategies that have been described in the scientific literature to identify unknown adjacent sequences to known DNA elements, with a focus on the development of more effective methods for UGMO identification. The advantages and disadvantages of the most promising enrichment strategies were evaluated and necessary adjustments of current methods to comply with the requirements for UGMO detection were determined. The aim of this chapter was to obtain an overview of relevant aspects of available genome walking (GW) methods that can be used to identify UGMOs.

In chapter 6, an NGS-based broad GMO screening approach was developed, and the applicability of the developed approach was evaluated by comparing its results with the results of the standard qPCR screening approach on the same samples. Five feed products, known to contain multiple GMOs in different quantities, were used to perform the comparison, a data analysing pipeline was developed to process the NGS data. The main goal of this chapter was to develop a broad NGS-based GMO screening approach along with a data analysis pipeline for efficient GMO identification, and to determine the practical efficiency of such a NGS screening strategy.

In chapter 7, the findings of *chapter 5* were applied to develop a detection method that can fulfil the requirements for UGMO identification. A new GW strategy for UGMO identification was developed by combining the advantageous aspects of available GW strategies that were discussed in *chapter 6*. The efficiency of the developed enrichment strategy was evaluated based on available, well-characterised reference materials for EU-approved GMOs with known sequences that were used to compose complex samples with multiple GMOs present in different percentages.

In chapter 8, the findings of this thesis are integrated. The impact of the research described in this thesis, together with the implications, limitations and recommendations for further research are discussed and the final conclusions of the thesis are presented.

Chapter 2

Advances in DNA metabarcoding for food and wildlife forensic species identification

This chapter was published as: Staats M, **Arulandhu AJ**, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E. "Advances in DNA metabarcoding for food and wildlife forensic species identification". *Analytical and Bioanalytical Chemistry* 2016; 408(17): 4615-4630.

Abstract

Species identification using DNA barcodes has been widely adopted by forensic scientists as an effective molecular tool for tracking adulterations in food and for analysing samples from alleged wildlife crime incidents. DNA barcoding is an approach that involves sequencing of short DNA sequences from standardized regions and comparison to a reference database as a molecular diagnostic tool in species identification. In recent years, remarkable progress has been made towards developing DNA metabarcoding strategies, which involves Next-Generation Sequencing of DNA barcodes for the simultaneous detection of multiple species in complex samples. Metabarcoding strategies can be used in processed materials containing highly degraded DNA e.g. for the identification of endangered and hazardous species in traditional medicine. This review aims to provide insight into advances of plant and animal DNA barcoding and highlights current practices and recent developments for DNA metabarcoding of food and wildlife forensic samples from a practical point of view. Special emphasis is placed on new developments for identifying species listed in the Convention on International Trade of Endangered Species (CITES) appendices for which reliable methods for species identification may signal and/or prevent illegal trade. Current technological developments and challenges of DNA metabarcoding for forensic scientists will be assessed in the light of stakeholders' needs.

Keywords Endangered species, Next-generation sequencing, Wildlife forensic samples, COI, CITES

2.1 Introduction

Genetic identification of species plays a key role in the investigation of illegal trade of protected or endangered wildlife [47] and in the detection of species mislabelling and fraud in the food industry [48]. Currently, DNA barcoding is an established molecular technique that is used for differentiating and assigning taxonomy to species using standardized short DNA sequences (Box 2.1). Application of DNA barcoding for food authentication has gained much attention because of food safety concerns, including incorrect food labelling, food substitutions or food contamination [49-51]. DNA barcoding has been effective in the traceability of many processed food products in particular seafood and meat products [48]. For instance, DNA barcoding has made impact by demonstrating widespread mislabelling or substitution of fish and seafood products in markets and restaurants in New York (USA) and Canada [50,51]. Proper identification of species present in food and food supplements is of vital importance to protect consumers against potential food adulteration, ingredient mislabelling or food poisoning. Given its utility, DNA barcoding is being used by the US Food and Drug Administration as a replacement for the time-consuming technique of protein isoelectric focusing for fish and fish products [52].

Another established application of DNA barcoding to forensic science is in investigations of wildlife crimes such as illegal collection and trade of flora and fauna. More than 35,000 species of flora and fauna are categorized as endangered by the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES). Protected species are listed in appendices I, II and III, according to how severe a certain population is threatened to extinction [53]**.** Besides the regulated legal trade, a significant portion of the trade in endangered flora and fauna is illegal. In the European Union (EU), enforcement is mainly focused at the borders, where illegally imported wildlife products, plants or animals will be seized by Customs and CITES Authorities. The mailings on the EU-TWIX (European Union Trade in Wildlife Information eXchange; www.eutwix.org) network of wildlife-protecting enforcement bodies are very clear; seizures of wildlife and products containing wildlife are everyday practice. In some cases, the species identification of seized specimens is not very difficult, because the specific morphological characteristics can be readily observed, though often requiring taxonomic expertise for decisive identification. Identification will be more difficult when only parts of an animal or plant without distinctive morphological characteristics are present, or when plant or animal parts have been pulverized and have become ingredients of food supplements (e.g. Traditional Medicines (TMs)). Currently, CITES list species encompasses a wide diversity of species of terrestrial plants such as cycads, cacti and orchids, in addition to vertebrates such as fish, amphibians, reptiles, birds, and mammals, and invertebrates such as lobsters, crabs, and corals [54]. Customs laboratories will obviously benefit from applying standardized, fast and reliable methods when dealing with samples of which no *a priori* knowledge on the species composition is known. It is these benefits that have made DNA barcoding the method of choice for customs laboratories when trying to establish the presence of biological material from endangered species within processed products [20,26,55,56]**.**

A complicating aspect for DNA barcoding in the analysis of food supplements such as TMs and other mixed products is that they are composed of more than one ingredient. Such samples often contain multiple species that can only be efficiently analysed if multiple DNA barcode templates can be sequenced in parallel; something that Next-Generation Sequencing (NGS) technologies do effectively [42]. Current NGS platforms yield millions of DNA reads in a relatively short period of time, and the sequencers' performance improves every year [57]. NGS combined with DNA barcoding is referred to

as metabarcoding [58]. Metabarcoding thus uses universal PCR primers to mass-amplify one or more taxonomically informative targets. The prefix 'meta' refers to the collection of barcode sequences from different species. The general strategy consists of (1) extracting DNA from food or (wildlife) forensic samples (2) amplifying a specific DNA barcode or other target region of taxonomic value, (3) sequencing the corresponding DNA amplicons using NGS technology, (4) analysing the sequences using appropriate bioinformatics pipelines, (5) identifying the species in the sample from which DNA has been extracted and (6) screening for CITES species among these [59,60]. Metabarcoding has been applied in many diverse environmental samples, such as faeces [58], soil [61], marine water [62] and bulk samples of tropical arthropods [63]. However, there are only a few published applications of metabarcoding to food and (wildlife) forensic samples. Coghlan *et al*. [8,44] demonstrated the power of metabarcoding in detecting species in complex Traditional Chinese Medicines (TCMs) samples presented in the form of powders, crystals, capsules, tablets, and herbal tea. Their screening revealed that some of the TCM samples contained CITES-listed species, including the Asiatic black bear (*Ursus thibetanus*) and the Saiga antelope (*Saiga tatarica*), as well as unlisted ingredients, and potentially toxic and allergenic plants. Cheng *et al*. [64] performed metabarcoding analyses on well-defined TCM preparations based on a six-herb formula named Liuwei Dihuang Wan, which is widely used in China. They concluded that there are significant differences in quality and safety among commercial TCM preparations, as the unlisted species *Senna obtusifolia* was identified in some preparations that may potentially pose safety risks to consumers. Tillmar *et al*. [65] developed a metabarcoding method for the identification of species of mammals in human forensic tissues, with which the presence of low quantities of DNA from the genus *Canis* could be identified.

Although metabarcoding may seem easy to apply, researchers often face limitations in obtaining a representative assessment of species composition. First, different pre-processing conditions and production procedures from samples with different composition and matrices (e.g. TMs and other processed and complex products) may result in highly variable DNA quality and concentration. DNA integrity has shown to have significant influence on the effectiveness of the metabarcoding and other molecular methodologies for species identification [34,64,66]. Secondly, while there are many bioinformatics methods available for the analysis of metabarcoding data, the discriminating power of these methods is directly related to prior choices on barcode marker and reference database composition [67,68]. PCR bias caused by variable primer-template mismatches across species may limit the quantitative potential of DNA metabarcoding, and may cause species to be missed [69,70]**.** Furthermore, DNA metabarcoding wholly relies on the presence of high-quality barcode sequence reference databases that are based on good taxonomy and barcode coverage. The goal of this paper is to review the advancements and current practices of plant and animal metabarcoding, with an emphasis on complex food and forensic wildlife samples for identifying, in particular, species listed by CITES. This effort is complementary to recent work focussing on metabarcoding for biodiversity assessments in environmental samples [25,60,71,72] and an extension of the work on DNA barcoding of food and forensic samples [47-49,73-76]. This overview will address the opportunities and challenges that must be faced to allow the customs laboratories and other routine laboratories to perform efficient and reliable metabarcoding analysis that can broadly identify any species present in a sample under investigation.

Box 2.1 DNA barcoding and the International Barcode of Life project (iBOL).

DNA barcoding is a rapid method of differentiating and assigning taxonomy to species using standardized short DNA sequences. For animals, the most commonly used sequence is a 658-base pair region of the mitochondrial cytochrome *c* oxidase subunit I gene (COI, COX1, CO1). DNA barcoding allows for fast, reliable, automatable, and cost-effective species identification by users with little or no taxonomic experience [77]. Identifications are usually made by comparing unknown sequences against known species DNA barcodes via alignment searching (BLAST) [78] or distance-based tree construction [79].

A suitable barcode for identification at the species level should be sufficiently variable between species (typically ≥ 3% difference between closely related species but this may vary amongst taxonomic groups) and display either low or no intraspecific variations. Also, barcodes should be widely studied for a large number of species to enable comparison of the nucleotide sequence from an unknown sample with reference sequences in a database. Accurate species identification wholly relies on the taxonomic coverage of barcodes in a reference database. If the query sequence lacks a conspecific (belonging to the same species) target sequence in the database, species-level barcoding-based identification of the query will fail. Instead, the closest matches in the database may be identified and the sample barcode scored as a "new" taxon (operational taxonomic unit; OTU). From a practical point of view, therefore, DNA barcoding requires a comprehensive reference database. Such reference data sets are being assembled by the barcoding campaigns initiated by the International Barcode Of Life project (iBOL; www.ibol.org), resulting in considerably improved species coverage for target taxa of such DNA barcoding campaigns (Kwong *et al*. 2012). Official barcode sequences generated by the iBOL initiatives are deposited and organized in the Barcode Of Life Data (BOLD) Systems (http://boldsystems.org; [80]). BOLD is a large-scale and rigorously curated DNA barcode storage database and most of the sequences information contained within BOLD have all been derived from voucher specimens with authoritative taxonomic identifications. Barcoding campaigns focussing on fish, birds, mammals, insects and fungi have been initiated e.g. the "Fish Barcode of Life" Initiative (FISH-BOL, www.fishbol.org), the "Marine Barcode Of Life" Initiative (MarBOL, www.marinebarcoding.org), the "Shark Barcode Of Life" project (SharkBOL; www.sharkbol.org), and the "Barcode of Wildlife Project" (BWP; www.barcodeofwildlife.org). For plants there are initiatives to barcode e.g. the world's tree species Barcoding Of Life (TreeBOL), and grasses and grass-like plants Barcoding Of Life (GrassBOL).

Barcodes and a variety of alternative taxonomically informative genes that have been generated from general scientific research are deposited in the International Nucleotide Sequence Database Collaboration (INSDC), can be used for taxonomic assignment in barcoding studies. The iBOL initiative aims to create a database of 5 million standardized DNA sequences, which can be used to identify 500,000 species, by 2015.

Scientific literature on the utility of DNA barcoding in the recognition, discrimination, and discovery of plant and animal species has been reviewed extensively by Savolainen *et al*. [81], Kress and Erickson [82], Bucklin *et al*. [83], Hollingsworth *et al*. [67], Fazekas *et al*.[35], Ortea *et al*. [73], Nicolè *et al*. [74], Bhargava and Sharma [84], Kvist [85] and Sandionigi *et al*. [86].

2.2 DNA extraction and DNA integrity

The initial sample preparation and extraction step in the analysis of DNA from food products is probably the most crucial step in the process of species identification in complex forensic samples. This step can be very difficult to standardize and optimize due to the complexity and diversity of the matrices encountered, each presenting different problems. For instance, it can be difficult to ensure that a representative sample is obtained from heterogeneous samples that are composed of many ingredients (e.g. TMs), and in such cases sufficient homogenization is particular critical prior to DNA extraction. Forensic samples, such as food samples and traditional medicine may contain only very low amounts of DNA, or contain ingredients that have been subjected to various treatments during the production process (e.g. cooking, high pressure, pH modification, grinding or drying), which may cause the DNA to be highly degraded [64,87-89]. Furthermore, failure to eliminate potential inhibitory components and interfering substances from the material under investigation (e.g. protein, lipids, polyphenols, polysaccharides) may severely influence PCR analysis. Needless to say that any factor that may contribute to downstream bias needs to be minimized.

Different DNA extraction methods, which can be used for analysis of forensic samples are now available; extraction is either based on in-house developed protocols or commercially available kits. Commercial kits offer a means for standardizing DNA extraction from forensic samples, as the protocol can be easily implemented in any laboratory. However, in many laboratories user-specific protocols have been developed to improve DNA extraction efficiency on a case-by-case basis. DNA extraction using cetyltrimethylammonium bromide (CTAB) extraction buffer combined with additional silica or resin-based purification step have found to be efficient for a wide range of plants and plant-derived products, in particular for separation of polysaccharides from DNA [66,89,90]. Ivanova *et al*. [91] developed a cost-efficient and automation-friendly DNA extraction protocol for animal tissues that consists of a tissue lysis step (SDS and proteinase K) followed by silica-based purification of DNA using inexpensive glass fibre filtration plates. The latter method has been used to process thousands of animal species at the Canadian Centre for DNA Barcoding (CCDB) as part of the iBOL initiative. Despite these efforts in standardizing the DNA extraction method, the most suitable method is generally found to be strongly dependent on the matrix, and there is no "universal" method that could be used for all food and (wildlife) forensic samples [87].

As suspect samples may often contain degraded DNA, it is a requirement that metabarcoding methods are able to identify species on the basis of short DNA sequences that may still be present in highly processed materials [89]. In such forensic samples, DNA degradation often prevents the amplification of PCR fragments longer than ~300 base pair (bp) [34,66,92,93]. The use of shorter barcode regions, so-called mini-barcodes, may overcome this problem. Due to their reduced size, mini-barcodes are often amplified with higher efficiency in degraded samples than standard, full-length barcodes, which are typically between 650 – 900 bp in length [38,66]. On the other hand, the rate of taxonomic discrimination is generally positively correlated with the length of the mini-barcode. The use of universal mini-barcodes that will only allow identification of taxa above the species-level, due to saturation of the taxonomic discrimination, should generally be avoided unless identification at the genus or family-level is warranted.

2.3 Animal DNA barcodes and mini-barcodes

For animals, the standard barcode is a 658 bp region in the gene encoding mitochondrial cytochrome *c* oxidase I (COI or COX1, CO1) [94]. COI has long been used in animal molecular systematics to study relationships of closely related species because of its high level of interspecific variation [95]. Its popularity within the barcoding community is clearly reflected in the large public databases such as National Centre for Biotechnology Information (NCBI) GenBank (www.ncbi.nlm.nih.gov/genbank) and BOLD (Box 2.1). Universal primer sets for amplifying the COI barcode across major taxonomic groups have been developed by Ivanova *et al*. [31] and primer cocktails have been reported that are effective in fish, mammals, amphibians and reptiles (Table 2.1). A good discriminatory power in the identification of birds (98 to 100% identification success rate [79]), fish (93% to 98% identification success rate [96]), spiders (100% identification success rate [97]), butterflies (97.9% identification success rate [98]), and reptiles (72.7% to 100% identification success rate [99]) has been shown for the COI barcode.

Despite its proven effectiveness, COI is not always suitable and effective in identifying all animal species. For endangered organisms such as sea snails (the mollusc class Gastropoda) and corals the COI barcoding region and other mitochondrial markers were found to offer insufficient resolution to allow for reliable discrimination between closely related species [100-103]. Using a DNA metabarcoding approach, Elbrecht *et al*. [70] demonstrated that species may go undetected in complex artificial mixtures of freshwater invertebrate taxa because of universal COI primer-template mismatches. The use of group-specific primers or alternative degenerate primers may prevent species from being missed using COI [69,70].

The traceability of mammalian meat including meat of ranched and hunted game species heavily relies on the use of the mitochondrial cytochrome *b* (cyt*b*) region [48,104]. The choice of cyt*b* instead of COI is due mainly to practical reasons. The early availability of universal primers for cyt*b* [60-61], long before the use of COI became popular, lead to the deposition of several thousand cyt*b* sequences of a large range of edible mammalian species in public databases. Thus, its use became well established. Nonetheless, DNA barcoding based on COI has also proven effective in the identification of edible meat, including bush meat species [48,104-107]. The FishTrace consortium (www.fishtrace.org) have promoted the use of cyt*b* through the development of universal cyt*b* primers for teleost fish species and the release of validated sequence data of many hundreds of European marine fish species [108].

Additional activities have taken place in finding suitable short DNA regions and related PCR primers for barcoding of species in widely diverse food and forensic samples, but so far no truly mini-barcode standard has been adopted. Efforts in designing short broad coverage COI barcodes (i.e. mini-barcodes) to accommodate identification of a diversity of animal species in samples with degraded DNA has proven to be difficult. The use of the 130-bp COI mini-barcode primers designed by Meusnier *et al*. [34] has been limited [109], because the priming sites in the COI gene to accommodate the mini-barcode design have shown not to be conserved enough to cover a broad range of taxa [68,110]. Leray *et al*. [67] have taken a thorough approach and used the COI barcodes provided by the Moorea BIOCODE project, an "All Taxa Biotic Inventory" (www.mooreabiocode.org), consisting of > 64,000 sequences across all phyla to design conserved universal COI mini-barcoding primers to target a 313 bp region. The newly designed primers have been reported to perform well across metazoan diversity, with a higher success rate than the versatile primer sets traditionally used for DNA barcoding, i.e. the "Folmer primers" HCO2198 and LCO1490 [111] (Table 2.1).

Mitochondrial cyt*b*, 12S and 16S rRNA genes are the most commonly used genetic markers for species discrimination in degraded samples [112]. Universal primers for the amplification of short regions of cyt*b* have been developed for various animal taxa [113,114]. Their use has been demonstrated in different problematic forensic samples that may contain degraded DNA including hair shafts, bones, feathers, and meat products [104,114].

Mini-barcodes based on the 12S and 16S rDNA mitochondrial genes have recently been demonstrated by several studies to be suited for identifying a wide range of animal species in environmental samples [115] and processed food and wildlife forensic products including TMs [8,33,65,116]. The 12S and 16S rDNA contain internal regions that are strongly conserved across taxa, suitable for designing universal primers, alternated with short hyper-variable regions that are species-specific. Sarri *et al*. [33] developed an approximately 250 bp barcode marker (Table 2.1), which allowed for the successful amplification of the 16S region across different sample types (e.g. cheese, processed meats, frozen fish fillets) and the correct identification of a wide range of animals in food products, including fishes, birds, reptiles, crustaceans and European mammals. Kitano *et al*. [117] developed 12S and 16S mini-barcodes for the identification of a large number of vertebrates (mammals, birds, reptiles, amphibians, and fish). Similarly, Karlsson and Holmlund [112] used short 12S and 16S regions to identify a total of 28 different mammals including domestic and game species.

2.4 Plant DNA barcodes and mini-barcodes

In plants, the COI gene and other mitochondrial regions are a poor choice for species identification because the mitochondrial genome in plants has evolved too slowly to allow it to be used for DNA barcoding [118]. The research for a COI analogue in plants has focused on the plastid genome, but the selection of a standard plant barcode marker has been complicated by the trade-off that arises between the high requirements of universality and high variability among plants [82]. So far, no single barcode marker has been found that is expected to discriminate all of the > 200,000 species of plants. The Consortium for the Barcode Of Life (CBOL) plant working group has opted for the use of a core set of two (*rbc*L and *mat*K) coding sequences from plastids as the "core" DNA barcode (Table 2.2) [67]. The *rbc*L barcode consists of a 599 bp region at the 5' end of the gene. It is easy to amplify, sequence and align in most land plants, but it has only modest discriminatory power. Newmaster *et al*. [119] analysed over 10,000 *rbc*L sequences from GenBank and found that *rbc*L could discriminate samples in approximately 85% of pairwise comparisons of congeneric species. The *mat*K barcode region consists of a ca. 841-bp region at the centre of the gene, which is one of the most rapidly evolving regions of the plastid genome. The *mat*K is perhaps the closest plant analogue to the COI animal barcode [120]. Ogden *et al*. [121] developed a Single-nucleotide polymorphism (SNP) genotyping approach based on *mat*K DNA barcodes to distinguish between traded timber products of Ramin (*Gonostylus*) species, which are all CITES protected. Unfortunately, *mat*K can be difficult to amplify, particularly in non-angiosperms, due to the lack of sufficiently universal primers [35,122].

The two most widely-used supplementary loci are the nuclear ribosomal ITS (nrITS) [123] and plastid intergenic spacer *psb*A-*trn*H region [124]. The nrITS region had previously been discounted as a standard DNA barcode due to concerns over paralogy and the presence of putative pseudogenes which led to sequencing difficulties in many plant groups [125]. However, the increased resolution of nrITS over plastid DNA barcodes in many studies suggested that it should continue to be explored as part of the plant DNA barcode [123,126]. Some authors have noted that just using a subset of the ribosomal

cassette (nrITS2) can lead to greater amplification and sequencing success compared to the entire nrITS region [123]. By testing the discriminating ability of nrITS2 in more than 6,600 medicinal plants and closely related samples, Chen *et al*. [123] found that the rate of successful identification was 92.7% at the species level, and they proposed that the nrITS2 region should be the standard barcode for investigating forensic samples containing medicinal plants. Newmaster *et al*. [119] used *rbc*L and nrITS2 DNA barcodes to highlight species substitution and contamination in herbal products.

The *psb*A-*trn*H region is straightforward to amplify across land plants, and is one of the more variable intergenic spacers in plants [127]. It has been used successfully in a range of barcoding studies [128,129]. One of the main concerns associated with the use of *psb*A-*trn*H as a standard barcode is the premature termination of sequence reads by mononucleotide repeats leading to unidirectional reads in up to 30% of sequences [130].

In plants, the design of suitable universal mini-barcode markers has proven difficult. The length constraints to allow working with highly degraded DNA severely limit the taxonomic resolution of mini-barcodes compared to that of the 500-800 bp long standardized barcodes (*rbc*L, *mat*K). Primers for the amplification of a ~180-bp region of chloroplast *rbc*L have been used, but this system only allows in most cases the identification of families, not genera or species [131]. Little [38] *in silico* evaluated a variety of *rbc*L primers and found the discriminatory power of the best *rbc*L mini-barcode to be less than 38.2%. Taberlet *et al*. [132] have used the chloroplast $tRNA^{Leu}$ (UAA) intron sequences [*trn*L (UAA): 254 - 767 bp] and a shorter fragment of this intron (the P6-loop, 10 - 143 bp) for identifying plant species in processed food and ancient permafrost samples. The number of *trn*L (UAA) intron sequences available in databases is high, by far the most numerous among non-coding chloroplast DNA sequences. The *trn*L (UAA) region had overall low resolution. However, Taberlet *et al*. [132] concluded that only closely related species are not resolved and that the region can effectively be used to identify commonly eaten plants (e.g. potato, tomato, maize, but not almond). The *trn*L (UAA) has been extensively used in food industry [133], forensic sciences [20] and diet studies based on faeces [134].

Table 2.1 Non-exhaustive list of primers for amplifying animal DNA barcodes and mini-barcodes.

DNA marker	Target taxonomic group	Primer name	Primer sequences (5′ - 3′)	Amplicon length (bp)	Remark	Reference
COI	Various phyla	LCO1490 HCO2198	GGTCAACAAATCATAAAGATATTGG TAAACTTCAGGGTGACCAAAAAATCA	648		Folmer *et al*. [111]
COI	Reptiles	RepCOI-F RepCOI-R	TNTTMTCAACNAACCACAAAGA ACTTCTGGRTGKCCAAARAATCA	664		Nagy *et al*. [99]
COI-1	Birds	BirdF1 BirdR1	TTCTCCAACCACAAAGACATTGGCAC ACGTGGGAGATAATTCCAAATCCTG	648	Forward Reverse	Hebert *et al*. [79]
COI	Insects and amphibians	LepF1 LepR1 MLepF1 MLepR1	ATTCAACCAATCATAAAGATATTGG TAAACTTCTGGATGTCCAAAAAATCA GCTTTCCCACGAATAAATAATA (use with LepR1) CCTGTTCCAGCTCCATTTTC (use with LepF1)	648		Hebert *et al*. [135] Hajibabaei *et al*. [98]
COI-2	Mammals, fish reptiles and amphibians	LepF1_t1 VF1_t1 VF1d_t1 VF1i_t1 LepR1_t1 VR1d_t1 VR1_t1 VR1i_t1	TGTAAAACGACGGCCAGTATTCAACCAATCATAAAGATATTGG TGTAAAACGACGGCCAGTTCTCAACCAACCACAAAGACATTGG TGTAAAACGACGGCCAGTTCTCAACCAACCACAARGAYATYGG TGTAAAACGACGGCCAGTTCTCAACCAACCAIAAIGAIATIGG CAGGAAACAGCTATGACTAAACTTCTGGATGTCCAAAAAATCA CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCRAARAAYCA CAGGAAACAGCTATGACTAGACTTCTGGGTGGCCAAAGAATCA CAGGAAACAGCTATGACTAGACTTCTGGGTGICCIAAIAAICA	648	M13-tailed cocktail; mix ratio 1:1:1:3:1:1:1:3	Ivanova *et al*. [31]
COI-3	Fish and mammals	VF2_t1 FishF2_t1 FishR2_t1 FR1d_t1	TGTAAAACGACGGCCAGTCAACCAACCACAAAGACATTGGCAC TGTAAAACGACGGCCAGTCGACTAATCATAAAGATATCGGCAC CAGGAAACAGCTATGACACTTCAGGGTGACCGAAGAATCAGAA CAGGAAACAGCTATGACACCTCAGGGTGTCCGAARAAYCARAA		M13-tailed cocktail; mix ratio 1:1:1:1	Ivanova *et al*. [31]
COI	Sharks	FishR2 Shark-int	ACTTCAGGGTGACCGAAGAATCAGAA ATCTTTGGTGCATGAGCAGGAATAGT	550		Ward *et al*. [96]
COI	Echinodermata phylum	COIceF COIceR	ACTGCCCACGCCCTAGTAATGATATTTTTTATGGTNATGCC TCGTGTGTCTACGTCCATTCCTACTGTRAACATRTG	> 550		Hoareau and Boissin [136]
COI	Universal animal mini-barcode	mlCOIintF jgHCO2198	GGWACWGGWTGAACWGTWTAYCCYCC TAIACYTCIGGRTGICCRAARAAYCA	313		Leray *et al*. [110] Geller *et al*. [137]

COI	Universal animal mini-barcode	Uni-MinibarR1 Uni-MinibarF1	GAAAATCATAATGAAGGCATGAGC TCCACTAATCACAARGATATTGGTAC	130		Meusnier *et al*. [34]
cyt*b*	Universal mammal	L14724 H15915	CGAAGCTTGATATGAAAAACCATCGTTG AACTGCAGTCATCTCCGGTTTACAAGAC	1140	Full-length cyt*b*	Irwin *et al*. [138]
cyt*b*	Universal fish	FishcytB-F CytB1-5R	ACCACCGTTGTTATTCAACTACAAGAAC GGTCTTTGTAGGAGAAGTATGGGTGGAA	750	cyt*b*-5′ fragment	Sevilla *et al*. [108]
cyt*b*	Universal vertebrate animal mini-barcode	L14816 H15173	CCATCCAACATCTCAGCATGATGAAA CCCCTCGAATGATATTTGTCCTCA	357		Parson *et al*. [114]
cyt*b*	Universal animal mini-barcode	L14841 H15149	AAAAAGCTTCCATCCAACATCTCAGCATGATGAAA AAACTGCAGCCCCTCAGAATGATATTTGTCCTCA	307		Kocher *et al*. [113]
16S	Universal animal	16sar-L 16sbr-H	CGCCTGTTTATCAAAAACAT CCGGTCTGAACTCAGATCACGT	500 - 650	Forward Reverse	Palumbi [139]
16S	Universal animal mini-barcode	16S-forward 16S-reverse	AYAAGACGAGAAGACCC GATTGCGCTGTTATTCC	250		Sarri *et al*. [33]
16S	Fish, cephalopods and crustaceans	16S1F 16S2R	GACGAKAAGACCCTA CGCTGTTATCCCTADRGTAACT	250		Deagle *et al*. [140]
16S	Universal mammal mini-barcode	16S-forward 16S-reverse	GACGAGAAGACCCTATGGAGC TCCGAGGTCACCCCAACCTCCG	100		Tillmar *et al*. [65]
16S	Universal vertebrate mini-barcode	L2513 H2714	GCCTGTTTACCAAAAACATCAC CTCCATAGGGTCTTCTCGTCTT	244		Kitano *et al*. [117]
16S	Universal animal mini-barcode	16S-HF 16S-HR1 16S-HR2	ATAACACGAGAAGACCCT CCCACGGTCGCCCCAAC CCCGCGGTCGCCCCAAC	80 - 125		Horreo *et al*. [116]
12S	Universal vertebrate mini-barcode	L1085 H1259	CCCAAACTGGGATTAGATACCC GTTTGCTGAAGATGGCGGTA	215		Kitano *et al*. [117]
12S	Universal vertebrate mini-barcode	12SV5-F 12SV5B2	TTAGATACCCCACTATGC TAGAACAGGCTCCTCTAG	98		Riaz *et al*. [141]

Table 2.2 Non-exhaustive list of primers for amplifying plant DNA barcodes and mini-barcodes.

*rbc*L	Universal plant	rbcL a-F rbcL a-R	ATGTCACCACAAACAGAGACTAAAGC GTAAAATCAAGTCCACCRCG	654		Levin *et al*. [142] Kress and Erickson, [122]
*mat*K	Angiosperms & Gymnosperms	matK-KIM1R matK-KIM3F	ACCCAGTCCATCTGGAAATCTTGGTTC CGTACAGTACTTTTGTGTTTACGAG	656 - 889	Forward Reverse	Fazekas *et al*. [143]
*mat*K	Angiosperms & Gymnosperms	matK-390f matK-1326r	CGATCTATTCATTCAATATTTC TCTAGCACACGAAAGTCGAAGT	656 - 889	Forward Reverse	Cuenoud *et al*. [144]
*mat*K	Gymnosperms	NY552F NY1150R	CTGGATYCAAGATGCTCCTT GGTCTTTGAGAAGAACGGAGA	656 - 889	Forward Reverse	Fazekas *et al*. [143]
*mat*K	Gymnosperms	matKpkF4 matKpkR1	CCCTATTCTATTCAYCCNGA CGTATCGTGCTTTTRTGYTT	656 - 889	Forward Reverse	Fazekas *et al*. [35]
nrITS2	Universal plant	S2F ITS4	ATGCGATACTTGGTGTGAAT TCCTCCGCTTATTGATATGC		Forward Reverse	Chen *et al*. [123] White *et al*. [145]
nrITS2	Universal plant	S2F S3R	ATGCGATACTTGGTGTGAAT GACGCTTCTCCAGACTACAAT	160-320	Forward Reverse	Chen *et al*. [123]
nrITS	Universal angiosperm	17SE 26SE	ACGAATTCATGGTCCGGTGAAGTGTTCG TAGAATTCCCCGGTTCGCTCGCCGTTAC	800	Forward Reverse	Sun *et al*. [146]
*trn*H-*psb*A	Universal plant	psbAF trnH2	GTTATGCATGAACGTAATGCTC CGCGCATGGTGGATTCACAATCC	264 - 792		Sang *et al*. [147] Tate and Simpson [148]
*trn*L (UAA)	Universal plant mini-barcode	g h	GGGCAATCCTGAGCCAA CCATTGAGTCTCTGCACCTATC	10 - 143	p-loop region of *trn*L	Taberlet *et al*. [132]
*trn*L (UAA)	Universal plant	c d	CGAAATCGGTAGACGCTACG GGGGATAGAGGGACTTGAAC	767		Taberlet *et al*. [149]
*trn*L (UAA)	Universal plant mini-barcode	c h	CGAAATCGGTAGACGCTACG CCATTGAGTCTCTGCACCTATC	250		Taberlet *et al*. [149] Taberlet *et al*. [132]

For some applications, a plant mini-barcode with relatively modest discriminatory power at the genus or higher taxonomic level can be useful. For example, it is often an entire genus or family that is listed by CITES, rather than individual plant species. For many plant families listed by CITES (e.g. Cycadaceae, Orchidaceae, Cactaceae, Euphorbia) identification to a larger group is therefore all that is required. This does not apply to all illegally traded plant genera though, such as tree ferns of the genus *Cibotium*, of which only *C. regale* is legally protected. In such cases, an alternative approach could be to design species specific mini-barcodes to distinguish between closely related species from the CITES listed species, as was done for instance for *Rauvolfia serpentina* [20]

Box 2.2 Approximate number of sequences of DNA barcodes and other taxonomically informative genes available in GenBank (December 2014).

GenBank sequences were retrieved with a query of the sequence annotations using the nucleotide database, for example 'COI' OR 'cytochrome *c* oxidase' AND eukaryote'. After which the query headers were downloaded and additionally filtered using the GNU/Linux command line tools (e.g. awk and grep). The number of unique genera and species were estimated from the sequence annotations, and should be considered only as an approximation.

Number of barcoding sequences deposited in GenBank								
	COI	**16S**	**cyt*b***	***mat*K**	***rbc*L**	***trn*L**	***psb*A-*trn*H**	**nrITS**
Approx. number of accessions	940,687	264,931	324,769	94,246	134,784	172,493	44,581	378,711
Approx. number of species	102,919	60,928	34,230	43,039	47,675	63,172	20,891	84,670
Approx. number of genera	30,923	21,691	10,822	8,759	10,978	10,895	3,836	14,338

2.5 Sequencing of DNA barcodes using NGS technology

There are many excellent reviews on NGS platforms, and also their fundamentals and broad characteristics are described elsewhere [42,57,150,151]. We will focus on the important steps in the NGS workflow, and only provide a brief overview of NGS technologies relevant for DNA metabarcoding. Early DNA metabarcoding studies have employed the 454 pyrosequencing technology of Roche because it was the first commercially available NGS system and because of its longer sequence read-outs allowing for a more informative fraction of DNA barcodes to be sequenced. Pyrosequencing has been used for DNA metabarcoding of raw materials of the diet of several animals [110,134], environmental monitoring [58,152,153], as well as for analysing ancient DNA extracted from museum specimens [109]. The 454 technology is however no longer mainstream and Roche announced that 454 sequencers will be phased out in mid-2016.

Recently, benchtop sequencers have emerged that due to their compact format, lower set-up and running costs, and faster data turnaround times have made NGS accessible for routine testing laboratories. The 454 GS Junior System (Roche), the MiSeq and MiniSeq (Illumina®), the NextSeq 500 (Illumina®), the Ion Proton™ System (Ion Torrrent™) and Ion PGM™ System (Ion Torrrent™) have sequencing capacities large enough for most metabarcoding projects (Box 2.3). Tillmar *et al*. [65] used to Roche 454 GS Junior system for the detection of animal species using the 16S rRNA gene. The same benchtop sequencer and the 454 GS-Titanium sequencer were used to identify plant and animal species in TMs [8,9,64]. Bertolini *et al*. [154] used the Ion Torrent PGM™ System for the identification of DNA from meat species using 12S and 16S rRNA genes.

Box 2.3 Benchtop next-generation sequencing systems and their characteristics.

Benchtop instruments are scaled-down, economical NGS platforms driven by the need for cheaper and faster sequencing, and which are suited for metabarcoding of typical food and forensic samples. The system specificities are listed with expected maximum performance by beginning 2016.

Instrument	Company	Machine run time (h)	Reads/run	Read length (base)	Output
454 GS Junior Plus[1]	Roche	18	70,000	~700	70 Mb
MiniSeq[2]	Illumina®	24	44 – 50 million	2x150	6.6 – 7.5 Gb
MiSeq[3]	Illumina®	56	44 – 50 million	2x300	13.2 – 15 Gb
NextSeq 500[4]	Illumina®	29	Up to 800 million	2x150	100 – 120 Gb
Ion PGM™ System[5]	Ion Torrent™	7.3	4 – 5.5 million	400	1.2 – 2.0 Gb
Ion Proton™ System[6]	Ion Torrent™	4	60 – 80 million	200	Up to 10 Gb

1: Adopted from http://454.com/products/gs-junior-plus-system/index.asp. Roche announced that 454 sequencers will be phased out in mid-2016.
2: adopted from http://www.illumina.com/systems/miniseq/specifications.html
3: Adopted from http://www.illumina.com/systems/miseq/performance_specifications.html
4: Adopted from http://www.illumina.com/systems/nextseq-sequencer/performance-specifications.html
5: Adopted from https://tools.lifetechnologies.com/content/sfs/brochures/PGM-Specification-Sheet.pdf
6: Adopted from https://tools.lifetechnologies.com/content/sfs/brochures/CO06326_Proton_Spec_Sheet_FHR.pdf

The choice of NGS technology for DNA metabarcoding, may depend on several parameters such as the barcode length, the number of barcodes used and the number of samples that need to be analysed. An advantage of Illumina® sequencing is that sequencing data with very low error rates (> 0.1%) are produced, compared to 454 and Ion Torrent™ sequencing [155]. The most common error types on the 454 and Ion Torrent™ platforms are insertions and deletions (indels), in particular when reading homopolymer regions. This results in an overall error rate of ~1.5% [155,156]. Sequencing errors can lead to spurious identification of species. Bertolini *et al*. [154] reported that when Ion Torrent data are quality filtered during downstream bioinformatics processing, the error rates do not introduce any bias that could prevent the correct assignment of meat specie.

The high output combined with relatively short length have limited the use of Illumina sequencing technology mainly to profiling of bacterial communities using short 16S rDNA hypervariable regions [157,158]. However, recent developments allowed the MiSeq platform to double the amount of output per flow cell by producing read lengths of 300 bp (Box 2.3). Because the Illumina platform can generate amplicon sequences in a paired-end format, paired reads can be directly matched and assembled into amplicons of up to ~550 bp.

This development has allowed the MiSeq sequencer to compete with 454 sequencing technology as it allows for generating sequence data from barcode regions with sufficient taxonomic resolution for animal and plant species identification.

An important step in the NGS workflow is to generate a library of the amplicons of interest. Fundamental for library construction is the modification of the DNA amplicons into a form that is compatible with the NGS platform to be used. The library is constructed by enzymatically ligating adapter sequences to the DNA amplicons or by adding them by PCR. The adaptors include specific sequences that are required for clonal amplification of the library on a solid surface (bead or glass slide). The choice of these adapter sequences is dictated by the NGS platform (Box 2.3). The adapter sequences may additionally contain a 6 - 10 nucleotide-long multiplex identifier (MID) that is used to pool amplicons from several independent samples in one run. MIDs are typically added to make more efficient use of the sequencing capacity of the NGS sequencers i.e. the number of reads generated by each NGS technology is usually higher than required per sample. Adapter sequences with different MIDs need to be used for each sample when multiple samples are sequenced in a single NGS experiment. The number of samples that can be pooled depends on (1) the number of available MIDs, (2) the sequencing capacity of the NGS platform, (3) the number of amplicons per sample and (4) the required sequencing depth [159,160]. After NGS, the resulting combined sequence data from different samples are subsequently sorted *in silico* by MID using bioinformatics tools.

2.6 Bioinformatics tools

Bioinformatics has played a crucial role in the advancement of metabarcoding. In recent years, many bioinformatics tools have been developed and are constantly being improved to efficiently and effectively perform various steps involved in the metabarcoding process. After obtaining NGS data, quality filtering is the first essential step, because it removes erroneous data that may otherwise potentially lead to misidentification of species. Sequencing errors introduced during NGS can be recognized because raw reads have predicted error probabilities for each base indicated by Phred quality scores. Sequence errors can be removed during quality-filter and –trimming e.g. by truncating reads at the position where their quality begins to drop. A Phred score of 20, which corresponds to a 1% error rate in base calling, is often used as a minimum threshold in quality filtering. Bokulich *et al*. [161] have published guidelines for quality-filtering strategies to enable efficient extracting of high-quality data from Illumina amplicon sequencing data. In their studies on TMs, Coghlan *et al*. [9,44] used the commercially available software Geneious [162]. Other software tools for quality filtering of reads include e.g. PRINSEQ [163] and Trimmomatic [164].

Following quality control, the sequences can either be directly matched to a reference library of DNA barcodes or processed further using clustering analysis. Clustering analysis is often performed to improve throughput by removing redundancy in the data such that the input can be used for the more computationally intensive analysis of assigning taxonomy. Clustering methods group reads into operational taxonomic units (OTUs) based on their similarity to other sequences in the samples, and

from which representative or consensus sequences are selected. Commonly-used clustering algorithms are CD-HIT [165] , BlastClust [166] and UPARSE [167]. An OTU is commonly defined as a cluster of reads with 97% similarity, which would be considered as belonging to a unique species according to the DNA barcoding standard [94]. However, the traditionally used 97% similarity threshold is only an approximation. Sometimes two closely related species may have identical barcode sequences [168] or conversely single species may have two or more copies of a DNA barcode marker that differ by more than 3% [169].

Next, tree-based methods and similarity-based methods are most commonly used for assigning query sequences to taxonomy. Tree-based methods assign query sequences to species based on their membership of clusters (or clades) in a barcode tree. This approach is usually based on neighbour joining (NJ) developed by Saitou and Nei [170], and is implemented in BOLD by Ratnasingham and Hebert [80]. The underlying assumption in NJ barcode matching is that distinct species form discrete clusters in a NJ tree [94]. For identification, query sequences are induced in the NJ tree to see in which cluster they appear. Similarity-based BLAST (Basic Local Alignment Search Tool [166]) is probably the most widely used method for classifying DNA sequences in practice. BLAST aligns the query sequence against those present in a selected target database using nearly exact matches of short nucleotide strings (e.g. 10 nucleotides). A similarity score is computed from the portion of the query aligned to the reference sequence. The reference sequence(s) with the highest similarity score is presented along with an indication of the Expect value (E-value), which is the number of hits one can "expect" to see by chance when searching a database of a particular size.

A number of dedicated software pipelines exist that allow processing of metabarcoding datasets followed by taxonomic annotation, including jMOTU and Taxonerator [171], CLOTU [172], QIIME [173], Mothur [174] and UPARSE [167]**.** These software tools have been developed for studying microbial communities using the 16S rRNA gene fragment, but they can also be used for metabarcoding samples containing plants and animals [8,175]. The HTS barcode checker pipeline is an application for automated processing of NGS data to determine whether these contain DNA barcodes obtained from species listed on the CITES appendices [59]. DNA metabarcodes are automatically converted into taxonomic identifications by matching with names on the CITES appendices. By inclusion of a blacklist and additional names database, the HTS barcode checker pipeline prevents false positives and resolves taxonomic heterogeneity.

In DNA metabarcoding, the availability of curated reference databases is of major importance to the assignment of sequences to species. A prerequisite is that reference database should contain accurate sequences that are correctly assigned to taxa with adequate sampling and taxon coverage to fully evaluate both the intraspecific and interspecific variations. Unbalanced representation of certain species, which is expected when dealing with CITES species may greatly affect the analysis. Currently, there are many barcoding campaigns initiated by iBOL to generate DNA barcode data from well-identified and vouchered samples (Box 2.1). Worldwide sequencing efforts have already resulted in more than 2 million COI records from nearly 170,000 species in BOLD. The Barcode Index Number System (BINs) introduced by BOLD is an online framework that automatically clusters animal COI barcode sequences, generating a wiki web page for each cluster [176]. Since clusters show high concordance with species, the framework can be used to verify species identifications as well as document potential new animal species without taxonomic information. BOLD has already reached a good level of standardization and accuracy in terms of the identification of animals but the situation for plants is quite different. The debate about the correct marker(s) to be used as universal barcode has led to a delay in the introduction of plant sequences in the BOLD database [67]. There is also valuable

sequence data archived by the International Nucleotide Sequence Database Collaboration (INSDC; www.insdc.org), which, besides the COI region, is particularly extensive for nrITS and cyt*b* (Box 2.2). However, many of the existing INSDC sequences lack validation in the form of voucher information, making it difficult to detect and remove misidentified specimens or contaminated sequences. Currently the use of local curated reference datasets is often preferred when DNA barcoding is used in plants [177]. Luo *et al*. [178] developed a custom DNA barcoding database for medicinal plant materials, and it accepts plastid DNA markers and nuclear nrITS regions as input (www.cuhk.edu.hk/icm/mmdbd.htm). Furthermore, an online identification module for herbal plant materials has been developed (www.tcmbarcode.cn), which is based around a selection of nrITS2 and *psb*A-*trn*H barcodes from selected medicinal species and their adulterants, substitutes and closely related species. Non-exhaustive list of software available for DNA metabarcoding.

	Description	**Reference**
Software for quality filtering of reads		
PRINSEQ	Application for filtering, reformatting and quality trimming of metagenomic datasets. The software is publicly available through a user-friendly web interface and as stand-alone version.	Schmieder and Edwards [163] http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi
Geneious	Commercially available suite of molecular tools	Kearse *et al*. [162] www.geneious.com
Trimmomatic	A flexible read trimming tool for Illumina NGS data	Bolger *et al*. [164] http://www.usadellab.org/cms/?page=trimmomatic
Software for cluster analysis of reads		
CD-HIT	A fast program for clustering of next-generation sequencing data. The software is publically available through a user-friendly interface and as stand-alone version.	Fu *et al*. [165] http://weizhongli-lab.org/cdhit_suite/cgi-bin/index.cgi
BLASTclust	A program to make non-redundant sequence sets.	Altschul *et al*. [166] http://toolkit.tuebingen.mpg.de/blastclust
Software for assigning reads to taxonomy		
BOLD identification	Species identification system of the Barcode of Life Data Systems (BOLD)	Ratnasingham and Hebert [80] http://www.boldsystems.org/
BLAST	The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program is publically available through a user-friendly web interface and as stand-alone version.	Altschul *et al*. [166] http://blast.ncbi.nlm.nih.gov/Blast.cgi
HTS-barcode-checker	A tool for automated detection of illegally traded species from high-throughput sequencing data	Lammers *et al*. [59] https://github.com/naturalis/HTS-barcode-checker

Software pipelines for DNA metabarcoding		
jMOTU and Taxonerator	Software for turning DNA barcode sequences into annotated OTUs.	Jones *et al*. [171]
QIIME	Quantitative Insights Into Microbial Ecology: bioinformatics pipeline for microbiome analysis from raw DNA sequence data.	Caporaso *et al*. [173] http://qiime.org/
CLOTU	Software for processing amplicon reads followed by taxonomic annotation.	Kumar *et al*. [172]
UPARSE	Pipeline for clustering NGS amplicon reads into OTUs.	Edgar *et al*. [167] http://drive5.com/uparse/
Mothur	Open-source, platform-independent, community-supported software for describing and comparing microbial communities.	Schloss *et al*. [174] http://www.mothur.org/

2.7 Outlook

Next-Generation Sequencing of DNA barcodes, commonly referred to as DNA metabarcoding, is more and more becoming a standard approach for the simultaneous identification and detection of multiple species in complex samples. The approach is similar for both species identification to prevent food fraud and for tracing possible cases of illegal trade of CITES species. A large variety of informative barcodes and mini-barcodes in both the animal and plant area is available, potentially allowing for a clear-cut identification of species present in a sample of interest. However, comprehensive identification of (endangered and/or protected) species in complex forensic samples is not yet fully feasible at this moment. This is due to a number of reasons. In the first place, no truly universal DNA isolation method is available for all the different matrixes seized by the customs and CITES authorities. In-house developed protocols or commercially available kits or a combination of both are typically assessed in an attempt to obtain amplifiable DNA from forensic samples, which will increase time and cost. In many cases, the poor success of extraction and PCR amplification of DNA from forensic samples hinder effective identification of species. Accordingly, systematic studies are needed to optimise DNA isolation methods and efficiency to satisfy the stakeholders needs, which is to obtain a robust and rapid DNA isolation method that can be applied across on a wide range of (wildlife) forensic samples, and which would maximize DNA purity and yield, whilst reducing any further DNA damage.

Secondly, forensic samples are often heavily processed and may contain severely fragmented DNA, thus hampering the ability to PCR amplify full-length barcodes. In such cases, mini-barcodes are often the only alternative, but these do not always provide species level resolution and truly universal primers for mini-barcode amplification have been found difficult to design. Universal primers should be used that minimize PCR bias caused by variable primer-template mismatches across species to ensure that all species can be detected [69,70]. Several mini-barcodes have been proposed, but especially for plants no universal mini-barcode standard to provide species-level resolution has so far been adopted. The power of DNA metabarcoding is that a panel of different barcodes and mini-barcodes can efficiently be analysed in parallel. Such a strategy will provide improved resolution at the species level when some barcodes fail to resolve, while verifying species with multiple barcodes contributes to enhanced quality assurance.

Thirdly, the current underrepresentation of DNA barcodes from species protected under CITES and closely related species critically hamper their identification. This will improve as DNA barcoding campaigns continue, in particular through initiative such as the Barcode of Wildlife Project (BWP; www.barcodeofwildlife.org). The latter project aims to construct a public DNA barcode reference library for 2,000 endangered plant and animal species, thereby paving the way for the use of DNA barcodes in a court of law to provide strong evidence against those involved in poaching and trafficking of species protected by CITES.

Finally, it will be necessary to develop and validate bioinformatics pipelines for the detection and identification of endangered species using DNA metabarcoding strategies. Several dedicated software tools have been developed, but there is a need to validate pipelines for clustering of reads into OTUs, using benchmarked algorithms for quality control, de-noising, chimera removal, and OTU picking.

Concluding, the DNA metabarcoding approach holds great promise for detecting and identifying endangered plant and animal species in complex forensic samples. However, validation of the approach should be performed before DNA metabarcoding can be applied in a routine setup. By making use of DNA-barcoded reference species in well-characterized experiment complex products, or as internal controls in real-life samples it can be assessed whether the DNA metabarcoding procedure is able to accurately and concurrently identify various target plant and animal species. Only when DNA metabarcoding has been demonstrated to be robust and transferable across laboratories can the method truly be implemented in routine testing. In that sense, we are just at the beginning of exploring the broad applications of DNA metabarcoding to reveal the composition of complex products in the light of, for instance, food fraud and the illegal trading of endangered plant and animal species.

Chapter 3

Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples

This chapter was published as: **Arulandhu AJ**, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Arne Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale C, Niekamp H, Perri E, Barbante A, Rosec JP, Seyfarth R, Sovová T, van Moorleghem C, van Ruth S, Peelen T and Kok EJ. "Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples". *GigaScience* 2017; 6(10): 1-8.

Abstract

Background DNA metabarcoding provides great potential for species identification in complex samples such as food supplements and traditional medicines. Such a method would aid CITES (the Convention on International Trade in Endangered Species of Wild Fauna and Flora) enforcement officers to combat wildlife crime by preventing illegal trade of endangered plant and animal species. The objective of this research was to develop a multi-locus DNA metabarcoding method for forensic wildlife species identification and to evaluate the applicability and reproducibility of this approach across different laboratories.

Results A DNA metabarcoding method was developed that makes use of 12 DNA barcode markers that have demonstrated universal applicability across a wide range of plant and animal taxa, and that facilitate the identification of species in samples containing degraded DNA. The DNA metabarcoding method was developed based on Illumina MiSeq amplicon sequencing of well-defined experimental mixtures, for which a bioinformatics pipeline with user-friendly web interface was developed. The performance of the DNA metabarcoding method was assessed in an international validation trial by 16 laboratories, in which the method was found to be highly reproducible and sensitive enough to identify species present in a mixture at 1% dry weight content.

Conclusion The advanced multi-locus DNA metabarcoding method assessed in this study provides reliable and detailed data on the composition of complex food products, including information on the presence of CITES-listed species. The method can provide improved resolution for species identification, while verifying species with multiple DNA barcodes contributes to an enhanced quality assurance.

Keywords Endangered species, CITES, Traditional medicines, DNA metabarcoding, Customs agencies, COI, *matK*, *rbcL*, cyt *b*, mini-barcodes.

3.1 Background

The demand for endangered species as ingredients in traditional medicines (TMs) has become one of the major threats to the survival of a range of endangered species such as seahorse (*Hippocampus* sp.), agarwood (*Aquilaria* sp.), and Saiga antelope (*Saiga tatarica*) [10,179,180]. The Convention on the International Trade in Endangered Species of Wild Fauna and Flora (CITES) is one of the best supported conservation agreements to regulate trading of animal and plant species (www.cites.org) and thereby conserve biodiversity. Currently, ~35,000 species are classified and listed by CITES in three categories based on their extinction level (CITES Appendix I, II and III) by which the trade in endangered species is regulated. The success of CITES is dependent upon the ability of customs inspectors to recognize and identify components and ingredients derived from endangered species, for which a wide range of morphological, chromatographic and DNA-based identification techniques can be applied [4,5].

Recent studies have shown the potential of DNA metabarcoding for identifying endangered species in TMs and other wildlife forensic samples [8,9,64,181]. DNA metabarcoding is an approach that combines DNA barcoding with next-generation sequencing (NGS), which enables sensitive high-throughput multispecies identification on the basis of DNA extracted from complex samples [60]. DNA metabarcoding uses more or less universal PCR primers to mass-amplify informative DNA barcode sequences [18,182]. Subsequently, the obtained DNA barcodes are sequenced and compared to a DNA sequence reference database from well-characterized species for taxonomic assignment [60,182]. The main advantage of DNA metabarcoding over other identification techniques is that it permits the identification of all animal and plant species within samples that are composed of multiple ingredients, which would not be possible through morphological means and time-consuming with traditional DNA barcoding [8,9,64]. Furthermore, the use of mini-barcode markers in DNA metabarcoding facilitate the identification of species in highly processed samples containing heavily degraded DNA [8,9]. Such a molecular approach could aid the Customs Authorities to identify materials derived from endangered species in a wide variety of complex samples, such as food supplements and TMs [183].

Before routine DNA metabarcoding can be applied, there are some key issues that need to be taken into account. First, complex products seized by Customs, such as TM products, may contain plant and animal components that are highly processed, and from which the isolation of good quality DNA is challenging. Second, the universal DNA barcodes employed may not result in amplification of the related barcode for each species contained in a complex sample, due to DNA degradation or the lack of PCR primer sequence universality. For plants, for example, different sets of DNA barcodes have been suggested for different fields of application (i.e. general taxonomic identification of land plants, identification of medicinal plants, etc.), and none of them meet the true requirements of universal barcodes [184]. Also, whilst PCR primers can be designed to accommodate shorter DNA barcode regions for degraded DNA samples, such mini-barcodes contain less information and their primers are more restrictive, often making them unsuitable for universal species barcoding [38,64]. The third challenge is the reference sequence database quality and integrity, which is particularly problematic for law enforcement issues, where high quality and reliability are essential. The current underrepresentation of DNA barcodes from species protected under CITES and closely related species critically hampers their identification. The fourth challenge is that a dedicated bioinformatics pipeline is necessary to process raw NGS data for accurate and sensitive identification of CITES-listed species [18]. Finally, studies using the DNA metabarcoding approach are scarce and none of these methods have been truly validated [18,185]. Therefore, before implementing DNA metabarcoding by Customs and other enforcement agencies, the above-mentioned challenges need to be thoroughly assessed to ensure accurate taxonomic identifications. [18]

The objective of this research was to develop a multi-locus DNA metabarcoding method for (endangered) species identification and to evaluate the applicability and reproducibility of this approach in an international interlaboratory study. The research was part of a larger programme on

the development of advanced DNA-based methods from the DECATHLON project (www.decathlon-project.eu), within the European Union's Framework Programme 7. In the process of establishing the standard operating procedure (SOP) for multi-locus DNA metabarcoding, all important aspects of the procedure (i.e. DNA isolation procedure, DNA barcode marker, barcode primers, NGS strategy and bioinformatics) were evaluated. The challenges concerning the quality and integrity of the DNA reference database(s) are discussed. The first step was aimed at identifying an ideal DNA isolation method to extract DNA from complex mixtures consisting of both animal and plant tissues. Secondly, animal and plant DNA barcode markers and corresponding primer sets were identified from literature that allowed good resolution for identifying (endangered) species from a wide taxonomic range. Thirdly, a panel of universal plant and animal DNA barcodes was selected and a single optimal PCR protocol was identified for efficient amplification of a panel of DNA barcode markers. Finally, the suitability of the Illumina MiSeq NGS technology was evaluated, and a bioinformatics pipeline with a user-friendly web interface was established to allow stakeholders to perform the NGS data analysis without expert bioinformatics skills.

The DNA metabarcoding method was developed and tested based on data generated for 15 well-defined complex mixtures. The use of well-characterised mixtures allowed for optimising the bioinformatics procedure and subsequent robustness testing of multiple parameter settings and thresholds. The practical performance and reproducibility of the DNA metabarcoding strategy was assessed in an international validation trial by 16 laboratories from 11 countries, on the basis of eight other newly composed complex mixtures and two seized TMs, which were suspected to contain ingredients derived from CITES species. In this study, the multi-locus DNA metabarcoding method is presented and it is assessed whether the method can improve the compositional analysis of complex and real-life samples by enabling the sensitive and reproducible identification of CITES-listed taxa by enforcement agencies and other laboratories.

3.2 Data description

To constitute well-defined complex mixtures, 46 reference specimens were commercially purchased from shops or were provided by the Dutch Custom Laboratory. In addition, two TMs that were suspected to comprise endangered species material were also obtained from Dutch Customs Laboratory. Each reference specimen was identified morphologically. Genomic DNA was extracted from 29 animal and 17 plant reference species for DNA barcoding. Standard cytochrome c oxidase I (COI) barcodes for all animal specimens were generated and individually sequenced using the Sanger method, and compared against the Barcode of Life Data Systems and NCBI database for taxonomic confirmation. For plant species, the DNA barcodes *rbcL* and *matK* were sequenced to confirm species identity. For a number of plant and animal species the generated barcode sequence information was deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and LT718651 (Additional file 3A; Table S1).

The complex mixtures for the pilot study and interlaboratory validation trial were prepared with 2 to 11 taxonomically well-characterised species present in relative concentrations (dry mass: dry mass) from 1% to 47%. For all experimental mixtures in the interlaboratory trial, internal control species were used to verify the efficiency of homogenization and to check for possible sample cross-contamination using species specific qPCR assays. DNA was isolated from the complex mixtures and the concentration and purity of extracted DNA was determined using spectrophotometer (NanoDrop 1000, Thermo Fisher Scientific Inc.). Subsequently, PCR amplifications using 12 DNA barcode primer sets were performed. The pooled and purified amplicons of each sample were sequenced using an Illumina MiSeq paired-end 300 technology, following the manufacturer's instructions (Illumina, Inc.). The NGS datasets were analysed using the CITESspeciesDetect pipeline. All raw NGS datasets from both analyses were deposited in ENA under accession numbers ERS1545972 to ERS1545988,

ERS1546502 to ERS1546533, ERS1546540 to ERS1546619, ERS1546624 to ERS1546639, ERS1546742 to ERS1546757, ERS1546759 to ERS1546774, and study number PRJEB18620 (Additional file 3C; Table S1). A web interface was developed for the CITESspeciesDetect pipeline to allow stakeholders to perform the NGS data analysis of their own samples. The web interface can be globally accessed via the SURFsara high-performance computing and data infrastructure (http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/).

3.3 Analyses

3.3.1 Establishing a laboratory procedure for multi-locus DNA barcode amplification

Based on the previous studies on DNA isolation for TMs [44,64] and from the comparison between modified Qiagen DNeasy plant mini kit [186] and CTAB isolation [187] (unpublished results), we identified that the CTAB isolation method in general yields better DNA purity and provides better PCR amplification success. Therefore, the CTAB DNA isolation method was selected for successive experiments.

The DNA barcode markers included in this study were selected based on Staats et al. [18] supplemented with additional primers from literature [38] (Table 3.1). DNA barcode markers were selected based on the availability of universal primer sets and DNA sequence information in public repositories [18]. Important considerations in selecting suitable primer sets were that, preferably, they are used in DNA barcoding campaigns and studies, and as such have demonstrated universal applicability across a wide range of taxa. Furthermore, primer sets for both the amplification of full-length barcodes and their respective mini-barcodes (i.e. short barcode regions < 300 nt within existing ones) were selected when available. This was done to facilitate PCR amplification from a range of wildlife forensic samples containing relatively intact DNA (using full-length barcodes) and/or degraded DNA (mini-barcodes). Based on these criteria, PCR primer sets for the following animal DNA barcodes were selected: regions of the mitochondrial genes encoding 16S rRNA gene (16S), cytochrome c oxidase I (COI) and cytochrome *b* (cyt *b*). For plant species identification, primer sets for the following DNA barcodes were selected: regions of the plastidial genes encoding maturase K (*matK*), ribulose-1,5-bisphosphate carboxylase (*rbcL*), $tRNA^{Leu}$ (UAA) intron sequence (*trnL* (UAA)), *psbA-trnH* intergenic spacer region (*psbA-trnH*), and the nuclear internal transcribed spacer 2 (ITS2) region (Table 3.1). The selected primers sets were modified to include the Illumina adapter sequence at the 5′ end of the locus-specific sequence to facilitate efficient NGS library preparation. A gradient PCR experiment was performed to identify the optimal PCR annealing temperature. While the selected PCR primer sets had previously been published with their own annealing temperatures and conditions, the identification of a single optimal annealing temperature for all PCR primer sets would allow for increased efficiency of analysis. Initially, a thermal gradient of 49.0 ˚C to 55.0 ˚C was tested on the *Bos taurus* reference material with the primer sets for COI, 16S, mini-16S, and cyt *b*. The amplification efficiency across the PCR primers sets was determined by comparing the intensity of the amplicons across the thermal gradient. An optimal annealing temperature of 49.5 ˚C was identified, but additional non-specific amplicons were observed with some primers (not shown). To reduce the amounts of non-specific amplification products, the PCR program was modified to increase the annealing temperature after five cycles from 49.5 ˚C to 54.0 ˚C [31], and tested on all 15 PCR primer sets (Table 3.1). It was observed that certain PCR primer combinations still produced non-specific products (for *psbA-trnH* gene) or less intense PCR products (for *rbcL* gene with primers rbcLa-F and rbcLajf634R, and *matK* gene with primers matK-390f and matK-1326r). Consequently, these PCR primer sets were excluded from subsequent experiments.

Next, the selected PCR thermocycling protocol was evaluated with the remaining 12 PCR primer sets on a panel of 29 animal and 17 plant species, representing a phylogenetically wide range of taxa (Mammalia, Actinopterygii, Malacostraca, Bivalvia, Aves, Reptilia, Amphibia, Insecta,

Angiospermae, and Cycadopsida; Additional file 3A; Table S2 and S3). The overall PCR amplification success rates varied across reference species and across DNA barcode markers (Additional file 3A; Table S2). For instance, no PCR amplification was observed with cyt *b* for the CITES-listed species *Balaenoptera physalus*, whereas intense amplification was seen for the same species with 16S, COI, mini-16S and mini-COI (Additional file 3A; Table S2). Overall, at least one DNA barcode marker could successfully be amplified for each of the 46 plant and animal species (Additional file 3A; Table S2 and S3). For a number of plant and animal species the generated barcode sequence information was deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and LT718651 (Additional file 3A; Table S1).

Table 3.1 Overview of the PCR primer sets used in this study for amplifying plant and animal DNA barcodes and mini-barcodes.

DNA Marker	Primer name	Primer sequence 5′-3′	Amplicon length (nt)	Reference
Universal animal DNA barcodes and mini-barcodes				
16S	16sar-L	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGCCTGTTTATCAAAAACAT	500-600	Palumbi [32]
	16sar-H	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCGGTCTGAACTCAGATCACGT		
mini-16S	16S-forward	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAYAAGACGAGAAGACCC	250	Sarri *et al.*[33]
	16S-reverse	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATTGCGCTGTTATTCC		
COI*	LepF1_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATTCAACCAATCATAAAGATATTGG	648	Modified from Ivanova *et al.* [31]
	VF1_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCAACCAACCACAAAGACATTGG		
	VF1d_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCAACCAACCACAARGAYATYGG		
	VF1i_t1	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTTCTCAACCAACCAIAAIGAIATIGG		
	LepR1_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAAACTTCTGGATGTCCAAAAAATCA		
	VR1d_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACTTCTGGGTGGCCRAARAAYCA		
	VR1_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACTTCTGGGTGGCCAAAGAATCA		
	VR1i_t1	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAGACTTCTGGGTGICCIAAIAAICA		
mini-COI	mlCOIintF	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGWACWGGWTGAACWGTWTAYCCYCC	313	Leray *et al. [110].*, Geller *et al.*[137]
	jgHCO2198	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTAIACYTCIGGRTGICCRAARAAYCA		
cyt *b*	L14816	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATCCAACATCTCAGCATGATGAAA	743	Palumbi [32], Parson *et al.*[114]
	CB3-H	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGCAAATAGGAARTATCATTC		
mini-cyt *b*	L14816	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCATCCAACATCTCAGCATGATGAAA	357	Parson *et al.*[114]
	H15173	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCCTCAGAATGATATTTGTCCTCA		
Universal plant DNA barcodes and mini-barcodes				
matK	matK-KIM1R	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACCCAGTCCATCTGGAAATCTTGGTTC	656-889	Fazekas *et al.*[188]
	matK-KIM3F	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGTACAGTACTTTTGTGTTTACGAG		
matK&	matK-390f	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGATCTATTCATTCAATATTTC	656-889	Cuénoud *et al.*[144]
	matK-1326r	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTCTAGCACACGAAAGTCGAAGT		
rbcL	rbcLa-F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTCACCACAAACAGAGACTAAAGC	654	Levin *et al.* [142]
	rbcLa-R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTAAAATCAAGTCCACCRCG		Kress and Erickson[122]

rbcL&	rbcL a-F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGTCACCACAAACAGAGACTAAAGC	607	Levin *et al.* [142]
	rbcLajf634R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAAACGGTCTCTCCAACGCAT		Fazekas *et al.* [35]
mini-*rbcL*	F52	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTGGATTCAAAGCTGGTGTTA	140	Little[38]
	R193	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCVGTCCAMACAGTWGTCCATGT		
trnL (UAA)	c	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCGAAATCGGTAGACGCTACG	767	Taberlet *et al.*[60]
	d	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGGGGATAGAGGGACTTGAAC		
trnL (P6 loop)	g	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGCAATCCTGAGCCAA	10-143	Taberlet *et al.*[60]
	h	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCATTGAGTCTCTGCACCTATC		
ITS2	S2F	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGATGCGATACTTGGTGTGAAT	160-320	Chen *et al.*[123]
	S3R	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGACGCTTCTCCAGACTACAAT		
psbA-trnH&	psbAf	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTATGCATGAACGTAATGCTC	264-792	Sang *et al.* [147], Tate and Simpson [148]
	trnH2	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCGCGCATGGTGGATTCACAATCC		

The shaded text represents the sequence of the Illumina overhang adapters.
**Modified COI cocktail primers without M13-tails were used [31].*
& The primers were not included in the final panel of DNA barcodes.

3.3.2 Development and pre-validation of the CITESspeciesDetect bioinformatics pipeline

A dedicated bioinformatics pipeline, named CITESspeciesDetect, was developed for the purpose of rapid identification of CITES-listed species using Illumina paired-end sequencing technology. Illumina technology was selected because it produces NGS data with very low error rates, compared to other technologies [2, 19]. Furthermore, the Illumina MiSeq platform enables paired-end read lengths of up to 300 nt, allowing relatively long DNA barcode regions of up to ~550 nt to be assembled. Also, the multiplexing capabilities of Illumina technology are well developed, allowing for simultaneous sequencing of multiple samples in one run, thereby enabling more cost-efficient NGS. While NGS data analysis pipelines exist that allow processing of Illumina DNA metabarcoding datasets (e.g. CLOTU, QIIME, Mothur), the majority have been developed for specifically studying microbial communities using the 16S rRNA gene region. CITESspeciesDetect, developed in this study, extends on the frequently-used software tools developed within the USEARCH [189] and BLAST+ packages [166], and additionally includes dedicated steps for quality filtering, sorting of reads per barcode, and CITES species identification (Figure 3.1). The CITESspeciesDetect is composed of five linked tools and data analysis passes through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming and filtering of reads, followed by sorting by DNA barcode, 2) Operational Taxonomic Unit (OTU) clustering by barcode, and 3) taxonomy prediction and CITES identification.

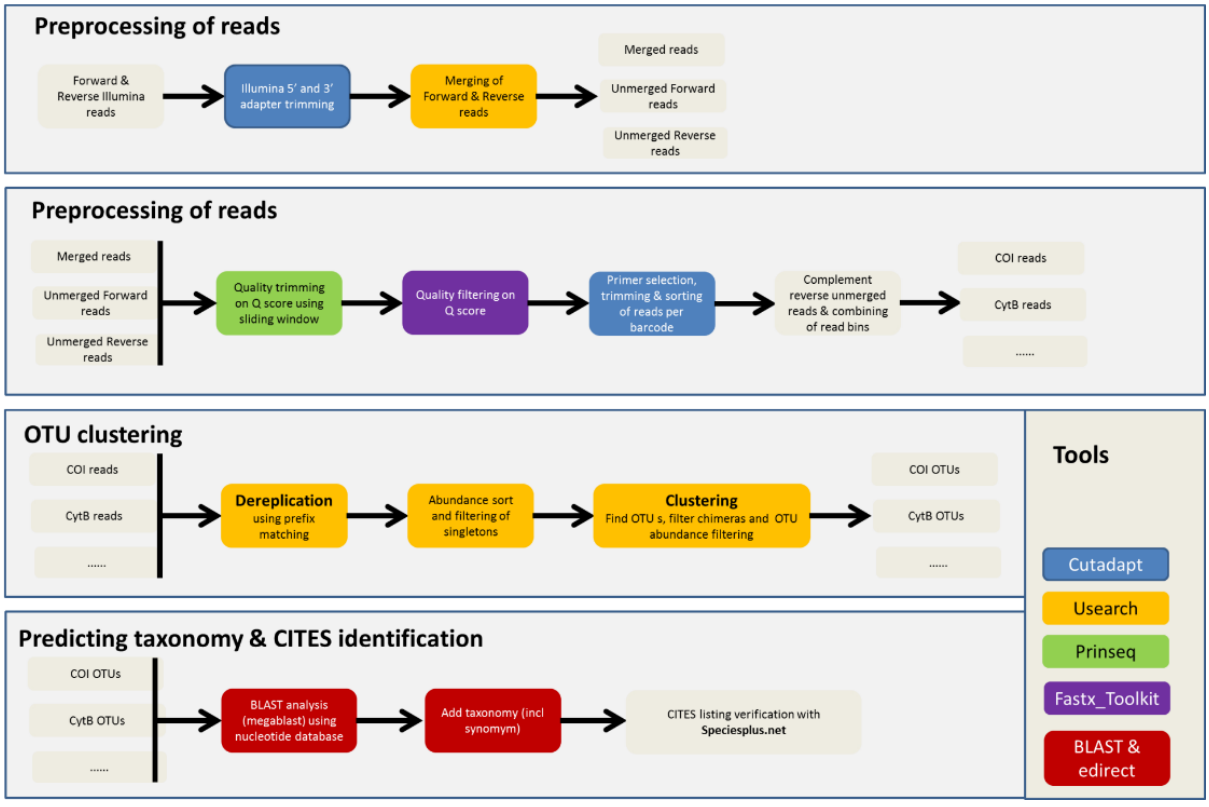


Figure 3.1 Schematic representation of the CITESspeciesDetect pipeline.

It was found that with the current setup of the pipeline, reads generated for cyt *b* and mini-cyt *b* could not be separated based on the forward PCR primer, as the forward primers are identical. It was therefore decided to combine (pool) the overlapping reads of cyt *b* and mini-cyt *b* during pre-processing (primer selection) of reads to prevent reads from being double selected. This means that the results of cyt *b* and mini-cyt *b* are presented by the CITESspeciesDetect pipeline as cyt *b*. The same issue was found for COI barcode and mini-barcode markers, for which the results are presented as COI.

A parameter scan was performed in order to assess the effect of software settings on the ability to identify species. The evaluation allowed for the identification of important parameters and their effect on the sensitivity, specificity and robustness of the procedure. Changing the base quality score has a major impact on the number of reads per barcode (Additional file 3A; Table S4). Increasing the strictness of the base quality score resulted in decreasing numbers of reads per barcode. Quality score values other than the default values (Q20 for 95% of bases) did not yield better identifications. When applying strict quality filtering settings (Q20 for 100% of bases, or Q30 for 99% of bases) the species *Pieris brassicae* and *Anguilla anguilla* could not be detected with cyt *b* and/or mini-COI, indicating these settings were too strict (Additional file 3A; Table S5). This is likely due to the resulting overall low read numbers for cyt *b* and mini-COI when applying these strict quality filtering settings (Additional file 3A; Table S4).

An OTU abundance threshold is generally applied to make DNA metabarcoding less sensitive to (potential) false-positive identifications. False-positives may occur e.g. as contaminants during pre-processing of samples (DNA extraction, PCR) or as cross-contamination during Illumina sequencing. Applying an OTU abundance threshold higher than zero generally results in loss of sensitivity. We have found, however, that applying an OTU abundance threshold of higher than zero may help in reducing noisy identifications and potential false-positive identifications (results not shown). It should be noted that applying filtering thresholds may always lead to false negative or false positive identifications. In this study, an OTU abundance threshold of 0.2% was set as default, however, the OTU abundance threshold may need re-evaluation for samples with expected very low species abundances (< 1% dry weight).

The effect of applying a minimum DNA barcode length revealed that allowing DNA barcodes of ≥ 10 nt did not lead to additional identification of species, compared with default settings (e.g. ≥ 200 nt). Increasing the minimal DNA barcode length to 250 nt, however, resulted in a failure to identify most plant species with mini-*rbcL* and *rbcL*. We implemented a minimum DNA barcode length of 200 nt, except for DNA barcodes with a basic length shorter than 200 nt, in which case the minimum expected DNA barcode length is set to 100 nt for ITS2, 140 nt for mini-*rbcL*, and 10 nt for the *trn*L (P6 loop) marker.

The results of the parameter scan resulted in specifying recommended parameter values (default setting) for analysing DNA metabarcoding datasets using the CITESspeciesDetect pipeline (see Methods section "Bioinformatics analysis"). An online version of the CITESspeciesDetect pipeline with a user-friendly web-interface was developed for skilled analysts with basic, but no expert level knowledge in bioinformatics and is made available via http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/.

3.3.3 Pilot study to assess the performance of the DNA metabarcoding procedure using experimental mixtures

The DNA metabarcoding procedure was assessed in a pilot study, for which 15 complex mixtures (EM1 to EM15) were prepared containing from 2 to 10 taxonomically well-characterised species with DNA barcode reference sequences available in the NCBI reference database (Table 3.2). The experimental mixtures 10 and 11 (EM10 and EM11) were independently analysed twice to verify repeatability of the

method (DNA isolation, barcode panel analysis and pooling). Only mixtures were used with well-characterised species (DNA Sanger barcoded and taxonomically verified) ingredients, at known dry weight concentrations, and with high quality DNA that would allow for an assessment of the performance of the DNA metabarcoding method under optimal conditions.

A total of 2.37 Gb of Illumina MiSeq sequencing data was generated for the 17 complex samples (15 complex mixtures along with the two replicates). On average, 464,648 raw forward and reverse Illumina reads were generated per sample, with minimum and maximum read numbers ranging between 273,104 (mixture EM4) and 723,130 (mixture EM10R; Table 3.3). During raw data pre-processing with the default settings of the CITESspeciesDetect pipeline, the reads were first quality filtered and overlapping paired-end Illumina reads were merged into pseudo-reads (Figure 3.1). The samples contained on average 269,099 quality controlled (QC) unmerged (forward and reverse) reads and merged pseudo-reads, collectively named (pseudo)reads. On average 88.27% (min = 77.38%, max = 96.26%) of raw reads passed the quality filtering and pre-processing steps, indicating that the overall quality of the Illumina data was high (not shown).

Next, the (pseudo)reads were assigned to DNA barcodes based on PCR primer sequences. On average, 96.44% (min = 88.78%, max = 98.21%) of QC pre-processed reads were assigned to DNA barcodes, indicating a high percentage of reads containing the locus-specific DNA barcode primers (Table 3.3). After this, the (pseudo)reads were clustered by 98% sequence similarity into OTUs. On average, 82.26% (min = 75.11%, max = 90.63%) of the DNA barcodes assigned reads were clustered into OTUs (Table 3.3). It was assumed that the small fraction of reads that was not assigned to OTUs contained non-informative (e.g. non-specific fragments, chimers) sequences that may have been generated during PCR amplification, and were filtered out during clustering.

For taxonomy prediction, OTUs were assigned to dataset sequences using BLAST when aligning with at least 98% sequence identity, a minimum of 90% query coverage, and an E-value of at least 0.001. Generally, the best match ("top hit") is used as best estimate of species identity. However, species identification using BLAST requires careful weighting of the evidence. To minimize erroneous taxonomic identifications a more conservative guideline was used that allowed a species to be assigned only when the best three matches identified the species. If the bit scores do not decrease after the top three hits, or if other species have identical bit scores, then identification was considered inconclusive. In such cases, OTUs were assigned to higher taxonomic levels (genus, family or order). All animal ingredients, except *Parapenaeopsis* sp. could be identified at the species-level with one or more DNA barcode marker using the default settings of the CITESspeciesDetect pipeline (Table 3.4 and 5). For plants, *Lactuca sativa* could be identified at the species-level using the *trn*L (P6 loop). All other plant taxa were identified at the genus or higher level (Table 3.4 and 3.5).

Table 3.2 Pilot study: Composition of the experimental mixtures, and taxa identified using the default setting of the CITESspeciesDetect pipeline.

		Experimental mixtures																
Species/Genus	**Common name**	**EM1**	**EM2**	**EM3**	**EM4**	**EM5**	**EM6**	**EM7**	**EM8**	**EM9**	**EM10**	**EM10R**	**EM11**	**EM11R**	**EM12**	**EM13**	**EM14**	**EM15**
Bos taurus	Cattle	99% (S)	90% (S)	1% (S)	10% (S)	99% (S)	95% (S)	85% (S)			10% (S)	10% (S)	46% (S)	46% (S)	95% (S)	85% (S)		
Parapenaeopsis sp.	Shrimp						1%	3%			10%	10%	1%	1%			1%	3%
Anguilla anguilla *	European eel						1%	3%			**10**% (S)	**10**% (S)	**1**% (S)	**1**% (S)			**1**% (S)	**3**% (S)
Crocodylus niloticus *	Nile crocodile						**1**% (S)	**3**% (S)									**1**% (S)	**3**% (S)
Gallus gallus	Domestic chicken						1% (S)	3% (S)			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
Pieris brassicae	Large white (caterpillar)						1% (S)	3% (S)			10% (S)	10% (S)	1% (S)	1% (S)			1% (S)	3% (S)
Echinocactus sp. *	Barrel cactus								**1**% (F)	**3**% (F)	**10**% (F)	**10**% (F)	**1**% (F)	**1**% (F)	**1**% (F)	**3**% (F)		
Euphorbia sp. *	Spurge								**1**% (F)	**3**% (F)	**10**% (F)	**10**% (F)	**1**% (F)	**1**% (F)	**1**% (F)	**3**% (F)		
*Aloe variegata *,&*	Tiger aloe					1% (F)			1% (F)	3% (F)	10% (F)	10% (F)	1% (F)	1% (F)	1% (F)	3% (F)		
Dendrobium sp. *	Dendrobium (orchid)								**1**% (F)	**3**% (G)					**1**% (G)	**3**% (G)		
Cycas revoluta *	Sago palm								1%	3%	**10**% (G)	**10**% (G)	1%	**1**% (G)	**1**% (G)	**3**% (G)		
Lactuca sativa	Lettuce	1% (S)	10% (S)	99% (S)	90% (S)				95% (S)	85% (S)	10% (S)	10% (G)	46% (S)	46% (S)			95% (S)	85% (S)

Taxa were identified at the species-level unless otherwise indicated in brackets. Cells highlighted in grey indicate taxa that were not identified. Identified taxa listed by CITES are highlighted in bold.

The symbol next to percentage indicates the taxonomic resolution of the identified taxon: (F) – Family level, (G) – Genus level and (S) – Species level

** Species listed by CITES. & Aloe variegata (synonym Gonialoe variegata) was recently assigned to the genus Gonialoe [190].*

Table 3.3 Pilot study: average number of Illumina MiSeq reads, the average number of (pseudo)reads that passed quality control (QC) and the percentage of QC (pseudo)reads that were assigned to DNA barcodes and Operational Taxonomic Units (OTUs) generated per sample.

Experimental mixture	Number of raw reads	Percentage of QC (pseudo)reads*	Percentage DNA barcode assigned (pseudo)reads*	Percentage OTU clustered (pseudo)reads*
EM1	466,108	88.07	95.68	83.86
EM2	448,428	86.04	97.24	84.04
EM3	496,328	87.46	96.61	84.34
EM4	273,104	77.38	95.74	80.54
EM5	582,254	96.26	97.84	90.63
EM6	442,574	92.81	97.54	81.48
EM7	394,354	93.04	97.14	80.70
EM8	455,172	79.62	95.66	82.35
EM9	434,326	86.23	97.30	83.60
EM10	387,816	87.73	97.00	75.11
EM10R	723,130	95.59	98.02	87.39
EM11	363,374	84.44	96.74	78.63
EM11R	635,304	91.11	98.21	87.01
EM12	355,634	92.55	97.54	76.54
EM13	405,742	89.46	96.49	77.31
EM14	480,772	85.74	95.98	81.91
EM15	554,602	87.05	88.78	82.98
Average**	464,648	88.27	96.44	82.26

* *(pseudo)reads are the combined quality controlled (QC) pseudo-reads, and the QC processed unmerged forward and reverse reads.*

** *Averaged across the 17 Illumina MiSeq datasets.*

Table 3.4 Taxonomic resolution provided by each DNA barcode marker for EM10 and EM10R.

Species/Genus	Species	Genus	Family
Anguilla anguilla	cyt *b*	**mini-16S**	
Parapenaeopsis sp.			
Bos taurus	**16S, mini-16S, cyt *b*, COI**		
Gallus gallus domesticus	**mini-16S, cyt *b*, COI**		
Pieris brassicae	**COI**		
Echinocactus sp.			***matK*, *rbcL*, mini-*rbcL*, ITS2**
Euphorbia sp.		***rbcL*, mini-*rbcL***	**ITS2**
Aloe variegata			*matK*, ***rbcL*, mini-*rbcL***, *trn*L (UAA)
Cycas revoluta		***rbcL*-mini, *trnL* (P6 loop)**	
Lactuca sativa	*trnL* (P6 loop)	*matK*, ***trnL* (UAA), ITS2**	***rbcL*, mini-*rbcL***

Highlighted in bold are DNA barcodes with the same taxonomic resolution in both sample

Putative contaminating species were observed in most of the experimental mixtures from multiple markers, detailed information about the identified cross-contained species in a sample and the related markers are specified in the Additional file 3B; Table S1. Even with the default OTU abundance threshold in place, the species *L. sativa*, *B. taurus* and *Gallus gallus* were identified in mixtures that were not supposed to contain these species. To verify whether these putative contaminations occurred during DNA isolation or Illumina sequencing, qPCR assays for the specific detection of *B. taurus* and *G. gallus* were performed on selected DNA extracts. The high Cq values above 39 indicated the presence of these

species, however, in low copy number, which suggests that for some experimental mixtures (EM8, EM9 and EM14) cross-contamination had occurred during sample preparation or DNA isolation, while for other experimental mixtures (EM15) cross-contamination may have occurred during PCR, Illumina library preparation or sequencing. In addition to these contaminants, a species of *Brassica* was identified in experimental mixtures containing *P. brassica.* This result is most likely not a false-positive, because the caterpillars used for this study had been fed on cabbage.

Table 3.5 Taxonomic resolution provided by each DNA barcode marker for EM11 and EM11R.

Species/Genus	Species	Genus	Family
Anguilla anguilla	**cyt *b***		
Parapenaeopsis sp.			
Bos taurus	**16S, mini-16S, cyt *b*, COI**		
Gallus gallus domesticus	**cyt *b***, COI		
Pieris brassicae	**COI**		
Echinocactus sp.			***matK*, *rbcL*, ITS2**
Euphorbia sp.		***rbcL*, mini-*rbcL***	
Aloe variegata			***matK*, *rbcL*, mini-*rbcL*, *trn*L (UAA)**
Cycas revoluta		mini-*rbcL*, *trnL* (P6 loop)	
Lactuca sativa	***trnL* (P6 loop)**	***matK*, *rbcL*, *trnL* (UAA), ITS2**	*rbcL*, **mini-*rbcL***

Highlighted in bold are DNA barcodes with the same taxonomic resolution in both samples

The DNA metabarcoding method was found to be sensitive enough to identify most plant and animal taxa at 1% (dry mass: dry mass) in mixtures of both low (EM1, EM3 and EM5; Table 3.2) and relatively high complexity (EM6, EM8, EM11, EM12, and EM14; Table 3.2). The exception being *Parapenaeopsis* sp. (all mixtures), *A. anguilla* in EM6, and *Cycas revoluta* in EM8 and EM11. Careful inspection of the NGS data revealed that in nearly all cases OTUs related to *Parapenaeopsis* sp., *A. anguilla*, and *C. revoluta* were present, but that these sequences had been filtered out by the CITESspeciesDetect pipeline because their cluster sizes did not fulfil the 0.2% OTU abundance threshold. There appeared to be no trend as to the type and length of DNA barcode marker that had been filtered out by the CITESspeciesDetect pipeline. For instance, *Parapenaeopsis* sp. was detected below the OTU threshold with cyt *b*, mini-16S, COI, and 16S markers (not shown). Lowering the OTU abundance threshold, however, would lead to (more) false-positive identifications, and this was therefore not implemented.

The repeatability of the laboratory procedure (excluding NGS) was assessed by analysing the experimental mixtures 10 and 11 (EM10R and EM11R; Table 3.2), which was independently performed twice, i.e. DNA isolation and PCR barcode amplification, but NGS was performed on the same MiSeq flow cell as the other samples of the pilot study. From the comparison, it was observed that the percentage of QC reads was nearly twice as high in the replicate analyses (Table 3.3). Also, the percentage of QC reads assigned to DNA barcodes varied among replicate analyses (Figure 3.2). Most notable were the observed differences among replicate analyses in the percentage reads assigned to *matK* and the *trnL* (P6 loop). For example, the percentage of QC reads assigned to *matK* were 6.11% (14081 reads) and 0.02% (97 reads) in EM10 and EM10R respectively (Figure 3.2). The low number of reads assigned to *matK* limited its use for taxonomy identification in EM10R (Table 3.4). The multi-locus approach, however, allowed for the repeatable identification of taxa in EM10 and EM11, though not in all cases with all DNA barcode markers (Table 3.4 and 3.5).

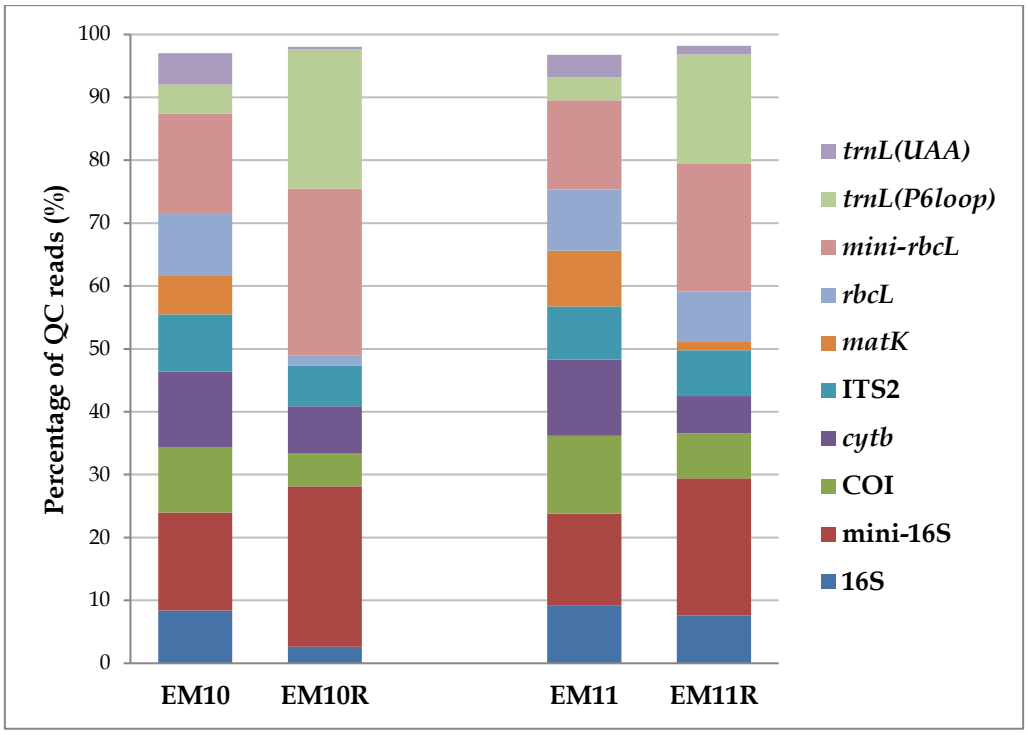


Figure 3.2 The percentage of QC reads assigned to DNA barcodes for samples EM10, EM10R, EM11 and EM11R of the pilot study.

Based on the results obtained from the pilot study, precautions were taken when grinding the freeze-dried materials and subsequent mixing to avoid cross-contamination during the laboratory handling of samples, which were used to improve the SOP for the interlaboratory trial (dx.doi.org/10.17504/protocols.io.ixbcfin). Also, control species were added to experimental mixtures that were prepared for the inter-laboratory trial to allow better confirmation of sample homogeneity and to verify that no cross-contamination had occurred during sample preparation.

3.3.4 Assessment of interlaboratory reproducibility of the DNA metabarcoding procedure

Altogether 16 laboratories from 11 countries (all experienced, well-equipped and proficient in advanced molecular analysis work), including two of the method developers, participated in the inter-laboratory trial (Table 3.6). The laboratories received ten anonymously labelled samples, each consisting of 250 mg powdered material. Two of the samples, labelled S3 and S8, were authentic TM products seized by the Dutch Customs Laboratory while the other eight samples were well-characterized mixtures of specimens from carefully identified taxa in relative dry weight concentrations from 1% to 47% (Table 3.7). In all experimental mixtures, 1% of *Zea mays* was added as quality control for homogeneity, which was confirmed with maize-specific *hmg* (high-mobility group gene) qPCR [186]. Also, tests performed with species-specific qPCR assays indicated that cross-contamination did not occur during sample preparation (Additional file 3A; Table S6). The qPCR assay for the detection of *Brassica napus*, however, also gave a positive signal for other *Brassica* sp. in the mixtures.

Table 3.6 Laboratories participating in the interlaboratory trial.

Laboratory	City and country
Agenzia delle Dogane E dei Monopoli	Genoa, Italy
AGES	Vienna, Austria
BaseClear BV	Leiden, The Netherlands
Biolytix AG	Witterswil, Switzerland
CREA-SCS sede di Tavazzano - Laboratorio	Tavazzano, Italy
Crop Research Institute	Prague, Czech Republic
Dutch Customs Laboratory	Amsterdam, The Netherlands
Eurofins GeneScan GmbH	Freiburg, Germany
Fera	Sand Hutton, United Kingdom
Generalzolldirektion	Hamburg, Germany
Laboratoire de Montpellier	Montpellier, France
Laboratorium Douane Accijnzen	Leuven, Belgium
LGC	Middlesex, United Kingdom
Livsmedelsverket	Uppsala, Sweden
RIKILT Wageningen University & Research	Wageningen, The Netherlands
U.S. Customs and Border Protection Laboratory	Newark, USA

Together with the sample materials, reagents for DNA extraction, and the complete set of barcode primers, the participants received an obligatory SOP. Any deviations from the SOP had to be reported. The participants were instructed to extract DNA, perform PCR using the barcode primers, purify the amplified DNA by removal of unincorporated primers and primer dimers, and assess the quality and quantity of the amplification products by gel electrophoresis and UV-spectrophotometry. The purified PCR products were then collected by the coordinator of the trial (RIKILT Wageningen University & Research, the Netherlands) and shipped to a sequencing laboratory (BaseClear, the Netherlands) for Illumina sequencing using MiSeq PE300 technology. The sequencing laboratory performed Index PCR and Illumina library preparation prior to MiSeq sequencing as specified in the Illumina 16S metagenomics sequencing library preparation guide. The altogether 160 PCR samples were sequenced using two Illumina flow cells with MiSeq reagent kit v3.

The interlaboratory trial should ideally have included the use of the online version of the pipeline, but unfortunately this was not possible due to shortage of time. Therefore, a single (developer) laboratory performed these bioinformatics analyses. The 160 individual samples contained on average 269,057 raw reads, and more than 150,000 reads per sample in 95% of the samples (Additional file 3A; Table S7). One sample contained less than 100,000 reads (51,750), which was considered more than sufficient for reliable species identification. After pre-processing, the samples contained on average 142,938 (pseudo)reads. On average 94.66% of the reads (min = 88.12%, max = 98.02%) passed the quality filtering indicating that the overall quality of the sequence data was consistently high across the 160 datasets.

Table 3.7 Interlaboratory trial study: Composition of the complex mixtures, and taxa identified using the default setting of the CITESspeciesDetect pipeline.

		Homogenized mixtures							
Species/Genus	**Common name**	**S1**	**S2**	**S4**	**S5**	**S6**	**S7**	**S9**	**S10**
Zea mays	Maize	1% (13) Poaceae	1% (14) Poaceae	1% (14) Poaceae	1% (15) Poaceae	1% (16) Poaceae	1% (15) Poaceae	1% (15) Poaceae	1% (14) Poaceae
Glycine max	Soy bean	1% (16) *Glycine* sp.							
Gossypium hirsutum	Cotton		1% (16) *Gossypium* sp.						
Brassica napus	Canola			1% (16) *Brassica* sp.					
Triticum aestivum	Wheat				1% (15) Poaceae				
Beta vulgaris	Sugar beet					1% (4) *Beta* sp.			
Meleagris gallopavo	Turkey						1% (16)		
Carica papaya	Papaya							1% (16)	
Solanum lycopersicum	Tomato								1% (16)
Aloe variegata *,&	Tiger aloe	1% (16) Xanthorrhoeaceae	2% (16) Xanthorrhoeaceae	3% (16) Xanthorrhoeaceae	4% (16) Xanthorrhoeaceae	1% (16) Xanthorrhoeaceae	2% (16) Xanthorrhoeaceae	3% (16) Xanthorrhoeaceae	4% (16) Xanthorrhoeaceae
Dendrobium sp. *	Dendrobium orchid	**1% (16) *Dendrobium* sp.**	**2% (16) *Dendrobium* sp.**	**3% (16) *Dendrobium* sp.**	**4% (16) *Dendrobium* sp.**	**1% (16) *Dendrobium* sp.**	**2% (16) *Dendrobium* sp.**	**3% (16) *Dendrobium* sp.**	**4% (16) *Dendrobium* sp.**
Huso dauricus *	Sturgeon/Kaluga	**1%(16)**	**2% (16)**	**3% (16)**	**4% (16)**	**1% (14)**	**2% (16)**	**3% (16)**	**4% (16)**
Crocodylus niloticus *	Nile crocodile	**1% (14)**	**2% (14)**	**3% (15)**	**4% (16)**	**1% (9)**	**2% (15)**	**3% (15)**	**4% (15)**
Lactuca sativa	Lettuce					10% (16)	10% (16)	10% (16)	10% (16)
Brassica oleracea	White cabbage	47% (16)	45% (16)	43% (16)	41% (16)	32% (16)	30% (16)	28% (16)	26% (16)
Sus scrofa	Pig					10% (16)	10% (16)	10% (16)	10% (16)
Bos taurus	Cattle	47% (16)	45% (16)	43% (16)	41% (16)	32% (16)	30% (16)	28% (16)	26% (16)
Pleuronectes platessa	European plaice					10% (16)	10% (16)	10% (16)	10% (16)

Taxa were identified at the species-level unless otherwise indicated. The number of laboratories that have identified a taxon at the species or higher level is provided in brackets. Identified taxa listed by CITES are highlighted in bold.

* *Species listed by CITES*

& *Aloe variegata (synonym Gonialoe variegata) was recently assigned to the genus Gonialoe [190].*

OTU-clustering at 98% sequence similarity on average assigned 78.14% of the pre-processed and DNA barcode assigned reads into OTUs (Additional file 3A; Table S7). Only two samples, both from the same laboratory, had a slightly lower percentage of the (pseudo-)reads assigned to OTUs (66.02% and 66.05%). This indicates that the pipeline correctly removed PCR artefacts in the clustering phase.

For taxonomy prediction, an OTU would be assigned to a database hit if they aligned with ≥ 98% sequence identity and ≥ 90% query coverage, and yielded an expect value (E-value) of at least 0.001. The BLAST output of the NGS data was interpreted by participants according to the guidelines in the SOP. Variation was observed among laboratories in interpreting the BLAST output: some laboratories consistently scored the top hits, irrespective of bitscore, while other labs selected all hits belonging to the top three bitscores, or interpreted only the first OTU of each DNA barcode, leading to large differences in identified taxa. Because of these inconsistencies, the BLAST results were re-interpreted by RIKILT Wageningen University & Research following the established guideline as mentioned in the SOP. These re-interpreted data are the data referred to in the following sections.

With one exception, all taxa mixed in at ≥ 1% (dry mass: dry mass) were reproducibly identified by at least 13 (81%) laboratories (Table 3.7). *Beta vulgaris* in sample S6 could only be identified by 4 out of 16 (25%) laboratories. *Beta vulgaris* specific sequences were present in all remaining datasets, but at very low read counts. So these clusters did not fulfil the 0.2% OTU abundance threshold (Additional file 3B; Table S2) . In order to provide insight into what alternative setting of the CITESspeciesDetect pipeline may have been better suited for identifying *Beta vulgaris*, three data sets with relatively low (S6 – laboratory 13), medium (S6 – laboratory 14) and high (S6 – laboratory 6) data volumes were reanalysed using a range of different settings for the OTU minimum cluster size and OTU abundance threshold (Additional file 3B; Table S3-S5). Setting the OTU minimum cluster size to 2, 4, or 6 has no effect on taxon identification, and *Beta vulgaris* is not identified at the species or higher taxonomic level in the data sets of laboratories 6 and 13. Setting the OTU abundance threshold to zero allows identifying *Beta vulgaris* in all three samples, but at the expense of many false positive identifications. Applying an OTU abundance threshold of 0.1% (default is 0.2%) allows identifying *Beta vulgaris* at the species or genus level irrespective of any differences in data volume between the three samples.

All six animal species could be identified to species level with at least one barcode marker (COI), while only four of the 12 plant species (*Brassica oleracea*, *Carica papaya*, *Gossypium hirsutum*, and *L. sativa*) could be identified to species level (Additional file 3B; Table S6). All other plant species were identified at the genus or higher level. For plants, no single barcode marker was best, and the most reliable data were obtained by combining the plant barcodes.

Three taxa that were misidentified or not intentionally included in the mixtures were reproducibly identified across all laboratories. *Acipenser schrenckii* co-occurred in all samples containing *Huso dauricus*. We have confirmed with DNA metabarcoding that the caviar used for preparing the experimental mixtures contains both *H. dauricus* and *A. schrenckii* (results not shown). Furthermore, *Brassica rapa* was identified by ITS2 in sample S4 by all 16 (100%) laboratories, instead of *Brassica napus*. We confirmed by Sanger sequencing *rbcL* and *matK* that our reference specimen is indeed *Brassica napus,* but that its ITS2 sequence is identical to *Brassica rapa* (LT718651). Finally, a taxon of the plant family Phellinaceae was reproducibly identified (by all laboratories) using the mini-*rbcL* marker in all samples containing *L. sativa* (S6, S7, S9, S10). Species of the family Phellinaceae and *L. sativa* both belong to the order Asterales. The evidence for Phellinaceae was not strong, i.e. the family-level identification was based on a single NCBI reference sequence only (GenBank: X69748). We therefore suspect a misidentification during the interpretation of the BLAST results.

Taxa that were identified to be the result of possible contaminations were scarcely observed, i.e. these were found in isolated cases and could possibly be explained by cross-sample contamination that may have occurred during any step of sample processing (DNA isolation, PCR, NGS library

preparation or NGS). For example, a contamination with *Gossypium* sp. was observed using *trn*L (P6 loop) in sample S1 of one of the participating labs. A total of 6 of such suspected cases of incidental cross-contaminations were observed (not shown)

Table 3.8 Sample S3 ingredients list and taxa (species, genus, family, order) identified.

Ingredients label:	**Common name**	**Species/genus**	**Family**	**(Infra)Order**
Herba Cistanches	Cistanche extract	*Cistanche* sp.	Orobanchaceae	Lamiales
Cauda cervi	Mature deer tail	*Cervus* sp.	Cervidae	Pecora
Radix Rehmanniae praeparata	Processed *Rehmannia root*	*Rehmanniae* sp.	Rehmanniaceae	Lamiales
Radix Ginseng	Dried root of *Panax ginseng*	*Panax ginseng*	Araliaceae	Apiales (8)
Radix morindae Officinalis	Morinda root	*Morinda officinalis*	Rubiaceae	Gentianales
Semen Cuscutae	Chinese dodder seed	*Cuscuta* sp. (14)	Convolvulaceae (2)	Solanales
Radix Achyranthis bidentatae	Dried root of *Achyranthis bidentatae*	*Achyranthes bidentatae*	Amaranthaceae	Caryophyllales
Rhizoma Cibotii	Root of *Cibotium barometz*	*Cibotium barometz*	Cibotiaceae	Cyatheales
Semen Platycladi	Dry ripe kernel of *Platycladus orientalis*	*Platycladus orientalis*	Cupressaceae	Cupressales
Cortex Eucommiae	Bark of *Eucommia ulmoides*	*Eucommia ulmoides*	Eucommiaceae	Garryales
Radix Astragali	Astragalus root	*Astragalus danicus* (16)	Fabaceae (16)	Fabales
Fructus Schisandrae chinensis	Chinese magnolia-vine fruit	*Schisandra chinensis*	Schisandraceae	Austrobaileyales
Cortex Cinnamomi	Dried inner bark of *Cinnamomum* sp.	*Cinnamomum* sp.	Lauraceae	Laurales
Cornu Cervi Pantotrichum	Antler of *Cervus* sp.	*Cervus* sp.	Cervidae	Pecora
Undeclared identified taxa *		*Bos taurus* (16) *Cullen* sp. (16) *Melilotus officinalis* (15) *Medicago* sp. (16) *Bupleurum* sp. (15) *Aspergillus fumigatus* (15) *Rubus* sp. (15) *Fusarium* sp. (15)		

The number of laboratories that have identified a taxon is provided in brackets. Species marked in grey are listed by CITES.
** Species identified by at least 14 laboratories that were not mentioned on ingredients list*

For the authentic TMs S3 and S8, it was observed that only few labelled ingredients could reproducibly be identified (Table 3.8 and 3.9). For sample S3 (Ma pak leung sea-dog), only the listed ingredients *Cuscuta* sp. (Chinese dodder seed), and *Astragalus danicus* (Astragalus root) could be identified. For sample S8 (Cobra performance enhancer), only the listed ingredients *Epimedium* sp. (Horny goat weed; Berberidaceae), *Panax ginseng* (Korean ginseng; Araliaceae), and species of the plant families Arecaceae (*Serenoa repens*) and Rubiaceae (*Pausinystalia johimbe*) could be identified. While most declared taxa were not identified, many non-declared taxa were identified. For sample S3, the animal species *B. taurus*, and the plants *Cullen* sp. (Fabaceae), *Melilotus officinalis* (Fabaceae), *Medicago* sp. (Fabaceae), *Bupleurum* sp. (Apiaceae), and *Rubus* sp. (Rosaceae) were identified by at least 14 (88%) laboratories (Table 3.8). Furthermore, the fungi *Aspergillus fumigatus* (Aspergillaceae) and *Fusarium* sp. (Nectriaceae) were reproducibly identified, of which the former is also a known human pathogenic fungus. For sample S8, the animal species *B. taurus* and *Homo sapiens*, the plant species *Sanguisorba officinalis* and *Eleutherococcus sessiliflorus*, and members of the plant genera *Croton* and *Erythroxylum*, and families Meliaceae and Asteraceae, were reproducibly identified (Table 3.9).

Table 3.9 Sample S8 ingredients list and taxa (species, genus, family, order) identified.

Ingredients label:	Common name	Species/genus	Family	(Infra)Order
Kola nut	Fruit of kola nut	*Cola* sp.	Malvaceae	Malvales
Siberian ginseng	Siberian ginseng	*Eleutherococcus senticosus*	Araliaceae	Apiales
horny goat weed	Horny goat weed	*Epimedium* sp. (16)	Berberidaceae (16)	Ranunculales
Catuaba	Catuaba bark	*Calophyllum antillanum*	Calophyllaceae	Malpighiales
Muria puama	Marapuama, potency wood	*Ptychopetalum* sp.	Olacaceae	Santalales
Korean ginseng	Korean ginseng	*Panax ginseng* (16)	Araliaceae (16)	Apiales
Damiana	Damiana leaves	*Turnera diffusa*	Passifloraceae	Malpighiales
Saw palmetto	Extract of fruit the of *Serenoa repens*	*Serenoa repens*	Arecaceae (16)	Arecales
Yohimbe	Extract from the bark of *Pausinystalia johimbe*	*Pausinystalia johimbe*	Rubiaceae (16)	Gentianales
Magnesium stearate				
		Bos taurus (16) *Homo sapiens* (15) *Eleutherococcus sessiliflorus* (16) *Croton* sp. (16) *Erythroxylum* sp. (15) *Sanguisorba officinalis* (15)		
			Asteraceae (16) Meliaceae (16)	

The number of laboratories that have identified a taxon is provided in brackets. Species marked in grey are listed by CITES. Species identified by at least 14 laboratories that were not mentioned on ingredients list.*

3.4 Discussion

In this study, a DNA metabarcoding method was developed using a multi-locus panel of DNA barcodes for the identification of CITES protected species in highly complex products such as TMs. As a first step, a CTAB DNA isolation method was selected for efficiently extracting high quality DNA from pure plant and animal reference materials as well as from complex mixtures. DNA isolation can be very difficult to standardise and optimise because of the complexity and diversity of wild life forensic samples, and a more systematic comparison of different DNA extraction methods is required. Secondly, a single PCR protocol, suitable for all the barcodes included, i.e. multiple universal plant and animal barcode and mini-barcode markers, was identified. This facilitated the design of a multi-locus panel of DNA barcodes. Furthermore, the developed DNA metabarcoding method includes a dedicated bioinformatics workflow, named CITESspeciesDetect that was specifically developed for the analysis of Illumina paired-end reads. The developed pipeline requires skilled experts in bioinformatics, and applies scripts for command-line processing. NGS data analysis pipelines may provide a lot of flexibility to the user, as modifications are easily implemented by expert users. The design of the pipeline prevented cyt *b* and COI full-length barcodes to be separated from their corresponding mini-barcodes, as they have identical forward primers. Since, the 300 PE reads can read through the cyt *b* and COI mini-barcodes, and therefore contain both 5'primer and 3'primer information, separation should be feasible.

To simplify the inter-laboratory validation of the pipeline, a user-friendly and intuitive web-interface with associated "Help" functions and "FAQs" was developed for the CITESspeciesDetect pipeline. The web interface was, however, not available in the course of the interlaboratory trial. Therefore, the sequence data generated in the interlaboratory study could not be analysed by the individual laboratories using the CITESspeciesDetect pipeline. A single (developer) laboratory therefore performed these analyses. Upon the availability of the online web-interface, individual participants were later given the opportunity to reanalyse their DNA metabarcoding data.

Observations made in this part demonstrated concordance of results with those obtained by the developing laboratory, reinforcing the perception of CITESspeciesDetect as a user-friendly and reliable pipeline that may readily be used by enforcement agencies and other laboratories.

The performance of the DNA metabarcoding method was assessed in an interlaboratory trial in which the method was found to be highly reproducible across laboratories, and sensitive enough to identify species present at 1% dry weight content in experimental samples containing up to 11 different species as ingredients. However, not all laboratories could identify all specified ingredients (species) in the analysed experimental samples. From the current study, we demonstrate that diverse animal taxa could be identified at the species level, which highlights the object of the method to target a wide range of animal species. COI (full-length COI and mini-COI) was found to be the most effective DNA barcode marker for animal species identification. This is not surprising considering that COI is the standard barcode for almost all animal groups [94]. Nearly all animal species identifications were supported by multiple DNA barcodes, thereby giving strong confidence to the correctness of the animal species identifications. In contrast, plants could mainly be identified at the family level, and no single DNA barcode marker was found to provide best resolution for identifying plant taxa. Ideally, adequate plant species discrimination would require the combined use of multiple DNA barcode markers, e.g. *rbcL* + *matK* [125], but this is technically not possible due to the nature of the target samples (heavily processed) and with the current Illumina Miseq technology. For the identification of plant taxa listed by CITES, the use of DNA barcodes with relatively modest discriminatory power at the genus or higher taxonomic level can still be useful, as it is often an entire plant genus or family that is listed by CITES, rather than individual plant species. This was the case for e.g. Orchidaceae and Cactaceae in this study. Yet, for some plant species (e.g. *Aloe variegata*) the resolution provided by the used plant DNA barcodes may still be too low for unambiguous CITES identification. It is important to note that the maximum achievable Illumina NGS read length limits the taxonomic resolution of DNA barcodes that are longer than ~550 nt. This particularly limited the discriminatory power of the full-length plant barcodes *matK* and *rbcL*. The DNA metabarcoding method may therefore benefit from (currently unavailable) Illumina read lengths longer than 300 nt, or other long-read sequencing technologies. Alternatively, full-length barcodes may be resolved using an advanced bioinformatics strategy (SOAPBarcode) to assemble Illumina shotgun sequences of PCR amplicons [191]. Single barcodes in several cases failed to amplify or provide resolution. The latter is likely to be caused mainly by database incompleteness, lack of genetic variability within some loci/target sequences, and sample composition. However, combining multiple barcodes into a multi-locus metabarcoding method mitigated the problems observed for individual barcodes. A high degree of confidence in the taxonomic assignments based on the combined barcodes were therefore observed, providing for enhanced quality assurance compared to the use of single barcodes.

While the use of well-characterised experimental mixtures allowed for an assessment of the performance of the DNA metabarcoding method under ideal conditions, the amplifiable DNA content of real-life samples encountered in routine diagnostic work are often of an unpredictable and variable quality. An analysis of two authentic TM products seized by the Dutch Customs Laboratory demonstrated that only few ingredients listed on the labels could be reproducibly identified. This does not mean that the undetected species were not used as ingredients. Ingredients may have been processed in such a way that the DNA is either degraded or effectively removed. This is e.g. the case with refined oils or cooked ingredients [89]. A PCR-free targeted DNA capturing approach coupled with shotgun sequencing was recently proposed for biodiversity assessments which may potentially also be suitable for enhancing species identification in difficult wildlife forensic samples [191,192]. The quality of the sequence reference database also strongly affects the ability to correctly identify species. Without correct references that also exhibit the necessary intraspecific variation, it is not possible to match and discriminate sequence reads correctly. It is well-known that accurate DNA barcoding

depends on the use of a reference database that provides good taxonomic coverage [8,18]. The current underrepresentation of DNA barcodes from species protected by CITES and closely related species critically hampers their identification. We estimate that only 18.8% of species on the CITES list contain one or more DNA barcodes (COI for animals, and *matK* or *rbcL* for plants). This will improve as DNA barcoding campaigns continue, in particular through initiatives such as the Barcode of Wildlife Project (BWP; www.barcodeofwildlife.org). Only by expansion of the sequence reference database of endangered and illegally-traded species can DNA barcoding provide the definitiveness required in a court of law.

A noteworthy observation was that most species that were reproducibly identified did not appear on the ingredients lists on the labels of the analysed TMs. This is possibly due to mislabelling. If the identifications are correct this also indicates that consumption may pose health risks. These findings corroborate earlier reports that DNA metabarcoding may provide valuable information about the quality and safety of TMs [8,9].

3.5 Potential implications

Overall, our findings demonstrate that the multi-locus DNA metabarcoding method assessed in this study can provide reliable and detailed data on the composition of highly complex food products and supplements. This study highlights the necessity of a multi-locus DNA metabarcoding strategy for species identification in complex samples, since the use of multiple barcode markers can enable an increased resolution and quality assurance, even in heavily processed samples. The developed robust bioinformatics pipeline for Illumina data analysis with user-friendly web interface allows the method to be directly applied in various fields such as: a) food mislabelling and fraud in the food industry [48], b) environmental monitoring of species [25], and c) wildlife forensics [47]. Furthermore, the pipeline can be readily used to analyse different types of Illumina paired-end datasets, even the future Illumina datasets (read length > 300 nt). Additionally, the web interface provides an opportunity for the global audience with limited expertise in bioinformatics, to analyse their own data. It also provides the liberty to select different primer sets and customise the settings for the selected purposes. As a result, the range of potential applications of the method to identify plant and animal species is diverse, the pipeline is versatile and adjustable to the user's needs, thus providing a powerful tool for research as well as enforcement purposes.

3.6 Methods

3.6.1 Reference materials and preparation of experimental mixtures

All reference specimens were obtained from a local shop in the Netherlands or provided by the Dutch Customs Laboratory (Additional file 3A; Table S2 and Table S3). The reference specimens were taxonomically characterised to the finest possible taxonomic level. For each species, it was checked whether reference sequences were present in NCBI GenBank. For taxonomic confirmation, standard COI barcodes for all animal specimens were generated and individually Sanger sequenced, and compared against the NCBI and BOLD nucleotide database. For plant species, the DNA barcodes *rbcL* and *matK* were Sanger sequenced to confirm species identity. For a number of plant and animal species the generated barcode sequence information was deposited in the European Nucleotide Archive (ENA) under accession numbers LT009695 to LT009705, and LT718651 (Additional file 3A; Table S1).

For the initial pilot study, in which the SOP for the DNA metabarcoding approach was established and tested, 15 well-defined complex mixtures were artificially prepared (Table 3.2). These experimental mixtures were prepared with 2 to 10 taxonomically well-characterised species (Table 3.2). The ingredients were mixed based on dry weight ratio, for which individual materials were freeze-

dried for 78 hours. The lyophilized ingredients were ground using an autoclaved mortar and pestle or blender in a cleaned fume hood, and subsequently stored at -20 °C ˚C. The individual ingredients of each complex mixture were weighted and mixed thoroughly using a tumbler (Heidolph Reax 2) for 20 hours and stored at -20 ˚C until further use.

For the interlaboratory validation trial, in which the applicability and reproducibility of the DNA metabarcoding method was assessed, eight additional well-characterised mixtures were artificially prepared using the above procedure. These complex mixtures were prepared with 8 to 11 taxonomically well-characterised species present at dry weight concentrations from 1% to 47% (Table 3.7). These complex mixtures were prepared in such a way that the efficiency of homogenization and possibility of sample cross-contamination could be verified using species-specific qPCR assays. In all samples, 1% of *Zea mays* was added as quality control for homogeneity. The presence of *Z. mays* was checked after sample mixing using maize-specific *hmg* qPCR along with a positive and negative control. A unique species was added at 1% dry weight to each mixture (S1-*Glycine max,* S2-*Gossypium* sp.,S4-*Brassica napus*, S5-*Triticum aestivum,* S6-*Beta vulgaris,* S7-*Meleagris gallopavo,* S9-*Carica papaya,* S10-*Solanum lycopersicum*) (Table 3.7). Species-specific qPCR was performed in duplex (together with positive and negative controls) in all samples, to check for possible cross-contamination between samples after sample preparation. Information about the qPCR primers and probes, and qPCR procedure can be found in the Additional file 3A; Table S8-S10. In addition to the eight experimental mixtures, two TMs were included that were obtained from the Dutch Customs Laboratory: a) Ma pak leung sea-dog hard capsules (MA PAK LEUNG CO, LTD, Hong Kong), was labelled to contain among others rhizoma Cibotii (*Cibotium barometz*, CITES Appendix II), and Herba Cistanches (*Cistanche* sp., CITES Appendix II) and b) Cobra performance enhancer hard capsules (Gold caps, USA), was labelled to contain among others Siberian ginseng (*Eleutherococcus senticosus*) and Korean ginseng (*Panax ginseng*). In both TMs, the medicine powder was encapsulated in a hard-capsule shell. All capsules were opened and the powder inside the capsules were stored in air-sealed and sterilized containers. The powdered medicines were thoroughly mixed using tumbler (Heidolph Reax 2) for 20 hours and stored at -20 ˚C until further use.

3.6.2 DNA isolation method

A cetyltrimethylammonium bromide (CTAB) extraction method [187] was assessed for its ability to efficiently extract DNA from a range of plant and animal materials (SOP). In brief, the CTAB method consists of an initial step to separate polysaccharides and organic soluble molecules using a CTAB extraction buffer (1X CTAB, 1.4M NaCl, 0.1 M Tris-HCl [pH 8.0], and 20mM NA_2EDTA) and chloroform. Next, the DNA was precipitated with 96% ethanol, purified with 70% ethanol, and the obtained DNA was stored at 4 ˚C until further use. DNA was extracted from 100 mg reference materials (plant and animal), artificially made complex mixtures, and real-life samples (TMs) along with an extraction control. The concentration and purity ($OD_{260/280}$ and $OD_{260/230}$ ratios) of the obtained DNA was determined by spectrophotometer (NanoDrop 1000 instrument, Thermo Fisher Scientific Inc.). The $OD_{260/280}$ ratios between 1.7 and 2.0 were considered to indicate purity of the obtained DNA. In case the extraction control contained DNA, the DNA isolation procedure was repeated.

3.6.3 Barcode markers

Candidate universal DNA barcode and mini-barcode markers and primer sets were identified using the information provided in Staats et al. (2016) [18], supplemented with additional primer sets from literature (Table 3.1). The PCR primer sets were modified to have an additional Illumina tail sequence at 5′ end of the primers (Table 3.1).

3.6.4 PCR

A gradient PCR was performed with all PCR primer combinations using 10 ng of DNA. The tested PCR conditions programme were according to the following protocol: 95 ˚C for 15 min, five cycles of 94 ˚C for 30 s, annealing range (49-55 ˚C) for 40 s, and 72 ˚C for 60 s, followed by 35 cycles of 94 ˚C for 30 s, 54 ˚C for 40 s, and 72 ˚C for 60 s, with a final extension at 72 ˚C for 10 min. The total volume of the PCR mixture was 25 µl, which included 12.5 µl of HotStarTaq Master Mix (Qiagen), 0.5 µl of 10 µM each sense and antisense primer, 7 µl of RNase-free water (Qiagen) and 5 µl of 10 ng/µl of represented species DNA. PCR was performed in the CFX96 thermal cycler (Bio-Rad) and the amplified products from all the analysed reference specimens, artificially made complex mixtures, and real-life samples (TMs) together with the positive and negative control reactions were visualised on 1% agarose gels. If amplification was observed in the negative control, the PCR analysis was repeated. Prior to NGS library preparation, 8 µl of PCR product of each target (12 in total) per sample was pooled and mixed. Next, the pooled PCR products were purified using the QIAquick PCR purification kit (Qiagen) according to manufacturer's protocol, and the purified amplicons were visualized on 1% agarose gels for all the artificially made complex mixtures, and real-life samples (TMs).

3.6.5 Next Generation Sequencing

The pooled and purified PCR amplicons were sequenced using Illumina MiSeq paired-end 300 technology. Prior to MiSeq sequencing, Index PCR and Illumina library preparation were performed as specified in the Illumina 16S metagenomics sequencing library preparation guide (Illumina document 15044223). All the DNA barcode amplicons of each sample were treated as one sample during library preparation i.e. all DNA barcode amplicons of each sample were tagged with the addition of the same, unique identifier, or index sequence, during library preparation. The Index PCR was performed to add dual indices (multiplex identifiers) and Illumina sequencing adapters using the Nextera XT Index Kit (Illumina, FC-131-1001). The prepared Illumina libraries from each sample were quantified using the Quant-iT dsDNA broad range assay (Life Technologies). Furthermore, the normalised library pools were prepared and their concentration was quantified using KAPA library quantification kit (KAPA Biosystems) and pooled prior to MiSeq sequencing using MiSeq reagent kit v3.

3.6.6 Bioinformatics analysis

The raw demultiplexed Illumina reads with Illumina 1.8+ encoding were processed using a bioinformatics pipeline, called CITESspeciesDetect. The CITESspeciesDetect is composed of five linked tools with data analysis passing through three phases: 1) pre-processing of paired-end Illumina data involving quality trimming and filtering of reads, followed by sorting by DNA barcode, 2) OTU clustering by barcode, and 3) taxonomy prediction and CITES identification (Figure 3.1).
During preprocessing of reads, the 5' and 3' Illumina adapter sequences are trimmed using Cutadapt v1.9.1 [193] using the respective substrings TGTGTATAAGAGACAG and CTGTCTCTTATACACA. After Illumina adapter trimming, reads ≤ 10 bp are removed using Cutadapt. Then, the forward and reverse reads are merged to convert a pair into a single pseudoread containing one sequence and one set of quality score using USEARCH v8.1.1861 [189].

Next, the merged pseudo-reads, unmerged forward reads and unmerged reverse reads are processed separately during quality filtering using a sliding window method implemented in PRINSEQ [163]. During this procedure, low quality bases with Phred scores lower than 20 are trimmed from 3'-end using a window size of 15 nt and a step size of 5 nt. After PRINSEQ, reads with a minimum of 95% per base quality ≥ 20 are kept, while the remaining reads are removed using FASTX_Toolkit

v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/). Then, reads are successively selected, trimmed and sorted per DNA barcode marker using Cutadapt [193]. The following steps are followed for each DNA barcode marker separately during this procedure. First, reads containing an anchored 5′ forward primer or anchored 5′ reverse primer (or their reverse complement) are selected with a maximum error tolerance of 0.2 (=20%) and with the overlap parameter specified to 6 to ensure specific selection of reads. Also, reads ≤ 10 nt are removed. The anchored 5′ primer sequences are subsequently trimmed. Second, primer sequences that are present at the 3′ end of the selected reads are also removed. For each DNA barcode, the primer-selected and unmerged reverse reads are reverse complemented and combined with primer-selected merged and unmerged forward reads.

The following procedure is used to cluster the quality trimmed reads of each DNA barcode into OTUs using the UPARSE pipeline implemented in USEARCH [189] with the following modifications: reads are dereplicated using the derep_prefix command. Also, singleton reads and reads with minimum cluster size smaller than 4 are discarded. Representative OTUs are generated using an OTU radius of 2 (98% identity threshold) and 0.2% OTU abundance threshold with minimum barcode length per primer set. Filtering of chimeric reads is performed using the default settings of the UPARSE-REF algorithm implemented in the cluster_otus command of USEARCH.

To assign OTUs to taxonomy, standalone BLASTn megablast searches [166] of representative OTUs are performed on the National Centre for Biotechnology Information (NCBI) GenBank nucleotide database using an Expectation value (E-value) threshold of 0.001 and a maximum of 20 aligned sequences. OTUs are assigned to the database sequence to which they align, based on bit score, and having at least 98% sequence identity and minimum of 90% query coverage. To identify putative CITES-listed taxa, the taxon ID first was matched against the NCBI taxonomy database using Entrez Direct (edirect) functions (available at ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/) to retrieve scientific name (species, genus, family, order and synonym name). The scientific, synonym and/or family names are then matched against a local CITES database that is retrieved from https://speciesplus.net. The final results are presented as a tab-separated values file (TSV) containing the BLAST hit metadata (i.e. bit-score, e-value, accession numbers etc.), the scientific name, synonym name, and in case a CITES-listed taxon was found, also the CITES Appendix listing and taxonomic group (i.e. species, genus, family or order name) under which the taxon is listed by CITES.

The BLAST output was interpreted by following guidelines: first, to minimize the chance of erroneous species identifications, the same species should have at least three top hits, i.e. highest bit scores. Secondly, if multiple hits are obtained with identical quality results, but with different assigned species, or with less than three top hits with same species designation, the OTU fragment was considered to lack the discriminatory power to refer the hit to species level. In such cases, the OTU would then be downgraded to a genus-level identification. Thirdly, if multiple hits are obtained with identical quality results, but with different assigned genera, the OTU fragment lacks the discriminatory power to describe the hit to genus level. In such cases, the OTU would then be downgraded to a family-level identification. An online web-interface based application for the CITESspeciesDetect pipeline was developed which is available from http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/. The web-interface facilitates intuitive BLAST identification of species listed by speciesplus.net by highlighting species on CITES Appendix I in red. Species listed on CITES Appendix II and II are highlighted in orange and yellow, respectively.

3.6.7 Pre-validation in-house of the CITESspeciesDetect pipeline

A parameter scan was performed in order to assess the effect of software settings on the ability to identify species. This evaluation allowed for identification of important parameters and their effects on the sensitivity, specificity and robustness of the procedure. This in turn resulted in specified, recommended (default) parameters values for analysing DNA metabarcoding datasets using the

CITESspeciesDetect pipeline. The effects of the following parameters were assessed: base quality scores, error tolerance for primer selection, OTU radius, OTU abundance threshold, expect E-value and query coverage threshold, percentage identity threshold, minimum DNA barcode length and BLAST database. The parameters scan was performed on experimental mixture 11 of the pilot study (Table 3.2). This mixture was selected because of its (relatively) high sample complexity, making it the most challenging complex mixture to analyse. Furthermore, the parameter scan was limited to four barcode primer sets: full-length cytochrome-B (cyt *b*), COI mini barcode (mini-COI), *rbcL* mini barcode (mini-*rbcL)* and the full-length *rbcL* (*rbcL)* barcode.

3.6.8 Inter-laboratory validation trial: participants and method

To assess the overall performance of the developed DNA metabarcoding approach, 16 laboratories from 11 countries participated in an international inter-laboratory validation. Only laboratories that regularly perform molecular analyses and have well-equipped laboratory facilities were selected to participate (Table 3.6). The majority are governmental or semi-official institutes and are considered highly authoritative within each respective country. Participants were requested to follow the SOP (dx.doi.org/10.17504/protocols.io.ixbcfin), and were asked to document any deviations that were made. The chemicals and reagents that were provided to the laboratories were: 10 samples (eight experimental mixtures and two TMs), *B. taurus* and *L. sativa* positive control DNA, CTAB extraction and precipitation buffer, 1.2 M NaCl solution, 12 universal plant and animal barcode and mini-barcode primer sets (Table 3.1), Qiagen HotStarTaq master mix, and Qiagen PCR purification kits. All reagents and samples were provided in quantities corresponding to 2.5× the amounts required for the planned experiments. After following the SOP from DNA isolation to purification of the amplified products, all the purified samples from all the laboratories (n=160) were collected and sequenced using Illumina MiSeq paired-end 300 technology (at BaseClear, Leiden, NL). The Index PCR and Illumina library preparation was performed according to the guideline and all 160 samples were sequenced on two Illumina flow cells. After Illumina MiSeq run, the raw NGS data was processed using the default settings of the CITESspeciesDetect pipeline. BLAST outputs for the samples were distributed back to the participating laboratories for interpretation of results. The laboratories interpreted the BLAST output based on the guideline provided in the SOP.

3.7 Availability of supporting data

All the sequence data obtained from the pilot study and the international interlaboratory validation trial, the CITESspeciesDetect pipeline and access to web interface are freely available. The generated barcode sequence information for some animal and plant species were deposited in GenBank under the accession numbers LT009695 to LT009705, and LT718651 (Additional file 3A; Table S1). The Illumina PE300 MiSeq data obtained from the pilot study and the international interlaboratory validation trial (n=177) were deposited to ENA with study ID PRJEB18620. The script for the CITESspeciesDetect pipeline is available at GitHub. The web interface for CITESspeciesDetect pipeline can be accessed via the following link: http://decathlon-fp7.citespipe-wur.surf-hosted.nl:8080/. The access to analysis via the web interface will be provided on request.

3.8 Availability and requirements

Project name: CITESspeciesDetect

Project home page: https://github.com/RIKILT/CITESspeciesDetect

Operating system(s): Linux

Programming language: Python and Bash

Other requirements: none

License: BSD 3-Clause License

Any restrictions to use by non-academics: none

3.9 Additional files (available with the publications)

Additional file 3A: Table S1 Accession numbers of DNA barcode sequences of plant and animal species. **Table S2** PCR success rate for animal reference species. **Table S3** PCR success rate for plant reference species. **Table S4** Statistics of different quality filtering settings for four DNA barcodes. **Table S5** BLAST identification of species with different quality filtering settings for four DNA barcodes. **Table S6** Results of species-specific qPCR performed on the experimental mixtures prepared for the inter-laboratory validation trial. **Table S7** Interlaboratory trial study: average number of Illumina reads per sample, the average number of (pseudo)reads that passed quality control (QC) and the percentage of QC (pseudo)reads that were assigned to DNA barcodes and Operational Taxonomic Units (OTUs). **Table S8** qPCR primer and probe information. **Table S9** qPCR reagent composition. **Table S10** qPCR thermocycling program. (*.docx).

Additional file 3B: Table S1 Pilot study: Composition of the experimental mixtures, and taxa identified using the default settings of the CITESspeciesDetect pipeline. **Table S2** Interlaboratory trial: *Beta vulgaris* observed in the sample S6 data sets generated by the 16 laboratories. **Table S3-S5** Interlaboratory trial: Assessment of the effect of different settings (OTU clusters size, OTU abundance threshold) of the CITESspeciesDetect pipeline on the identification of taxa using different data volume (low, medium and high) generated by three laboratory for S6. **Table S6** Interlaboratory trial: the taxonomic resolution provided by each DNA barcode marker for eight experimental mixtures (*.xlsx).

Additional file 3C: Table S1 ENA accession numbers of all raw NGS datasets obtained in this study (*.xlsx).

Chapter 4

The application of multi-locus DNA metabarcoding in traditional medicines

This chapter has been submitted for publication as: **Arulandhu AJ,** Staats M, Hagelaar R, Peelen T, Kok EJ.

Abstract

Traditional medicines (TMs) are globally traded and the consumer market is estimated to be $83 billion per annum. The diversity of TM matrices and poor quality of DNA extracted from highly processed TMs makes it challenging to apply standardized DNA-based procedures for ingredient analysis. In the present study, one standardized strategy was used to successfully obtain DNA from 18 TMs that were subsequently analyzed with a multi-locus DNA metabarcoding method to assess the species composition. In the analysis mini-barcodes accounted for the identification of most of the taxa in the TMs. The plant (ITS2) and animal (mini-16S) mini-barcode markers showed to allow species level identification of targets. In a few cases, full-length barcode markers, requiring higher quality DNA, proved to be critically informative at this level. The applied strategy resulted in the identification of a wide range of declared and undeclared ingredients, including endangered species (*Ursus arctos* and *Aloe* sp.). In 14 TMs less than 65% of the identified taxa matched the product label, and in two TMs none of the identified species matched the ingredients list. The current study shows that a multi-locus DNA metabarcoding approach is an informative analytical tool for species identification in TMs, including the potential identification of endangered species.

Keywords: Traditional medicines, multi-locus DNA metabarcoding, endangered species, CITES, Customs authority, DNA extraction

4.1 Introduction

The use of traditional medicines (TMs) containing naturally derived plant and/or animal compounds for therapeutic purposes is common practice in many countries. The global annual market for these products is estimated to be $83 billion [181]. TMs are typically plant-based mixtures of multiple species, sometimes supplemented with animal ingredients [9,8,18,27]. Some studies have found endangered species, such as *Ursus thibetanus* (Asiatic black bear; CITES Appendix I) and *Rauvolfia serpentina* (Indian snakeroot: CITES Appendix II), as TM ingredients in specific products [8,20]. The trade in endangered species, both legal and illegal, involves billions of dollars on a global scale [11,12,194]. To regulate the legal trade in endangered species worldwide, the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) was established in 1973 (www.cites.org). To implement CITES regulation, Custom authorities apply DNA-based or chemical identification methods to determine the biological origin of ingredients in cases where morphological identification of species is not possible [8,9,26,195-197]. This is particularly true for TMs for which the raw ingredients were highly processed and manufactured into powders, tablets or capsules [18].

DNA metabarcoding has been found to be a valuable molecular method for the authentication of animal and plant species used in, amongst others, TMs [8,27,9,181,198-200]. These studies have reported that TMs often contain undeclared plant and animal species, regularly including endangered species, and that mislabeling of ingredients is an issue [8,9,198]. Recently, Arulandhu et al. (2017) developed a multi-locus DNA metabarcoding strategy that enables reliable identification of plant and animal species in complex samples using a panel of different DNA barcode markers. The approach was successfully applied in well-characterized experimental mixtures and real-life TM samples in an inter-laboratory validation study. The success of a DNA metabarcoding approach depends on the amplifiability of the DNA obtained from samples. Previous studies have demonstrated that obtaining amplifiable DNA from TMs is a challenging due to the heavy processing of the ingredients (by physical and/or chemical treatments) and the presence of PCR inhibitory and interfering substances (e.g. protein, lipids, polyphenols, polysaccharides) [8,66,181]. Therefore, before the multi-locus metabarcoding method can confidently be applied in a routine set-up for screening, an adequate DNA extraction procedure needs to be selected to isolate good quality DNA from a wide range of TM matrices.

In this study, a systematic comparison of DNA extraction methods was made to identify an efficient procedure that can be applied across a wide range of TM samples. A number of commercially available and commonly-used DNA extraction methods (organic extraction, silica-based, and magnetic beads based) were compared for their ability to isolate DNA of sufficient quality from these matrices to allow subsequent PCR amplification. The best performing DNA extraction method was applied on TM samples that had been seized by the Customs authorities and were suspected to contain endangered species. All samples were DNA metabarcoded using the PE 300 MiSeq Illumina technology and data analysis was performed using CITESspeciesDetect (Arulandhu et al. 2017). Finally, the identified species from the DNA metabarcoding analyses were compared with the label information of the respective TMs for the identification of undeclared and endangered species.

4.2 Materials and Methods

4.2.1 Traditional Medicines (TMs) used in this study

In this study, 18 different TMs were analyzed that were either provided by the Dutch Customs laboratory or obtained from a local shop in the Netherlands (Table 2 and Additional file A: Table A.1). The TMs provided by the Dutch Customs laboratory were suspected to contain CITES listed species either based on labeling information or on other product-related intelligence. The samples were classified into 2 categories: plant-based TMs, or plant and animal-based TMs. The following TMs were

used to systematically compare DNA extraction methods: TMDW, TMGB, TMGW, TMSN, TMWW, TMYD and TMDP (Additional file A: Table A.1).

4.2.2 DNA extraction *methods*

The following DNA extraction methods were evaluated: Maxwell® 16 Tissue DNA Purification Kit [Promega], DNeasy *mericon* Food Kit [Qiagen], DNeasy plant mini kit [Qiagen], and the commonly-used cetyl trimethylammonium bromide (CTAB) DNA isolation method [187]. All commercially-available kits were used according to the manufacturer's instructions, except for the DNeasy plant mini kit where lysis buffer AP1 was replaced by a CTAB extraction buffer (20 g/L CTAB, 1.4 M NaCl, 0.1 M Tris, 20 mM Na_2 EDTA). Additionally, all 4 DNA extraction methods were tested in combination with an additional DNA purification step using the Promega-wizard® DNA Clean-up system. From each of the TMs 100 ± 10 mg was used to perform the DNA extraction, along with a separate plant and animal positive (*Zea mays, Acipenser schenckii/Huso dauricus*) and negative control (water). The yield and purity (OD260/280 and OD260/230 ratios) of the DNA was measured using Nanodrop (Nanodrop 1000 instrument, Thermo Fisher Scientific). The purified DNA was stored at -20^0 C until further use. The efficiency of the DNA extraction was also determined by performing a PCR using the plant mini-barcode marker (mini-*rbcL*), for which the following the PCR protocol was used: 95 ˚C for 15 min, five cycles of 94 ˚C for 30 s, annealing range 49.5 ˚C for 40 s, and 72 ˚C for 60 s, followed by 35 cycles of 94 ˚C for 30 s, 54 ˚C for 40 s, and 72 ˚C for 60 s, with a final extension at 72 ˚C for 10 min. The PCR was performed in an iCycler thermal cycler (Bio-Rad). After the PCR, 5 µl of each amplified DNA solution was loaded on a 1% agarose gel using the UV ChemiDoc™ XRS+ system (Bio-Rad). The PCR bands on the gel were used to determine the success of the DNA extraction methods.

4.2.3 Multi-locus DNA barcode panel: PCR and NGS

PCRs were performed using the DNA barcode markers and amplification conditions described in Arulandhu et al. (2017): cytochrome c oxidase I (COI) (648 nt), cytochrome *b* (cyt *b*) (743 nt), 16S (500-600 nt), mini-16S (250 nt), mini-COI (313 nt), mini-cyt *b* (357 nt), maturase K (*mat*K) (656-889 nt), ribulose-1,5-bisphosphate carboxylase (*rbcL*) (654 nt), $tRNA^{Leu}$ (UAA) intron sequence (*trn*L (UAA)) (767 nt), internal transcribed spacer 2 (ITS2) (160-320 nt), mini-*rbcL* (140 nt) and the *trn*L(P6-loop) (10-143 nt). For each TM sample, 8 µl of PCR product of each PCR target (12 in total) was pooled and purified using QIAquick PCR purification kit (Qiagen). The pooled and purified PCR amplicons were sequenced using Illumina MiSeq paired-end 300 technology as described by Arulandhu et al. (2017).

4.2.4 Bioinformatics analysis

The obtained raw Illumina reads from the TM samples were processed with the CITESspeciesDetect pipeline described in Arulandhu et al. (2017). The pipeline consists of the following steps: removal of Illumina adapters, merging the forward and reverse reads, quality filtering, segregation of reads based on barcodes primers and removal of the primers, the primer trimmed reads are clustered into Operational Taxonomic Units (OTUs) with a minimum cluster size of 100 [181]. Standalone megablast BLASTn search was performed for all the OTUs using the NCBI nucleotide database (downloaded on 14 July 2017) for taxonomical classification. Any alignment above 98% sequence identity, with a minimum of 90% query coverage and an E-value ≤ 0.001, was considered to be a match. The identified species were checked for CITES listing by using a local database extracted from Speciesplus.net and final results were interpreted following the guidelines described in Arulandhu et al. (2017). All raw NGS datasets were deposited in European nucleotide archive (ENA) under study accession number PRJEB25620.

4.3 Results

4.3.1 Comparison of DNA extraction *methods*

A comparison of DNA extraction methods was performed with seven TM samples (Additional file A: Table A.2). The DNA extraction protocols were assessed for their performance in terms of PCR amplification success using the mini-*rbcL* barcode marker. DNA yields prior to DNA clean-up system ranged from 1 to 222 ng/µl with A260/280 measurements between 1 and 1.9 (Additional file A: Table A.2). All DNA extracts were turbid. Applying the additional clean-up system resulted in clear DNA extracts with generally higher purity (Additional file A: Table A.2). The DNA extraction protocol with the best overall PCR success for the mini-*rbcL* barcode marker i.e. positive amplification for all TM samples, was the CTAB method in combination with the Wizard DNA clean-up system (Table 1). Using other DNA extraction methods positive amplification was observed, however, not in all the analyzed samples e.g. using DNeasy Plant Mini Kit + clean-up system a positive mini-*rbc*L amplification was observed in four out of seven TMs. In an additional experiment the reproducibility of the CTAB + clean-up system was confirmed by the successful PCR amplification in all samples (Table 1), and therefore this procedure was used in subsequent experiments.

Table 4.1 Effect of different DNA extraction methods on the PCR performance for seven TM samples.

	TMDW	TMGB	TMGW	TMSN	TMWW	TMYD	TMDP
CTAB						+	+
CTAB + clean-up	+*	+*	+*	+*	+*	+*	+*
DNeasy Plant Mini					+		+
DNeasy Plant Mini Kit + clean-up		+	+		+		+
Maxwell							+
Maxwell + clean-up	+				+	+	+
Mericon						+	+
Mericon + clean-up				+		+	+

"Clean-up" indicates that the DNA extraction method was coupled to the wizard® DNA Clean-up system. The symbol (+) indicates a positive PCR amplification using mini-rbcL and symbol (*) indicates that the results were successfully repeated in a second experiment.

4.3.2 Multi-locus DNA metabarcoding

DNA metabarcoding was performed on 18 authentic TM samples (Table 2). Twelve out of 18 TM samples listed only plant-based ingredients on the label and the remaining six TM samples listed both plant and animal-based ingredients (Table 2). Regardless of the TM plant or plant-animal classification, all 12 plant and animal DNA barcode markers described in Arulandhu et al. (2017) were used to analyze the TM samples. In total, 4.16 Gb of Illumina MiSeq sequencing data was generated for the 18 TM samples. On average, 769,985 raw reads were generated per sample (Additional file A: Table A.3). The raw datasets were processed using the default settings of the CITESspeciesDetect pipeline [198], except that the minimum cluster size after dereplication and Operational Taxonomic Units (OTU) clustering was set to 100. This was done to only keep clusters with relatively high support (read numbers), as was suggested by Ivanova et al. (2016). On average 95.74 % (min = 89.63%, max = 99.35%) of raw reads passed the QC steps, indicating that the overall quality of the Illumina data was high (Additional file A: Table A.3).

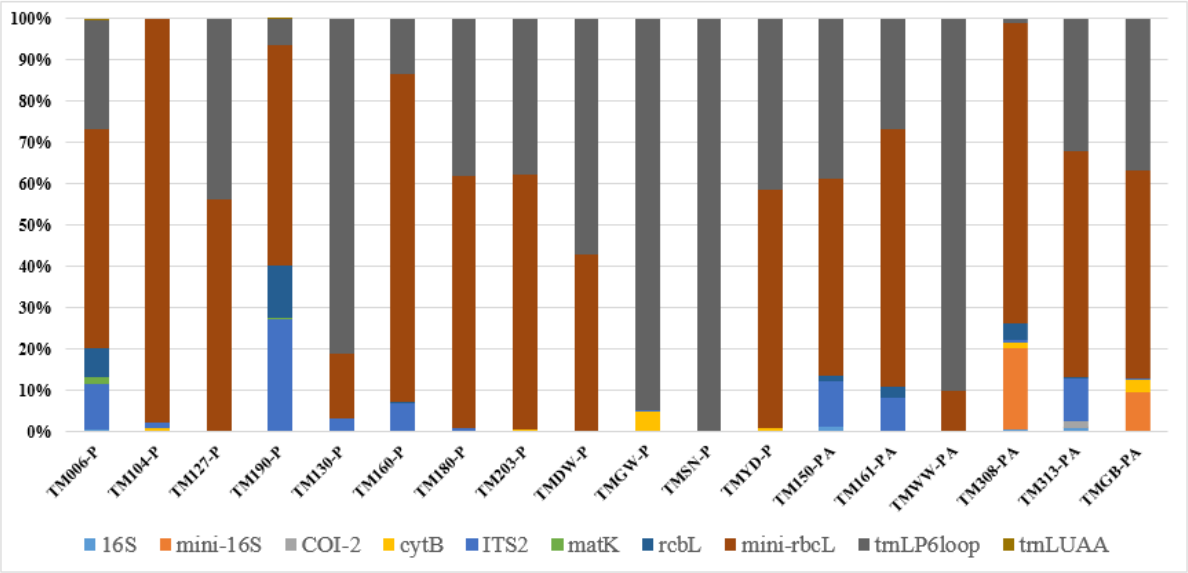


Figure 4.1: Percentage of OTU assigned reads per barcode marker. The Y-axis indicate the reads assigned to OTU per barcode marker, which is used for taxonomical identification. The symbol, P and PA indicates the plant and plant and animal based TMs analysed in the study.

For the plant-based TM samples, on average 99.3% (min = 94.97%, max = 100%) of reads were assigned to plant barcode markers (ITS2, *mat*K, *rbcL*, mini-*rbcL*, *trn*L(P6 loop) and *trn*L(UAA)) (Additional file A: Table A.4 and Figure 1). The highest percentage of reads were assigned to the plant mini-barcode markers mini-*rbcL* and *trn*L(P6loop) with on average 48.22% and 45.01% of the reads, respectively. The remaining plant-based barcode markers (ITS2, *mat*K, *rbcL*, and *trn*L(UAA)) contained on average less than 5% of the reads. The percentage of reads assigned to animal DNA barcodes was 0.7% across the 12 plant-based TM samples (Additional file A: Table A.4). For the plant and animal-based TM samples, on average 93.68% (min = 78.46%, max = 100%) of reads were assigned to plant DNA barcode markers, and on average 6.32% (min = 0.00%, max = 21.54%) of reads were assigned to animal DNA barcode markers (Additional file A; Table A.4). The majority of reads in these samples were assigned to mini-*rbcL* and/or *trn*L(P6loop), with on average 49.49% and 37.63% of the reads, respectively. For TM308 and TMGB, relatively high percentages of reads were assigned to mini-16S, namely 19.9% and 9.68% of reads, respectively (Additional file A: Table A.4 and Figure 1). The barcode markers 16S, cyt *b* and COI each contained on average less than 1% of assigned reads in plant- and animal-based TM samples. Overall, 70% of taxa were identified with a single DNA barcode (Additional file B: Table B.1-B.18). Mini-barcodes accounted for 69% of species-level identifications across the 18 TMs (Additional file B: Table B.1-B.18).

Table 4.2: Overview of CITES-listed taxa identified, the number of taxa matching the ingredients list, and the number of undeclared taxa identified in 18 TMs by DNA metabarcoding.

Sample name	Sample ID	TM classification	Target CITES-listed taxon based on labelling information.	Target CITES-listed taxon identified (yes/no)	Identified the endangered species with corresponding barcode markers	Number of species on label	Number of taxa identified matching the label	Number undeclared taxa identified.
Nongsuowan Xiangsha Liujun*	TM006	Plant-based	*Saussurea costus*	no		*9*	*1*	*9*
Kani Chitosan-Super Diet*	TM104	Plant-based	*Aloe* sp. *(excluding Aloe vera)*	*yes*	*Aloe* sp. (*mat*K)	*4*	*1*	*7*
Yaobitong Jiaonang*	TM127	Plant-based	*Cibotium barometz*	no		*6*	*1*	*5*
Shujin Huoxue Pian*	TM130	Plant-based	*Cibotium barometz*	no		*8*	*3*	*3*
One Night 8 Times*	TM150	Plant and animal based	*Hippocampus* sp.	no		*12*	*0*	*15*
Po Chai*	TM160	Plant-based	*Saussurea costus*	no		*14*	*11*	*3*
Kuku Bima TL*	TM161	Plant and animal based	*Hippocampus* sp.	no		*5*	*1*	*7*
Adutwumwaa*	TM180	Plant-based	*Aloe ferox*	no		*3*	*2*	*18*
Bu Shen Qiang	TM190	Plant-based	*Cibotium barometz*	no		*4*	*1*	*4*
Adutwumwaa BL*	TM203	Plant-based	*Aloe ferox*	no		*3*	*2*	*18*
Bear's Gall*	TM308	Plant and animal based	*Ursidae*	*yes*	*Ursus arctos* (16S and mini-16S)	*1*	*1*	*7*
Laryngitis Pills*	TM313	Plant and animal based	*Ursidae*	no		*7*	*1*	*14*
Du Huo Ji Sheng Wan	TMDW	Plant-based				*14*	*0*	*4*
Trassi Oedang	TMGB	Plant and animal based				*2*	*0*	*12*
Ge Xian Weng	TMGW	Plant-based				*2*	*2*	*5*
Seirogan	TMSN	Plant-based				*6*	*0*	*4*
Wu Ji Bai Feng Wan	TMWW	Plant and animal based				*19*	*2*	*6*
Yin Qiao Jie Du Pian	TMYD	Plant-based				*9*	*4*	*7*

**TMs provided by the Dutch Customs.*

4.3.3 Labelled and undeclared taxa identified in TMs

Overall, 31 plants and 17 animals were identified at species level, from those 19 plant and 5 animal species were unique. (Table 3). Most of the unique (11 of 19) plant species were identified using ITS2 and six plant species were identified with *mini-rbcL* (Additional file B: Table B.1-B.18). Two plant species *Saccharum hybrid cultivar* and *Zea mays* were identified with the full-length barcode marker *rbcL*. Three animal species were identified with mini-16S (i.e. *Homo sapiens*, *Ursus arctos*, *Pampus minor*), while *Sus scrofa* and *Gallus gallus* were identified with 16S and cyt *b,* respectively (Additional file B: Table B.1-B.18). *Homo sapiens* was the only species that could be identified with all animal barcode markers (16S, mini-16S, COI and cyt *b*) in all TMs. The identified taxa of a TM were compared to the ingredients list of the specific TM. For TM308 and TMGW all ingredients listed on the label could be identified (Additional file B: Table B.11 & B.15). For TM160, TM180, and TM203 more than 65% of the taxa on the ingredient list could be identified, either at genus or family level. For TM150, TMDW, TMGB, and TMSN none of the declared ingredients could be identified. For the remaining nine TMs less than 50% of the species listed on the label could be identified. Besides the declared species, many undeclared taxa were identified in all TM samples, and these were predominantly plant species (Additional file B: Table B.1-B.18). From the identified 31 plant and 17 animal species, 30 and 16 species were undeclared across the TMs, respectively. Some of the undeclared species were commonly identified, for example sugarcane (Saccharum hybrid cultivar) was found in six TMs. Additionally, for TMGB, *Manihot* sp. and one animal taxa (Shrimp) were listed on the label, however, in the analysis four fish related taxa and chicken were identified together with several plant taxa. Furthermore, in some TMs undeclared species were identified which were most likely the result of unintended contamination occurred during TM preparation or sample preparation (DNA isolation, PCR or library preparation), for instance the presence of *Homo sapiens* in plant-based TMs. In TM104 *Acipenseridae* was identified with cyt *b*, which was considered a cross-contamination during sample preprocessing, as this species was used as a positive control in during DNA extraction.

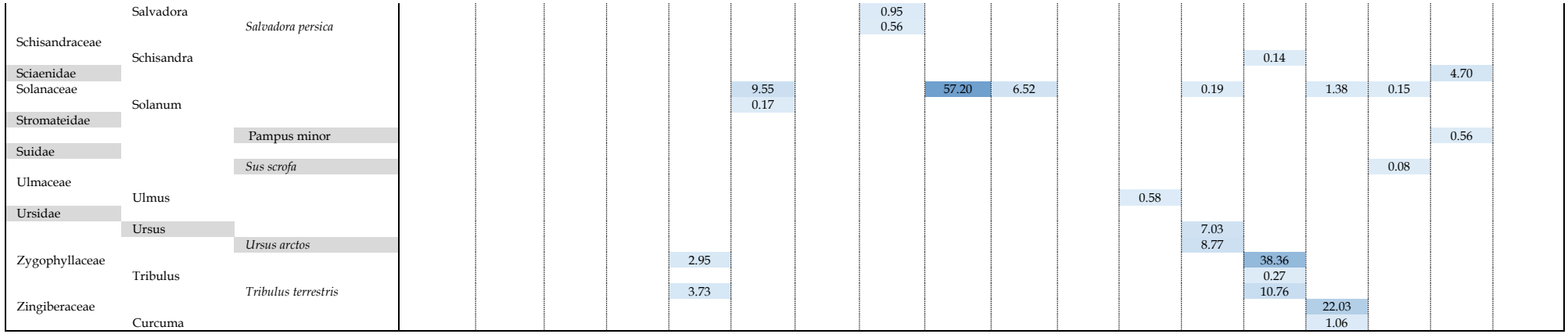
4.3.4 Endangered species identified

Of the 18 TMs analyzed in this study, 12 TMs had been seized by the Dutch Customs authority and were suspected of containing CITES-listed taxa (Table 2). These were *Saussurea costus* (CITES Appendix I), *Aloe ferox* (CITES Appendix II), *Aloe* sp. (CITES Appendix II), *Panax ginseng* (only the population of the Russian Federation in CITES Appendix II) and *Hippocampus* sp. (CITES Appendix II) and *Ursidae* sp. (CITES Appendix II) (Appendix A: Table A.1). TM127, TM130, and TM190 were suspected of containing the endangered species *Cibotium barometz* instead of Woodwardia (CITES Appendix II) based on Customs intelligence, i.e. the species name was not specified on the respective labels of these TM samples (Table 2). A survey of NCBI's nucleotide database learned that sufficient numbers of reference sequences for one or more relevant DNA barcode markers were present in the database for each of these endangered taxa (Additional file A: Table A.5). In TM308, the presence of *Ursus arctos* (Brown bear) could be confirmed with 16S and mini-16S, all with high read counts (Additional file B: Table B.11 and Table 3). TM104 contained a species of *Aloe* as identified using mat*K*. For TM006, TM150, TM160, TM161, TM180, TM203 and TM313 the suspected endangered species (*Saussurea costus, Hippocampus* sp., *Aloe ferox and Ursidae*) listed on the labels could not be identified (Table 2). Additionally, three TMs (TM127, TM130, and TM190) were seized based on the suspicion of containing *Cibotium barometz*, however, in the analysis, no *Cibotium barometz* or related genus or family could be identified with any of the plant barcode markers.

Table 4.3: Plant and animal species identified in 18 TMs by twelve universal plant and animal barcode and mini-barcode markers.

			Plant-based											Plant and animal-based					
Family	**Genus**	***Species***	**TM006**	**TM104**	**TM127**	**TM130**	**TM160**	**TM180**	**TM190**	**TM203**	**TMDW**	**TMGW**	**TMSN**	**TMYD**	**TM308**	**TM150**	**TM161**	**TM313**	**TMGB**	**TMWW**
Acanthaceae								2.72												
	Hypoestes									0.39										
Acipenseridae				0.50																
Amaranthaceae																				
	Chenopodium								2.60											
		Chenopodium album							0.53											
Amaryllidaceae																				
	Allium																1.21			
Anacardiaceae										1.15								0.14		
Apiaceae			36.15	0.61	0.10			0.34		0.18				0.48		0.09	11.18			62.54
		Angelica sinensis																0.49		
	Ligusticum		0.68																	
		Ligusticum jeholense	9.38																	
		Ligusticum acuminatum	0.05																	
		Sison amomum	0.83																	
Apocynaceae						15.69										1.27				11.30
Asparagaceae																				
	Asparagus											2.15								
Asphodelacaea				89.92				23.54		42.27										
	Aloe			0.13																
Asteraceae			43.71			64.03	8.17	2.93		0.42						1.65		0.57	10.95	
	Atractylodes		0.69				1.12													
		Atractylodes japonica	1.19																	
		Xanthium sibiricum														0.05				
	Lactuca															0.29			16.96	
Araliaceae					42.01			0.42		0.17				0.17	0.27					
	Eleutherococcus					3.22														
	Panax			4.12	53.07										0.46					
		Panax ginseng																		
Areaceae			1.09															0.35		
	Corypha		0.52																	
Arecaceae								0.34		14.53										
Aristolochiaceae																				
	Asarum																	70.77		
Bufonidae																				
	Bufo																	2.49		
Burseraceae								8.19		1.61										
Brassicaceae				0.67		10.29	0.73	0.61		1.60					0.19			0.38		
		Brassica oleraceae													0.23			0.07		
		Capsella bursa-pastoris																0.31		
	Isatis											31.30				0.11				
Campanulaceae														6.32						
Caprifoliaceae																				
	Lonicera													76.40						
Carangidae																			1.88	
Caricaceae								0.19					29.65							
Convolvulaceae									12.99							0.12				
	Ipomoea				1.94															
	Cuscuta								63.30							3.16				
Cordiaceae																				
	Cordia									1.81										
Cucurbitaceae							1.36											0.68		
Cyperus			0.07																	
Dioscoreaceae																				
	Dioscorea													0.41						

Engraulidae																					
	Setipinna																		3.13		
Euphorbiaceae																		23.53			
Fabaceae				0.42	2.02		7.37	4.34		0.77			6.09		36.92	17.06					
	Arachis							0.43		3.02											
	Astragalus				0.32											9.54					
		Astragalus danicus		0.77												3.69					
	Glycine				0.32			0.78	6.65	0.83								2.22			
	Glycyrrhiza												0.31		1.99						
	Vachellia									0.89											
	Vicia																0.16				
		Vicia sativa															0.37				
Fagaceae												2.26									
	Fagus																			9.49	
	Quercus											44.14								1.35	
Hominidae														0.08							
		Homo sapiens	0.38	0.28				0.15	13.79	0.41	0.17	5.03	0.18	0.98	5.66	1.07			2.12	0.27	
Hydrangeaceae																					
	Hydrangea				0.21																
Lamiaceae							2.08														
	Pogostemon						23.10														
Lauraceae																			1.86		
Linaceae																					
	Linum																	0.97			
		Linum bienne												0.12				1.42	0.47		
Meliaceae										0.23											
		Azadirachta indica						0.18													
Magnoliaceae							1.60														
Malvaceae															1.00						
	Gossypium											0.42								5.22	
Moraceae				2.57						0.32											
Musaceae																9.32					
Paeoniaceae																					
	Paeonia												0.12		0.60						
Pedaliaceae										0.18											
Phasianidae																					
		Gallus gallus																	0.18		
Pinaceae																					
	Larix										0.64										
Poaceae			0.24				5.63	23.84	0.13	12.88		54.41	23.78		0.56			12.83	17.72	9.84	
	Cenchrus							0.22													
	Oryza						5.94														
	Saccharum	*Saccharum hybrid cultivar*					35.95	20.51		14.81	41.99	0.09		46.27					16.35		
	Triticum							0.30										6.18			
		Sorghum bicolor						0.15													
		Zea mays										0.08		30.17					20.95		
Polygonaceae						6.77															
	Rheum														0.09						
	Rumex							0.21													
Rosaceae													7.89		0.90						
	Rubus														0.56						
Rubiaceae			3.32				0.19														
	Gardenia		1.68																		
	Rubia												0.23								
Ranunculaceae																					
	Coptis																	1.58			
Rutaceae															0.18						
	Zanthoxylum														0.43						
Salvadoraceae																					



Targets highlighted in grey are the animal species and related taxa. The number indicate the percentage of assigned read counts per taxon in a sample and blue colour gradient is applied to show the difference.

4.4 Discussion

In this study, the multi-locus DNA metabarcoding approach of Arulandhu et al. (2017) was used to assess the species composition of traditional medicines (TMs). The DNA metabarcoding approach makes use of twelve DNA barcode markers that have demonstrated universal applicability across a wide range of plant and animal taxa, and the use of mini-barcodes facilitates the identification of species in samples containing degraded DNA. The study by Arulandhu et al. (2017) showed that the DNA metabarcoding approach is a highly reproducible method for the identification of species in highly complex samples, with consistent results in a validation study involving 16 laboratories. In the current study the aims were to investigate labelling compliance and to assess the potential presence of endangered species in different TMs. Extracting good quality DNA from TM products is a crucial first step when using DNA-based methods. From previous studies it is known that isolating DNA from TMs is challenging and so far no single extraction method has been reported suitable for extracting DNA from all TM matrices [18,87]. Also, in cases where DNA can be extracted, it is often of insufficient quality for successful PCR [8]. To overcome the above-specified bottleneck, eight DNA extraction procedures were evaluated using seven TM samples representing a range of different matrices (sticky paste, crystal powders, fine powders, tablets, dry-balls, wet-balls). The performance of the DNA extraction methods were evaluated based on DNA purity, DNA yield, and PCR amplification success using the mini-*rbc*L barcode marker. This comparison showed that the CTAB DNA extraction method in combination with an additional DNA clean-up system provides the best overall DNA quality. The observed necessity of an additional DNA clean-up system is in line with the conclusions of previous studies, where it was found that potential inhibitory components and interfering substances derived from the samples (e.g. protein, lipids, polyphenols, polysaccharides) may hamper PCR efficiency despite high DNA yields [8,66].

DNA metabarcoding analyses revealed that the mini-barcode markers (ITS2, mini-*rbcL*, *trn*L(P6loop) and mini-16S were most informative in identifying plant and animal species in TM samples. This bias towards mini-barcode markers is most likely the reflection of the fragmentation of the DNA in the TMs due to the various treatments (e.g. cooking, high pressure, pH modification, grinding or drying). Here, ITS2 was shown to be the most informative marker to identify plants at species level. This is in line with Chen et al. (2010), who proposed that ITS2 can be a universal barcode to identify plant at species level, especially for medicinal plants. In a recent study, ITS2 barcode marker was efficiently used to identify *Veronica officinalis* and *Hypericum perforatum* from closely related species in herbal products using metabarcoding approach [199,200]. In animals, mini-16S has been shown to identify targets at species level, despite of the shorter barcode length (250 nt). Additional to species identification with mini-barcode markers, we found in a few cases full-length barcode (rbcL and 16S) providing valuable taxonomic information at species level. Although the full-length barcode markers were amplifiable in the TMs, the resolution, in this case, was limited to the maximum read length obtained in Illumnia sequencing (~300 bp). Therefore, using PacBio or MinION as an NGS technique, the full-length barcode could theoretically be sequenced completely, if the DNA quality permits so, which might provide a higher resolution for identification, especially for plants. Although, these barcodes could identify the target at species level, other barcodes also provided good resolution at genus level and facilitated the identification of additional taxa. The identification of different taxa with distinct barcode markers shows the necessity to use a multi-barcodes approach to identify the composition of complex samples.

In the assessment of the authenticity of the TM ingredients in terms of compliance with the label, it was confirmed in the present study that mislabeling of ingredients is a basic problem for TMs, as has also been reported in other studies [8,9,27,199,200]. In for four TMs ~35% of the identified species were undeclared, for another eight TMs, > 50% of the identified species did not match the labelling

information, of which five TMs none of the identified species matched the label information. This stresses the necessity to harmonize strict enforcement worldwide to correctly label biological ingredients of TMs on the label. Based on the World Health Organization TM regulation report, the main challenges lies with development and implementation of the regulation in the countries where the TMs are manufactured. As consequence, there is lack of safety assessment, quality control and knowledge about the TMs [198,201]. For example, in one-third of the analysed TMs contained *Saccharum hybrid cultivar* (sugar cane), which is a commonly used species to achieve a sweet taste in food products. However, only in TMGW, this species was listed on the label. Furthermore, in one case the undeclared species *Sus scrofa* (pig) was identified, while the endangered species *Ursus* sp. was listed on the label but not identified. Specifying an endangered species on the label might help to increase the presumed value of the product, however, as this example illustrates, the presence of the species in the TMs is not guaranteed.

The DNA metabarcoding analysis also focused on the identification of suspected endangered species. However, only in one case could the presence of an endangered species could be confirmed, namely *Ursus arctos* (CITES Appendix I) was positively identified in TM308. In the other TM samples, the suspected endangered taxon could not be confirmed. In TM104, the genus *Aloe* (CITES Appendix II) was identified using mat*K*, however *Aloe vera* is excluded in CITES, and therefore the identification of endangered species in this samples was considered inconclusive. It should be noted that from the identified endangered it is not possible to distinguish whether the species are obtained and used legally or illegally. Therefore, Customs authorities need to further investigate the origin of the identified endangered species, by requesting for an import, export or re-export permit for the specific species issued by the Management Authority of the State (www.cites.org). Furthermore, it is possible that the failure to identify certain species in these TMs might be due to the processing of the ingredients in such a way that the DNA was either degraded or effectively removed. Also, PCR amplification bias caused by variable primer-template mismatches may cause certain species to be missed [8,69,70]. Additional to the presence of endangered species, many studies have shown that TMs may contain fungal species, poisonous plants, toxic compounds and heavy metals, which can have adverse effects for human health [8,27,181,198].

All these factors raise concerns about the authenticity of TMs in general. In certain countries, such as United States, Canada, Australia, and European Union (EU) there is a regulatory framework to assess the quality and safety of TMs prior to allowing the product to enter the market [8,9,202], but in practice enforcement activities to control the authenticity and quality of products on the market seem to be limited. In many other countries, there is no established regulatory structure to assess the safety of TMs prior to their marketing [202]. However, countries following CITES convention are obliged to screen for the presence of endangered species in seized wildlife forensic samples. DNA metabarcoding could be an apt analytical tool for Customs authorities to address this issue.

4.5 Conclusion

In this study, we showed that the choice of DNA extraction method has a crucial influence on the PCR amplification success. The CTAB + clean-up system is the recommended DNA extraction method for use in DNA metabarcoding studies on TMs. We found that the DNA metabarcoding approach of Arulandhu et al. (2017) is suitable for providing valuable information about the authenticity and quality of TMs. Using this approach a wide range of plant and animal species, including endangered species, could be identified in different types of TMs. In the analysis, mini-barcodes barcode markers (ITS2, mini-*rbcL* and *trn*L(P6loop)) were accounted for the identification of most of the species in the TMs, reflecting the level of processing of the TM ingredients. Only in one TMs, the presence of endangered species could be confirmed, however, multiple undeclared species were identified across the TMs. These finding illustrate that the approach could support authorities to check the authenticity and quality

of TM products on the market, and will aid Customs authorities in the fight against the illegal use of endangered species in products such as TMs.

4.6 Appendices (available with the publication)

Additional file 4A: Table A1 Traditional Medicines (TMs) used in this study. **Table A2** Effect of extraction methods on DNA yield and purity. **Table A3** Summary of data quality obtained from TMs. **Table A4** Percentage of reads used to identify species in the TMs. **Table A5** Number of NCBI reference sequences available for TM ingredients listed by CITES.

Additional file 4B: Table B1 TM006, Nongsuowan Xiangsha Liujun Wan, ingredients and taxa (species, genus, family, and order) identified. **Table B.2** TM104, Kani Chitosan-Super Diet, ingredients and taxa (species, genus, family, and order) identified. **Table B.3** TM127, Yaobitong Jiaonang, ingredients and taxa (species, genus, family, and order) identified. **Table B.4** TM130, Shujin Huoxue Pian, ingredients and taxa (species, genus, family, and order) identified. **Table B.5** TM150, One Night 8 Times, ingredients and taxa (species, genus, family, and order) identified. **Table B.6** TM160, Po Chai, ingredients and taxa (species, genus, family, and order) identified. **Table B.7** TM161, Kuku Bima TL, ingredients and taxa (species, genus, family, and order) identified. **Table B.8** TM180, Adutwumwaa, ingredients and taxa (species, genus, family, and order) identified. **Table B.9** TM190, Bu Shen Qiang Shen Pian, ingredients and taxa (species, genus, family, and order) identified. **Table B.10** TM203, Adutwumwaa BT, ingredients and taxa (species, genus, family, and order) identified. **Table B.11** TM308, Bear's Gall, ingredients and taxa (species, genus, family, and order) identified. **Table B.12** TM313, Laryngitis Pills, ingredients and taxa (species, genus, family, and order) identified. **Table B.13** TMDW, Du Huo Ji Sheng Wan, ingredients and taxa (species, genus, family, and order) identified. **Table B.14** TMGB, Geroosterde Blachen, ingredients and taxa (species, genus, family, and order) identified. **Table B.15** TMGW, Ge Xian Weng, ingredients and taxa (species, genus, family, and order) identified. **Table B.16** TMSN, Seirogan, ingredients and taxa (species, genus, family, and order) identified. **Table B.17** TMWW, Wu Ji Bai Feng Wan, ingredients and taxa (species, genus, family, and order) identified. **Table B.18** TMYD, Yin Qiao Jie Du Pian, ingredients and taxa (species, genus, family, and order) identified.

Chapter 5

Critical review: DNA enrichment approaches to identify unauthorised genetically modified organisms (GMOs)

This chapter was published as: **Arulandhu AJ**, van Dijk JP, Dobnik D, Holst-Jensen A, Shi J, Zel J, Kok EJ. "DNA enrichment approaches to identify unauthorised genetically modified organisms (GMOs)". *Analytical and Bioanalytical Chemistry* 2016; 408(17): 4575-4593.

Abstract

With the increased global production of different genetically modified (GM) plant varieties, chances increase that unauthorised GM Organisms (UGMOs) may enter the food chain. At the same time, the detection of UGMOs is a challenging task because of the limited sequence information that will generally be available. PCR-based methods are available to detect and quantify known UGMOs in specific cases. If this approach is not feasible, DNA enrichment of the unknown adjacent sequences of known GMO elements is one way to detect the presence of UGMOs in a food or feed product. These enrichment approaches are also known as chromosome walking or gene walking (GW). In recent years, enrichment approaches have been coupled with Next Generation Sequencing (NGS) analysis and implemented in, amongst others, the medical and microbiological field. The present review will provide an overview of these approaches and an evaluation of their applicability in the identification of UGMOs in complex food or feed samples.

Keywords GMOs, UGMOs, PCR, NGS, Enrichment approaches.

5.1 Introduction

In 2014, over 28 countries were producing different increasing variety of genetically modified (GM) plants for commercial production [203]. These and other countries may have additional, large-scale field trials to test new GM plant varieties that are moving towards the world market. GM plants and derived products (food and feed) have been commercialised in many countries in the last two decades [15,39,204-206]. As a result, chances increase that unauthorised genetically modified organisms (UGMOs) may enter the market that have not been assessed for their food, feed and environmental safety. In some countries, regulations for the low-level presence (LLP) of UGMOs in food or feed products have been established in line with the Codex guideline, if the UGMO has already been approved in another country [207]. In the European Union (EU), this relates to only a limited group of UGMOs that meet a specific set of requirements. These LLP varieties are allowed to be present in feed products up to a level of 0.1% per ingredient (mass based) [208,209]. Other UGMOs are not allowed to be present in a product. As a result, new GMO identification strategies are required to identify UGMOs, and to achieve detection at a concentration of 0.1% (mass) per GMO relative to the ingredient.

Detection and identification are fundamentally different between UGMOs and authorised GMOs, at least in the EU, where producers are required to provide a detection method that fulfils strict regulations [210-212]. A validated event-specific quantitative polymerase chain reaction (qPCR) method is therefore available for every GMO that has been authorised in the EU (http://gmo-crl.jrc.ec.europa.eu/gmomethods/). This is not the case for UGMOs. Generally, there is no method available to detect UGMOs due to the lack of sequence information for both the genetic construct as well as for the flanking regions in the host genome. Moreover, reference materials for these UGMOs will not be available. As a result, it is not possible to develop methods to detect and identify UGMOs in a similar way as for authorised GMOs. Because of these reasons, there is a need to develop informative and cost-efficient approaches to detect and identify UGMOs, especially those UGMOs that have not yet been assessed for their safety.

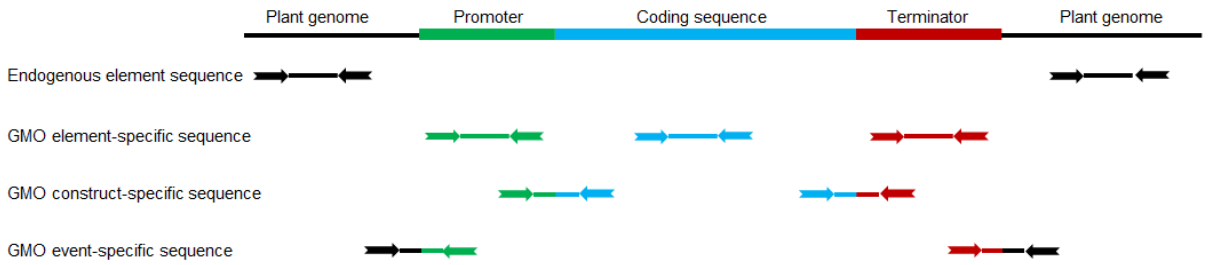


Figure 5.1 A schematic representation of an inserted GM construct in a plant genome and with available PCR-based assays to screen for GMOs and/or UGMOs. The black line indicates the plant genome and thick coloured bars (green, blue and maroon) indicates the parts of the entire GM construct. The primers and amplicons are indicated as arrows and solid lines, respectively. The colour of the arrows and solid lines corresponds to the plant genome or GM construct. Using four types of PCR-based assays simultaneously, the specificity of GMOs identification can be increased. This matrix approach is currently applied to screen for GMOs and/or UGMOs.

Currently, a matrix approach can be used to screen for UGMOs, Figure 5.1 represents the different types of PCR targets that can be used in such an approach [19,186,213-217]. By screening for the presence or absence of specific GMO elements, the presence of elements incompatible with the presence of only authorised GMOs may indicate the presence of one or multiple UGMOs. Subsequent sequencing of those unexplained GMO targets and their flanking regions might identify the UGMOs

present in the sample containing the unexplained GMO elements [216,218,219]. The obtained sequence information can be compared to the reference GMO sequences for identification [220]. Potential drawbacks of the matrix approach are the fact that it will become expensive when many GMO elements need to be tested for, and it requires a subsequent sequencing step for confirmation and identification in the case of most of the UGMOs. Technological advancements building on the matrix approach include applications of multiplexing and multi-target detection using chips [221], microarrays [222] and microfluidic arrays [205]. All these assays include a much higher number of targets than those covered by the original PCR based matrix assays. This results in a significant increase in the number of GMOs that can be detected simultaneously, including the potential to detect several UGMOs. However, none of these assays can provide information on the sequences adjacent to the detected elements. To identify the source of any unexplained GMO elements in a sample, it will still be necessary to subsequently characterise the related plant-GMO construct transition region through sequence analysis to confirm the presence of one or more UGMOs.

In complex samples the same detected element can be flanked by several adjacent motifs, each coming from a separate GMO. With the emergence of Next Generation Sequencing (NGS), it has become possible to search for multiple targets in parallel and use bioinformatics to infer and identify the adjacent sequence motifs. It is also possible to apply enrichment approaches coupled to NGS analysis to detect and identify GMOs in a complex sample, without the need for a prior screening. This review will discuss recent advances in DNA target enrichment approaches. Some of these enrichment strategies have already been coupled with NGS and applied in the medical and microbiological field [223-225]. The present review will provide an overview of such strategies and an evaluation of their applicability in the identification of UGMOs in complex food or feed samples. Coupling the most suitable NGS approach to the ideal enrichment approach may eventually allow for the identification of all GMOs, and UGMOs, in a single analysis.

5.2 DNA enrichment of the unknown sequence adjacent to a known element

Enriching the sequence adjacent to a known GMO element may lead to stronger supportive evidence of the presence of UGMOs than the re-sequencing of known amplicons. Ultimately, finding the event-specific genomic integration sites upstream or downstream of a GMO-associated element will provide conclusive evidence of the presence and identity of any GMO. The optimal enrichment approach should be able to enrich the GMO-related sequences in a complex mixture, even if the targets are present at low concentrations. Enrichment for a single element in a complex mixture could already lead to several different sequences with the same beginning. Obviously, multiplex target enrichment would even be more efficient than enriching for a single GMO element.

Multiple target enrichment approaches may be coupled to a suitable NGS approach for subsequent detection and identification of GMOs and UGMOs. An example is the SiteFinding PCR coupled with Sanger sequencing or NGS for GMO detection [219,226,227]. Several other genome walking/gene walking (GW) approaches have also been described in combination with NGS [223], though not yet adapted to the specific requirements for GMO detection, i.e. a 0.1% detection limit and multiple GMO-related targets. In all the GW approaches the enrichment starts using (a) a specific primer, e.g. targeting a known GM element, followed by amplification using a universal primer(s), or (b) a random or semi-random primer, followed by amplification using a target-specific primer, or (c) a specific primer and semi-random or adapter primer simultaneously (Table 5.1).

The different approaches can be linked to DNA pre-treatments, ranging from no pre-treatment to restriction enzyme based digestion to physical or chemical fragmentation of the DNA such as sonication or nebulization. Figure 5.2 summarizes the different enrichment approaches grouped based on the DNA pre-treatment.

Table 5.1 Available enrichment approaches for the detection of unknown sequences adjacent to a known DNA element

DNA enrichment approaches
a) Use of specific primers followed by universal primers
Long template-Rapid Amplification of Genomic DNA Ends (LT-RADE) [228]
Linear amplification-mediated PCR (LAM-PCR) [229]
Non-restrictive Linear amplification-mediated PCR (nrLAM-PCR) [225]
b) Use of semi-random primers followed by specific primers
SiteFinding-PCR [227]
c) Use of specific primer or semi-random and adapter primer simultaneously
Locus-finding PCR (LF PCR) [230]
High-throughput insertion tracking by deep sequencing (HITS) [224]
Randomly broken fragment PCR (RBF-PCR) [231]
sA-T linker adapter PCR [232]
Loop-linker PCR [232]

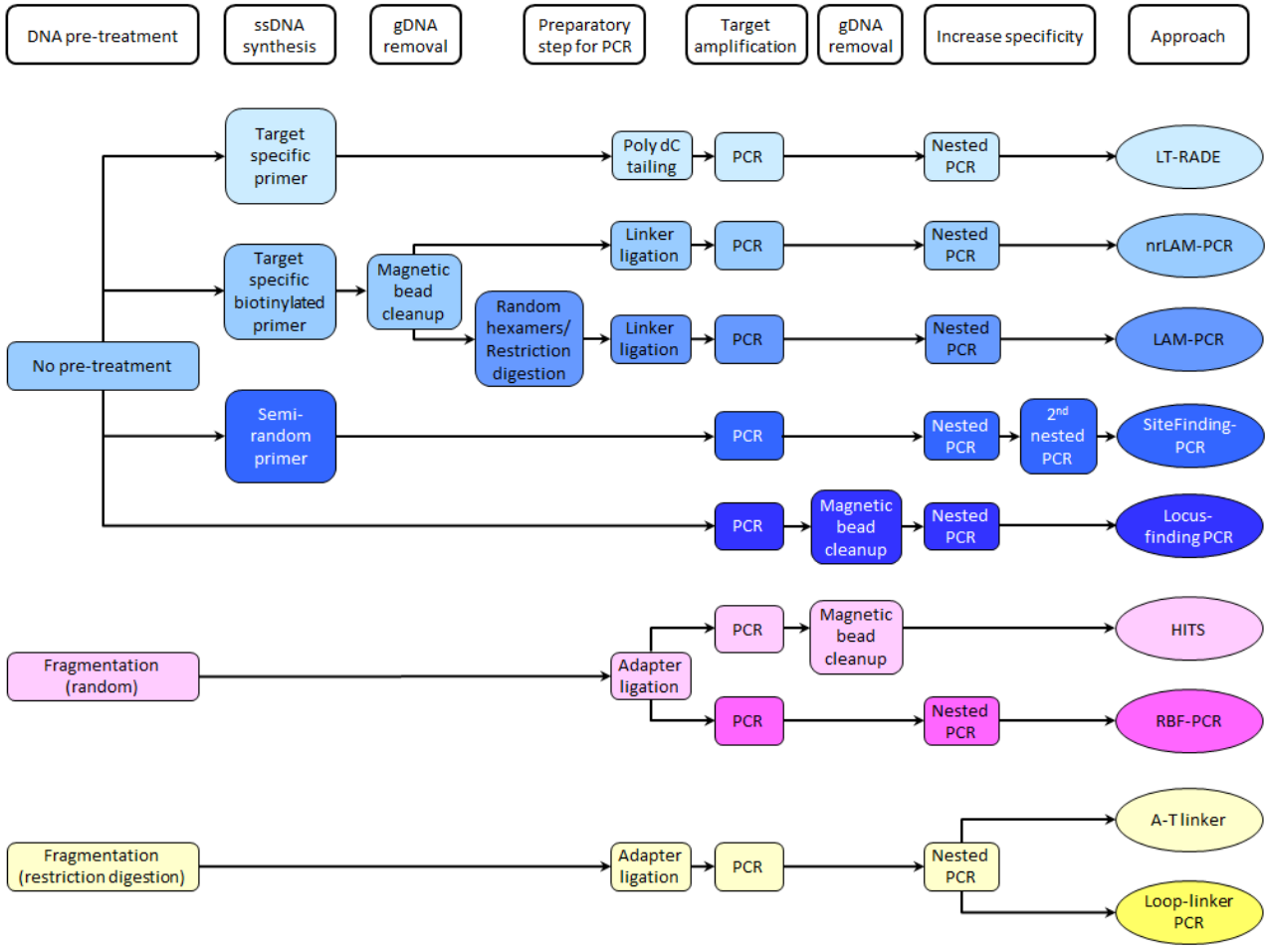


Figure 5.2 A schematic overview of different enrichment approaches grouped on the basis of DNA pre-treatment. From the left, three different DNA pre-treatments are indicated in three different colours: blue, yellow, and pink. Approaches following different DNA pre-treatments are grouped and shaded in colour corresponding to the DNA pre-treatment. Shade variation within a specific DNA pre-treatment group indicates the different enrichment approaches. To the right all the different enrichment approaches are specified in their corresponding colours.

Approaches that do not use DNA pre-treatment, require first the synthesis of a single stranded target molecule, and are hence referred to as primer extension based approaches. Primer extension can be performed in a target gene specific or semi-random way. In target primer extension based approaches, second strand synthesis and subsequent PCR can be achieved by tailing the single strand, as in LT-RADE [228], or by using random hexanucleotide as in LAM-PCR [229], or by ligation of a linker sequence, as in nrLAM-PCR [225]. In the semi-random primer extension approach, second strand synthesis and subsequent PCR make use of a target specific primer, as in SiteFinding-PCR [227], or by using specific and semi-random primers simultaneously as in LF PCR [230]. In restriction enzyme digestion approaches, the DNA is fragmented using a specific restriction enzyme to cut the DNA in a sequence specific manner. In physical or chemical fragmentation approaches, DNA is fragmented in a random manner. Both enzyme and physical fragmentation approaches require subsequent end repair procedures (tailing or linker or adapter ligation) prior to either specific PCR or direct sequencing. Loop-linker PCR [233], A-T linker [232], HITS [224], and RBF-PCR [231] come under one of the fragmentation approaches. Apart from the described approaches in the present paper, inverse PCR (I-PCR) [234,235], thermal asymmetric interlaced-PCR (TAIL-PCR) [236-240], ligation-mediated PCR (LM-PCR) [241], randomly primed PCR (RP-PCR) [236,242], vectorette PCR [243], boomerang PCR [244], TOPO vector ligation PCR [245], anchored PCR [246], cassette PCR [247] and T-Linker PCR [248] are some of the old enrichment approaches that were developed in the last two decades [39]. In fact, most of the described DNA enrichment approaches are recently developed by modifying the older enrichment approaches [15,16]. The most recent approaches relevant for the purpose of GMO detection/identification will be discussed in the following.

5.2.1 Long Template-Rapid Amplification of genomic DNA Ends (LT-RADE)

LT-RADE is a target primer extension based approach to identify the known and unknown adjacent regions of GMO elements (Figure 5.3) [228]. This approach is a modified version of previously described approaches [249,250]. In these papers RACE (Rapid Amplification of cDNA Ends) has been described, which was developed to amplify the sequences upstream and downstream of RNA transcripts after a reverse transcription reaction to convert the RNA into cDNA. RACE PCR has been applied to retrieve the sequence adjacent to the cDNA coding region in bacterial and plant RNA transcripts. When applying the RACE principle to genomic DNA, it was renamed RADE (Rapid Amplification of genomic DNA Ends) [228]. Using the non-proofreading *Taq* polymerase, RADE was first tested on the GM maize and rice events MON810 and LLRICE62 for enrichment of the right and left border of the *p35S* and *cry1Ab* for maize and *p35S* and *t35S* for rice [228]. The RADE procedure was further modified by combining the polymerases *Taq*+*Tgo* to enrich longer templates and renamed LT-RADE [228]. Single primer extension, product purification, homopolymeric tailing and nested PCR 1 and 2 are the five main steps in LT-RADE (Figure 5.3).

In the LT-RADE publication, single primer extension was performed in a PCR with a gene specific primer 1 (GSP1) for 35 cycles to obtain ssDNA reads. The reaction mixture was purified using a column purification kit and the poly-dC tailing reaction was carried out on the purified ssDNA at the 3′ end. The reaction was catalysed by template independent polymerase terminal deoxynucleotidyltransferase (TdT). The ssDNA was converted into dsDNA in a PCR reaction using the nested GSP2 primer and the universal Abridged Anchor Primer (AAP), followed by a nested PCR with the GSP3 and the Abridged Universal Amplification Primer (AUAP) to increase the specificity and the amount of final product. The obtained fragments were subsequently cloned for Sanger sequencing [228].

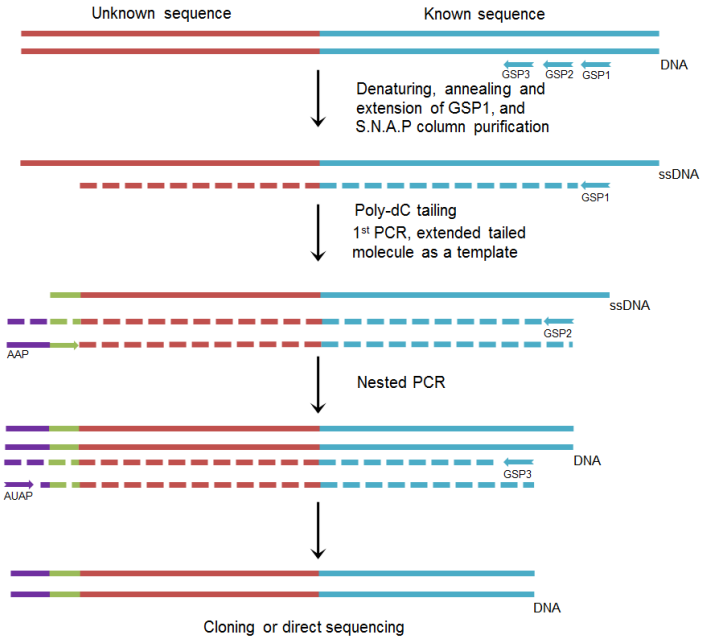


Figure 5.3 Overview of LT-RADE approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. Green solid lines indicate the poly-dC tailing. Enrichment of target sequence was performed using three gene specific primers (GSP1, GPS2, and GPS3), indicated in blue arrows, abridged anchor primer (AAP), and abridged universal amplification primer (AUAP), indicated in purple-green and purple arrows [228].

It was found that in a maize sample, the LT-RADE approach resulted in amplified fragments (1,018 bp upstream and 855 bp downstream) that were significantly longer than those obtained with the RADE protocol (577 bp and 564 bp, respectively). A rice sample was enriched only with the LT-RADE protocol and the amplified fragments were 601 bp (upstream) and 335 bp (downstream) in size [228]. The authors concluded that both approaches were simple and timesaving for the characterization of the insertion sites of GMOs and had a high specificity due to the nested PCR [228]. The authors did not discuss or provide an explanation for the observed differences in the enriched fragment lengths between the maize and rice GMOs. In a later study, LT-RADE approach was tested on three other GMOs (rapeseed, soybean and cotton) and also here the authors observed differences in the enriched fragment lengths between upstream and downstream enrichment and also between the GMOs [251]. LT-RADE has an advantage over RADE because of its ability to generate longer fragments. However, optimisation is necessary to obtain fragments >1 kb in all cases. Both approaches require large amounts of the amplified sequence for long fragment synthesis. Furthermore, a high background was observed due to nonspecific amplification. It was also observed that primers targeting GC-rich regions of the template amplified better than those targeting AT-rich templates. This is probably because the stability of G:C base pairs is higher than A:T base pairs. A limitation of the number of PCR cycles and a different template concentration at the start of the protocol was proposed to potentially reduce the competition between the abundant and less abundant target sequences and thus increase the sensitivity of the approach [228]. The LT-RADE approach has been applied in UGMO identification. The method was assessed on 100% GM plant leaf material [228,251], so the applicability of this approach in complex mixtures with some low abundance targets is still a matter of investigation. Nested PCR was used to increase the amount of targeted product. However, this also leads to a lengthy protocol that may increase the risk of contamination.

5.2.2 Nonrestrictive Linear Amplification-Mediated PCR (nrLAM-PCR)

nrLAM-PCR is one of the target primer extension based approaches for enrichment of unknown DNA sequences. nrLAM-PCR was derived from the Linear Amplification-Mediated PCR approach (LAM-PCR) [229]. In LAM-PCR publication (Figure 5.4), a gene specific biotinylated primer was used for ssDNA synthesis, followed by purification using streptavidin coated magnetic beads to capture the extended primers. Using a random hexanucleotide mixture, dsDNA was synthesised, followed by restriction enzyme digestion of the synthesised dsDNA. Fragmented DNA was ligated to linkers and a nested PCR was performed using gene specific primers (GSP1 and GSP2) and linker cassette primers (LC1 and LC2) to amplify the fragments of interest. The obtained amplicons were cloned for Sanger sequencing to identify the insertion region in the genome [229,252]. LAM-PCR was first used to characterize the retrovirus integration sites in the peripheral blood cells [229].

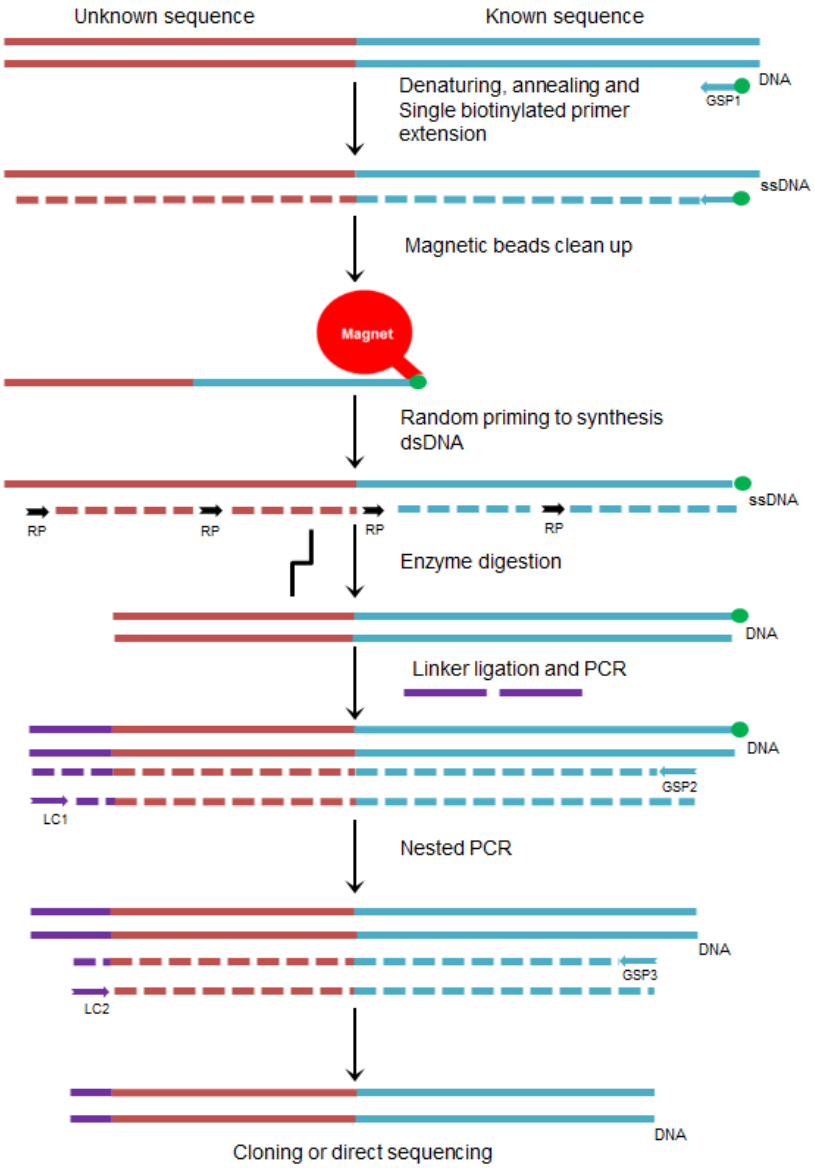


Figure 5.4 Overview of LAM-PCR approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. Red circle indicate magnetic beads clean up step, black arrows indicate random hexanucleotide, and purple solid lines indicate the linkers. Enrichment of target sequence was performed using three gene specific primers (biotinylated-GSP1, GPS2, and GPS3), indicated in green circle connected to blue arrow and blue arrows, and two linker cassette primers (LC1 and LC2), indicated in purple arrows [229].

nrLAM-PCR (Figure 5.5) is a modified non-restrictive version of the LAM-PCR [225], both approaches comprise the same steps: primer extension, magnetic beads clean-up, linked ligation, PCR and nested PCR. The nrLAM-PCR approach has been used in combination with NGS. Instead of using random hexanucleotide mixture, a ligation step was included to synthesise dsDNA. The 3'end of the ssDNA template were ligated with linkers and dsDNA was synthesised in a PCR, using the gene specific primers (GSP1, GSP2, and GSP3) and linker cassette primers (LC1 and LC2) (Figure 5.5) [223,225]. Adding a barcode to the primers was proposed for the generation of libraries for NGS [225,253]. nrLAM-PCR was performed to identify the viral vector-genome junction in the mouse SC1 embryonic fibroblast cells by coupling the approach with pyrosequencing, and most of the sequence obtained were ~250 bp [253].

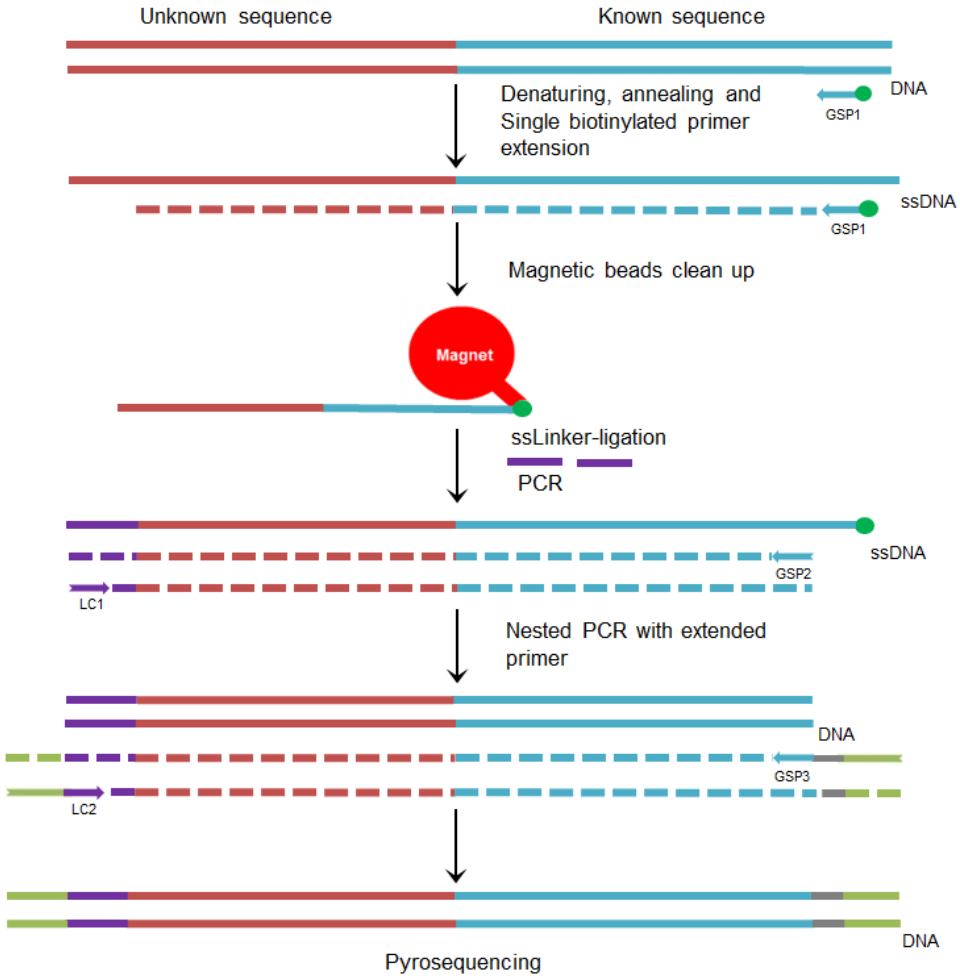


Figure 5.5 Overview of nrLAM-PCR approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. Red circle indicate magnetic beads clean up step, and purple solid lines indicate the linkers. Enrichment of target sequence was performed using three gene specific primers (biotinylated-GSP1, GPS2, and barcode-GPS3) indicated in green circle connected to blue arrow, blue arrow, and blue-gray-light green arrow. Linker cassette (LC1) and adapter linker cassette (LC2) primers are indicated by purple and light green-purple arrow [225].

LAM-PCR and nrLAM-PCR have not yet been applied to GMO and UGMO samples. The approaches may allow enrichment in a single reaction in both upstream and downstream directions. Single-stranded linker ligation to the ssDNA template is not as efficient as dsDNA ligation. This is a drawback of the nrLAM-PCR [253]. Both LAM-PCR and nrLAM-PCR products are suited for subsequent NGS analysis. Introducing barcode fusion primers would allow multiple products to be sequenced in a single NGS run.

5.2.3 SiteFinding-PCR

The SiteFinding-PCR approach (Figure 5.6) is a semi-random primer extension-based approach. This approach was used to enrich long DNA fragments [227]. The SiteFinding-PCR approach avoids ligation or tailing, restriction cleavage and complex multiple steps that could reduce the recovery. At the 5′ end SiteFinding primer 1 (SFP1) and SiteFinding primer 2 (SFP2) primer sites were linked to a random hexamer sequence and a 4-nucleotide motif common in the *Arabidopsis* genome. Three gene-specific (GSP) primers were designed from the known element sequence. After the initial denaturation of the genomic DNA the 4 nucleotides plus the hexamer hybridise to the genome in a sequence-specific manner and extension of ssDNA was initiated by *Taq* polymerase. A first PCR was performed to synthesize the dsDNA using GSP1 and SFP1 primers. Subsequently, two nested PCRs were performed using additional SFP2 and (GSP2 and GSP3) primers to gain in specificity and amount of amplified targets (Figure 5.6). The amplicons were ligated to a vector for subsequent screening and sequencing. The SiteFinding-PCR approach was first implemented to identify the known and unknown sequences adjacent to the inserted *Agrobacterium* derived T-DNA in the *Arabidopsis* genome and in the Cyanophage P4 genome [227]. The presence of the T-DNA insertion site was observed in 14 out of 15 samples of *Arabidopsis* mutants. The longest enriched fragment obtained was ~ 4.5 kb from Cyanophage and ~ 2.2 kb for *Arabidopsis* [227]. Only upstream enrichment was performed in both Cyanophage and *Arabidopsis*. The maximum enrichment in *Arabidopsis* was half of the enrichment observed in the Cyanophage. A likely explanation for this difference is a more frequent occurrence of complementary sites of the SiteFinder motif on genomic DNA in the *Arabidopsis* compared to the Cyanophage genome. This would lead to a generally shorter distance between the SiteFinder motif and the place of insertion, and hence to shorter enrichment lengths in *Arabidopsis*. As a result of using semi-random primers, multiple amplicons of different sizes could be obtained.

The SiteFinding-PCR approach was recently performed with some modifications of the SFP primers, to characterise the GM rice KMD1 and the maximum enrichment length obtained was ~ 300 bp [226]. To overcome this limitation, the SiteFinding-PCR was repeated several times with new gene-specific primers to obtain longer fragments, but this goal was not achieved. The purpose was to identify the integration site of the genetic construct, and use this information to design and test a successful event-specific qPCR assay [226]. The SiteFinding-PCR approach was applied to characterise the flanking sequence of the *vip3A* element in MIR162 maize as a model study for UGMO identification [219]. Sanger, Illumina and Pacific Biosciences (PacBio) sequencing approaches were used to analyse the obtained sequences. PacBio resulted in longer contigs both upstream and downstream (1326 bp and 1135 bp, respectively) when compared with the Illumina data (858 bp and 1038 bp, respectively). Both NGS approaches outperformed Sanger sequencing regarding the length of the newly obtained sequence information. PacBio showed lower sequence identity, 92-95%, compared to the 99% identity in Illumina data. However, the integration site of the genetic construct in MIR162 maize was not reached, due to the position of the targeted *vip3A* element, ~ 2 kb from the left border and ~ 4 kb from the right border.

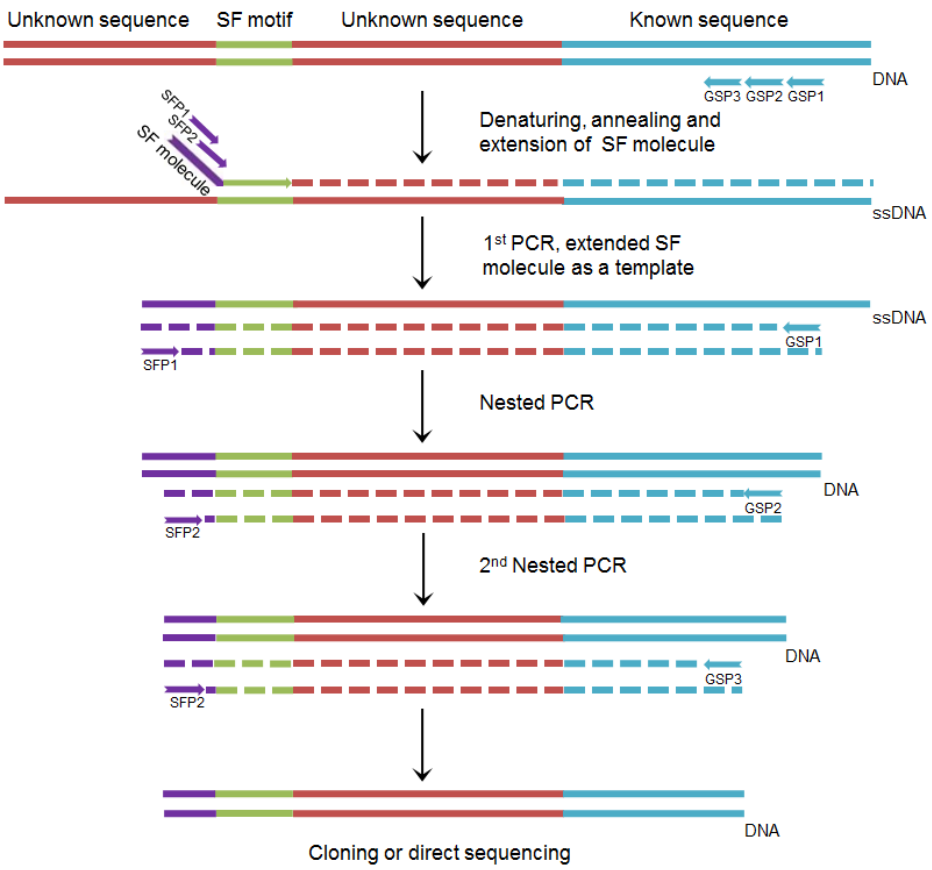


Figure 5.6 Overview of SiteFinding-PCR approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence, Maroon solid and broken lines indicate the unknown sequence. Green solid lines indicate the SiteFinder (SF) motif and the purple-green arrow indicates the SiteFinder (SF) molecule. Enrichment of target sequence was performed using three gene specific primers (GSP1, GPS2, and GPS3) indicated by blue arrows, and two SiteFinding primers (SFP1 and SFP2), indicated by purple arrows [227].

The SiteFinding-PCR approach can be used in UGMO identification based on the pervious results. However, the reported length of the amplified fragments was variable; this could be due to the species-specific frequencies of the SiteFinder motif in different genomes. In a situation where one SiteFinder motif is very near the gene of interest, it may be difficult to obtain also the sequence from a more distant second motif, since a bias for shorter fragments is expected in the subsequent PCR steps. In all of the respective studies, pure GM (100%) materials were used. Thus, the applicability of this approach in complex mixtures with some low abundance targets is yet to be demonstrated. Coupling the SiteFinding-PCR with NGS avoids the time-consuming and contamination-sensitive step of restriction digestion and ligation of the amplicons in a vector. However, the use of three PCR steps, including two nested PCRs, still leads to a rather lengthy protocol that may be prone to contamination. In the APAgene GOLD Genome Walking Kit the first two steps in the SiteFinding-PCR approach have been swapped [254]. In this approach, enrichment starts with GSP to obtain the ssDNA of the target sequence. The enrichment step was repeated separately four times in a parallel reaction with the same GSP. A first PCR reaction was performed to synthesize the dsDNA using GSP and SiteFinding molecule. This reaction was also performed four times in parallel in separate reactions with the same GSP and different SiteFinding molecules. Subsequently, two nested PCRs were performed for all the four reactions to gain in specificity and amount of amplified targets [254]. This approach was evaluated and validated on in-house made GM rice food mixtures and processed rice food [254,255].

5.2.4 Locus-finding PCR (LF PCR)

LF PCR (Figure 5.7) uses a genome walker (GWr) molecule, similar to the SiteFinder molecule, for primer extension [230]. Primer extension was performed in a PCR, combined with a target specific primer, instead of in a separate reaction prior to the PCR as is done in SiteFinding-PCR. LF PCR includes an affinity purification step to specifically capture the amplified target molecules. The final step prior to sequencing was PCR with a nested target primer and a primer that hybridises to the tail of the GWr molecule, again similar to SiteFinding-PCR. Four forward semi-random genome walker primers (GWs 1–4) were used in parallel with a construct-specific primer (CSP) to improve the chance of amplifying the desired location in the primary PCR. All four GWrs have four bases (A, T, G, and C) at their 3' ends. The authors state that they are random but do not show how the randomization was done. In fact, the four GWrs described are 4 of the 24 possible combinations of having four consecutive bases without repetition. Upstream of these four different four-nucleotide motifs are four degenerate nucleotides, and upstream of those are 18 or 19 bases reverse complementary to GW-5. This means that both the SiteFinder molecule and the GWr 1 to 4 share a similar structure of (from 5' to 3') a primer binding site, a stretch of degenerate nucleotides and a 4-nucleotide motif. This approach was successfully applied to find the transgene integration loci of 8 *Agrobacterium tumefaciens* transformed transgenic rice lines transformed with glyphosate-resistant genes, mutant *epsps* (enol pyruvyl shikimate phosphate synthase) and *gat* (glyphosate acetyl transferase [230]. The desired amplification in the rice study was obtained with the GWr-1 and CSP-1. The initial PCR was consequently performed using a GWr1 and CSP1 for all rice plants. The authors state that plants having more complex genome than rice may require optimization with the remaining GWrs. In order to select only the interesting amplicons for further processing, a biotinylated capture primer (CP) was designed based on the inserted vector region. The CP annealed to the desired amplicons and affinity purification was performed to separate the desired product from the rest of the mixture. Subsequent nested PCR was performed using GWr5 and CSP2 to increase the specificity. The efficiency of this approach was assessed by subsequent gel electrophoresis and Sanger sequencing. It was found that a maximum stretch of ~500 bp was enriched towards the left border of the insert as the experiment was performed only in one direction. To improve the length of the enrichment, high-fidelity enzymes and increased extension time (2 min) were tried in LF PCR, but failed to result in amplicons in the nested PCR. This was perhaps because PCR generally favours smaller size amplicons [230].

This approach seems very useful when the entire construct is known, or at least a considerable stretch of sequence close to one of the borders. The sample material used in the paper consisted of individual clones from a single transformation experiment. The authors retrieved three different insertion sites in chromosomes 1, 6, and 12 in 8 different clones. Seven of the clones showed two independent insertion sites per clone. These experiments relate to samples of limited complexity, no data was available on the application of LF PCR in complex samples.

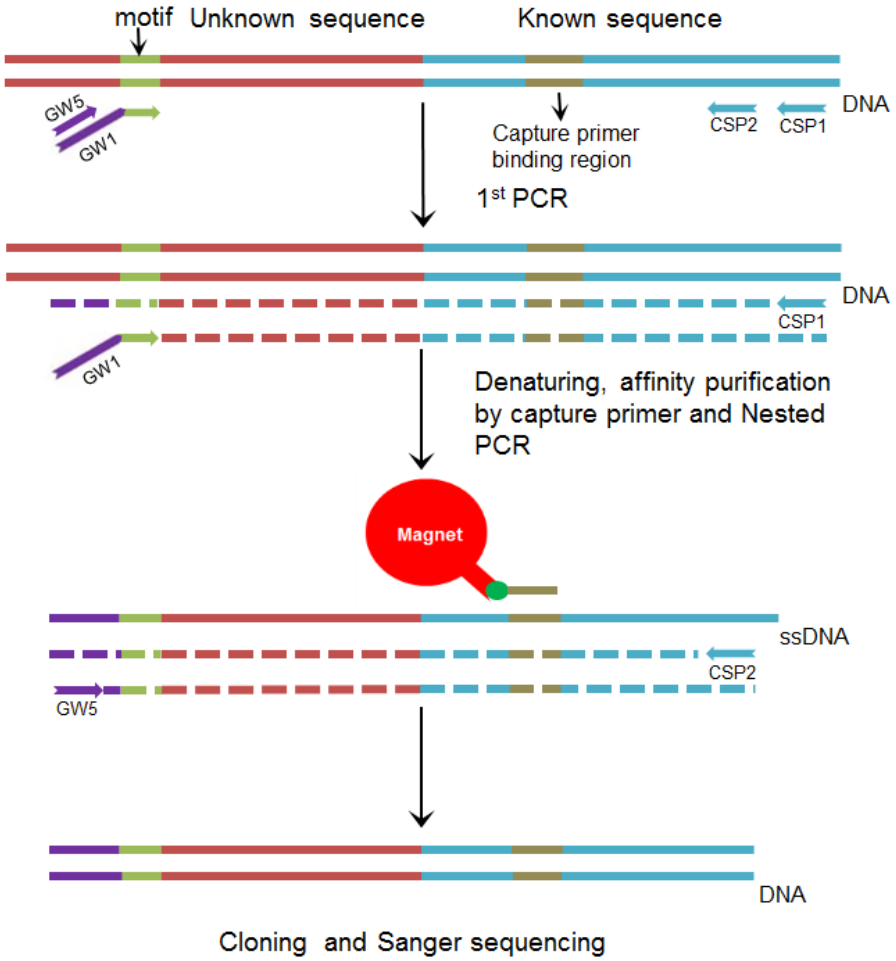


Figure 5.7 Overview of LF PCR approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. Green solid lines indicate the genome walker motif and the purple-green arrow indicate the genome walker (GW1) molecule. Red circle with extended green-brown part, indicate magnetic beads attached to biotinylated primer. Enrichment of target sequence was performed using two construct-specific primer (CSP1, and CSP2), indicated by blue arrows, and genome walker (GW5) primers, indicated by purple arrows [230].

5.3 Ultrasonic fragmentation based DNA enrichment

In ultrasonic fragmentation based DNA enrichment, the genomic DNA is randomly fragmented, repaired and ligated to an adapter for a target specific PCR. High-throughput Insertion Tracking by deep Sequencing (HITS), Randomly Broken Fragment PCR (RBF-PCR) and probe hybridisation are enrichment approaches that apply ultrasonic DNA fragmentation. The length of the fragmentation depends on the parameters and the probes used during ultrasonication.

5.3.1 High-throughput Insertion Tracking by deep Sequencing (HITS)

The HITS approach (Figure 5.8) was performed to identify the DNA junction between the integrated transposon of *Himar1-mariner* and *Haemophilus influenzae* [224]. The underlying goal was to analyse bacterial genes involved in pathogenesis, using a whole-genome transposon mutant bank in combination with NGS. DNA containing the randomly inserted transposons was fragmented using ultrasonication, and repaired and subsequently Illumina adapters were ligated to the ends. Using a biotinylated transposon-specific primer (SP) and adapter primer (LCP), transposon-genome junction

enrichment was performed. A magnetic bead-based clean up step was performed to obtain the specific sequences, prior to Illumina sequencing. NGS libraries were generated with a high coverage [224]. The same approach was applied in *Salmonella enterica* to enrich for Tn5-derived bacterial transposon insertion libraries. By adapting the Illumina approach, *piggyBac* PB insertion transposon libraries were generated in yeast and also Mutator transposon lines were identified in maize [223,256-258]. The HITS approach has not yet been applied to GMO and UGMO identification. Sequencing of transposon/chromosome junctions revealed independent insertions in nearly 56,000 genomic sites [224]. Random fragmentation may also occur in the inserted exogenous sequence, reducing the effectiveness to identify the plant-exogenous sequence junction in GMOs. Furthermore, the HITS approach combined with Illumina sequencing generated relatively short fragments, in between 200 bp – 400 bp.

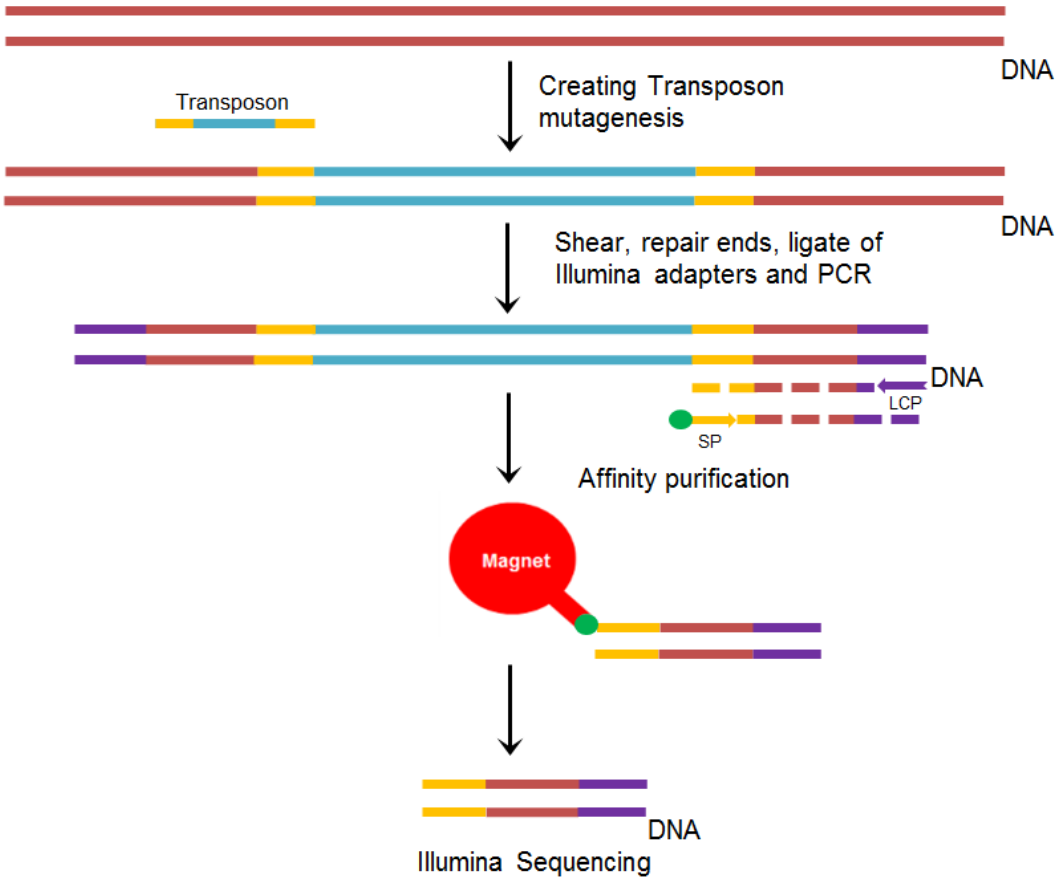


Figure 5.8 Overview of HITS approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. Enrichment of target sequence was performed using biotinylated transposon-specific primers (SP; indicated by green circle connected to yellow arrow), and Illumina adapter primer (LCP; indicated by purple arrow). Red circle indicate magnetic beads clean up step [224].

5.3.2 Randomly Broken Fragment PCR (RBF-PCR)

RBF-PCR approach (Figure 5.9) was successfully performed to identify the unknown sequence adjacent to known sequence motifs in the GMO maize LY038 [231]. In this approach, the genomic DNA was randomly fragmented by ultrasonication and repaired by addition of adenines to the 3′ end of the DNA-strands. An adapter with a T-overhang at the 3′ end was ligated to the fragmented sequences. The adaptors were designed to be not fully complementary, to avoid PCR amplification of all fragments that

contain adaptors by only the adaptor primer. Instead, by adding a primer adapter primer (AP) that has the same sequence as one of the non-complementary parts of the adaptor, PCR can only occur when the complementary strand is synthesized in the first round of PCR by elongation of the target sequence specific primer (SP1). Subsequent nested PCR was performed to increase the specificity (Figure 5.9) [231]. The RBF-PCR approach was performed with some modifications compared to HITS: a) the specific biotinylated primer was replaced by a specific primer (SP) and b) instead of Illumina NGS adapters, specifically designed APs used to avoid self-ligation (Figure 5.9).

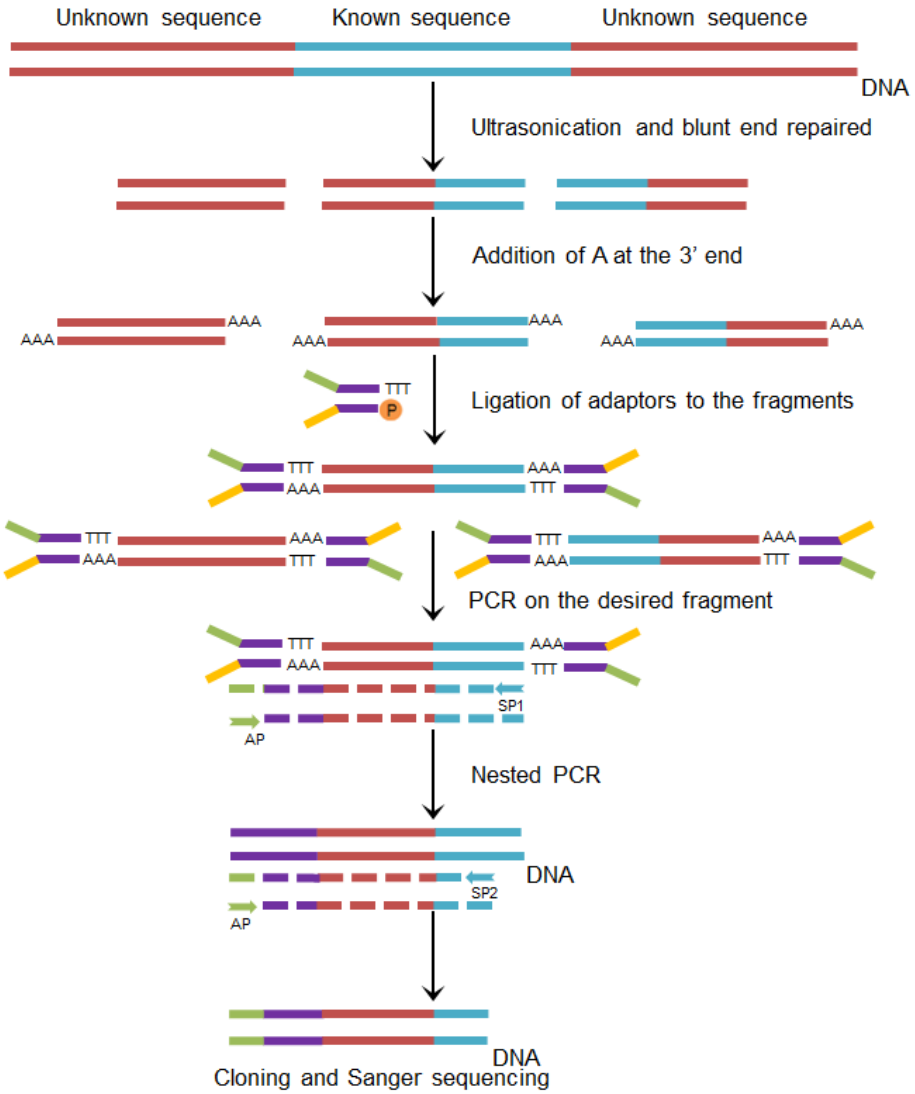


Figure 5.9 Overview of RBF-PCR approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. The modified adapters are indicated by purple-yellow and purple-green lines, and are partially complimentary to each other. Enrichment of target sequence is performed using specific primers (SP1 and SP2) indicated by blue arrows, and adapter primer (AP) is indicated by green arrow [231].

The fragmentation of the genomic DNA was shown to be a crucial step in this approach: changing the duration and frequency of ultrasonication influences the degree of fragmentation. The obtained amplicons using the RBF-PCR approach were reported to be between 500 bp and 2000 bp. After sequencing, 35 % of the obtained sequences did not match the target sequences. This implies that the applicability to mixed GMO samples is likely to be limited. In general, RBF-PCR may not be the

ideal approach because random fragmentation can occur in the inserted exogenous sequence or near to the event specific genome-insert junctions and thus hamper the identification of UGMOs.

5.3.3 Probe hybridisation approach

Recently, probe hybridisation combined with NGS was used to characterise GM insertion sites. In the probe hybridisation approach, the genomic DNA is randomly fragmented and probes are used to capture the target sequence, followed by NGS analysis to characterise the insertion sites. Contrary to all other approaches in this review, this approach does not include a PCR step. In a recent study, using this approach, three 70 nucleotides long biotinylated probes were designed to target the extremities of T-DNA sequences to rapidly locate the T-DNA insertion sites in 55 out of 64 mutant *Arabidopsis* plants [259]. A similar approach, called Southern-by-Sequencing (SbS), was used to characterise the insertion sites in GM crops [260]. In both the approaches, information of the construct close to the insertion sites is necessary for characterisation, which is typically not the case in unknown GMO identification. A further drawback is that the enrichment is possible only for a limited flanking region. Due to this, limited extra sequence information will be identified, making this approach not applicable for identification of long stretches of insert sequence of unknown GMOs.

5.4 Restriction enzyme based DNA enrichment

Using restriction enzyme, the genomic DNA is digested unevenly in a sequence specific manner. The ends of the fragmented DNA are repaired and ligated with the linker or adapter for a target-specific PCR. The obtained PCR fragments can then be used for either cloning or direct sequencing. Some of the enrichment approaches based on this principle are: (a) Classical restriction enzyme digestion followed by NGS approach, (b) A-T linker adapter PCR, (c) TOPO vector ligation PCR, and (d) Loop-linker PCR.

5.4.1 Classical restriction enzyme digestion followed by NGS approach

In this approach, after the restriction enzyme digestion linkers were ligated to the fragments, and target-specific PCR was performed using sequence specific primer (SP) and linker primer (LP). The obtained PCR fragments were then directly prepared for pyrosequencing by addition of sequencing adapters, or library were prepared by performing a second PCR using a barcode primer and linker primer (Figure 5.10) [223,261]. The combination of restriction enzyme digestion and NGS was first used to identify the insertion site of the HIV in the human genome [223,261]. Barcode primers were introduced in this approach to identify Mu transposon elements in different maize Mu-stocks [223,262]. An alternative way of restriction was performed by cutting the DNA of the transposon element with a specific enzyme that recognises the nucleotide sequence of the transposon [263]. Transposon mutants libraries of *Pseudomonas aeruginosa* have also been generated by adapting the inverse PCR approach in combination with NGS-Illumina [223,264]. Observed disadvantages of this approach were the possible amplification of non-target sequences, and less flexibility due to the use of specific restriction enzymes [265].

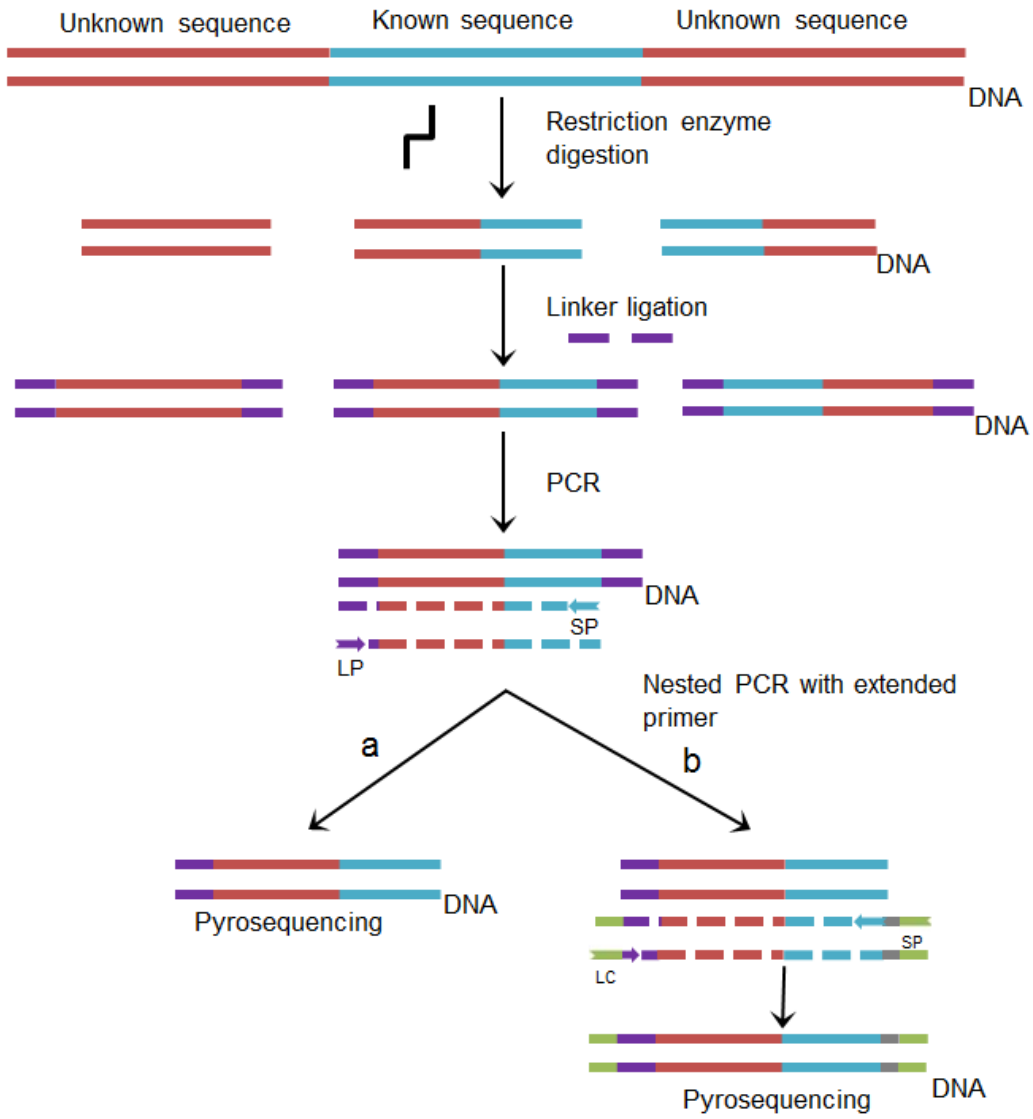


Figure 5.10 Overview of classical restriction enzyme digestion followed by NGS. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. Purple solid lines indicate the linkers. Initial enrichment of target sequence is performed using gene specific primers (SP) indicated by blue arrow and linker primer (LP) indicated by purple arrow. Amplification is either followed by (a) pyrosequencing or (b) a second PCR is performed with a barcode specific primer (SP) indicated by green-grey-blue arrow and linker primer indicated by green-purple arrow with the aim to generate an amplicon library with barcode prior to pyrosequencing [223].

5.4.2 A-T linker adapter PCR

The A-T linker adapter PCR (Figure 5.11) is a combination of Ligation-mediated (LM-PCR) and T linker PCR [232]. The A-T linker adapter was designed in such way that it binds to the A-tailed fragment with the NH_2 group, blocking elongation from the 3′ end of the short-strand primer to avoid the self-ligation of the adapter and nonspecific amplification. A-T linker adapter PCR was performed on 16 different *Arabidopsis* mutants with a T-DNA insert [232]. The template DNA was digested using 15 different restriction enzymes, yielding fragments with 5′ overhangs, a 3′ overhangs, or blunt-ends. *Taq* polymerase catalyses the A-tailing of the 5′ overhang and blunt-end DNA fragments. For the 3′ overhang fragments, a specific primer 1 (SP1) was used for extension of the target sequence (5′ to 3′ end)

followed by dA extension. The initial PCR was performed using adapter primer 1 (AP1) and SP1 primer and to increase the specificity and product yield, a nested PCR was performed (Figure 5.11). It was found that all the resulting fragments were < 1 kb, with low reproducibility [232].

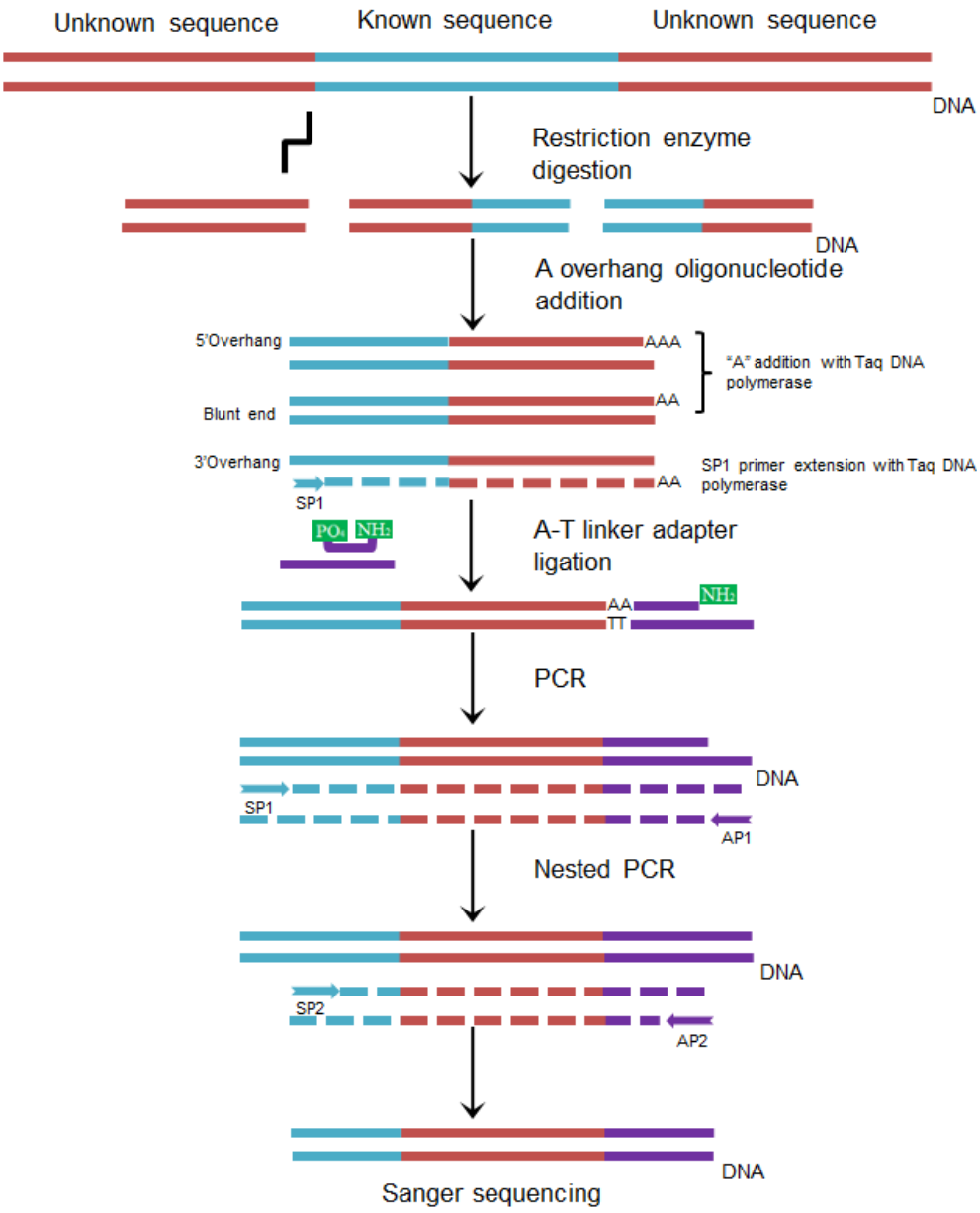


Figure 5.11 Overview of A-T linker adapter approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. The modified A-T linker adapter is indicated by two purple lines, and is partially complementary to each other. Enrichment of the target sequence is performed using specific primers (SP1 and SP2), indicated by blue arrows, and adapter primers (AP1 and AP2), indicated by purple arrows [232].

5.4.3 TOPO vector ligation PCR

The TOPO vector ligation PCR approach is related to the A-T linker adapter PCR [245]. The fragments from restriction digestion are ligated into a vector with flanking known adapter primer sequences. By combining a vector primer and a primer for the known sequence, the amplified product can include an

unknown sequence upstream or downstream of the known primer (depend upon the orientation of the cloned fragment in the vector). The selection of the restriction enzyme is critical to obtain fragments of amplifiable size suitable for successive sequencing [245]. This approach was adapted to NGS by modifying the 5′ tails to include Illumina adaptor sequences (HtStuf) [266]. In this approach, the insertion sites of 9 out of 10 transgenic soybean lines, as well as major transgene rearrangements in these soybean lines were characterised [266].

5.4.4 Loop-linker PCR

Loop-linker PCR (Figure 5.12) is another restriction enzyme based enrichment approach. In Loop-linker PCR, DNA was cleaved using multiple restriction enzymes that produce similar overhang sequences in the digested DNA [232]. The digested DNA fragments were subsequently ligated to a loop-linker adapter that was designed to form a nick site when ligated to the restricted DNA. The initial PCR was performed using specific primer 1 (SP1) and loop adapter primer 1 (LAP1), followed by a nested PCR using specific primer 2 (SP2) and loop adapter primer 2 (LAP2) to increase the specificity and the quantity of the desired amplicon (Figure 5.12). This approach was successfully evaluated in three GM maize LY038, DAS-59122-7 and Event 3272 and one GM soybean MON89788 [232].

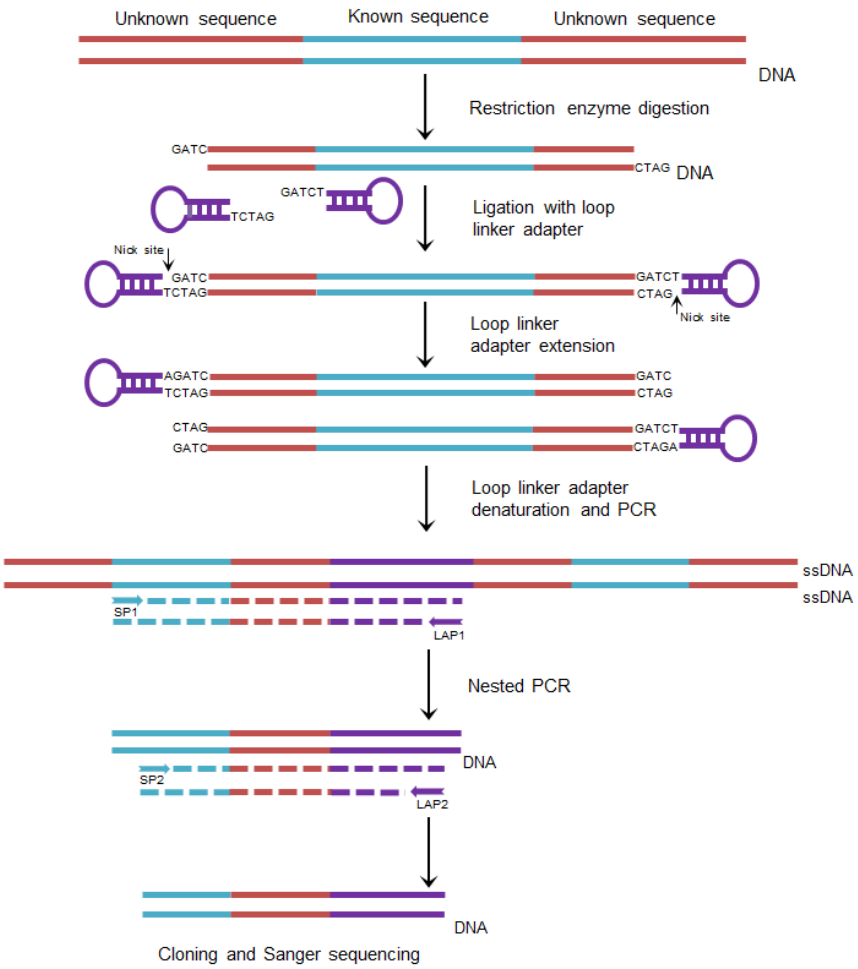


Figure 5.12 Overview of Loop-linker PCR approach. For each step, solid lines indicate the molecule present at the start and broken lines indicate the newly synthesised molecule. Blue solid and broken lines indicate the known sequence. Maroon solid and broken lines indicate the unknown sequence. The Loop-linker is indicated in purple. Enrichment of target sequence is performed using specific primers (SP1 and SP2), indicated by blue arrows, and loop adapter primers (LAP1 and LAP2), indicated by purple arrows [232].

It was reported that fragments of different lengths were obtained due to the use of different restriction enzymes and the distance between the specific primer and restriction site in the genome. The maximum length obtained in the left border adjacent sequence was 442 bp for GM maize LY038, 1830 bp for DAS-59122-7, 107 bp for Event 3272 and 512 bp for GM Soybean MON89788 [232]. A subsequent nested PCR was required in this approach due to the limited amplification of the target sequences and also to suppress the nonspecific amplification in the initial PCR [232].

5.5 Next generation sequencing (NGS) analysis

The enrichment approaches will always deliberately yield a mixture of sequences because they aim at simultaneous identification of several sequences flanking the same known element/sequence motif. Applying Sanger sequencing to enrichment products based on the described strategies is therefore not directly compatible with the mixtures of sequences obtained with these strategies. Sanger sequencing requires a single template per reaction, since mixtures will result in overlapping trace files that cannot be related to individual sequences. Therefore, extra steps are needed to separate the mixtures into distinct sequences. This can be done by cloning of the sequences and sequencing individual clones. This has the benefit of a high chance of obtaining clean, individual sequences. A disadvantage is a likely bias for sequences resulting in successful ligation into a cloning vector, severely reducing the diversity of the obtained mixture. Separation of the mixture on gel and cutting out distinct bands is less time consuming, but may still result in mixtures of sequences of similar length. Both cloning and gel separation require extra steps, handling PCR products. They are therefore extremely prone to causing contaminations in the lab, picking up contaminations that may be present in the lab, i.e. from a previous enrichment experiment, and cross-contaminations when analysing several samples in parallel.

The use of NGS, or parallel sequencing, may overcome these obstacles to find an *a priori* unknown sequence, as present in most UGMOs. Currently, the probably most cost-effective way of employing NGS is by the identification of enriched targets rather than by whole genome sequencing. Obviously, this method is limited to UGMOs that contain at least one known GMO element, otherwise whole genome sequencing is inevitable. The latter would currently be very costly to reach enough sequencing depth to cover all GMO related sequences down to the level of 0.1% for all ingredients in a complex mixture [46,267]. On top of that, elaborate analyses would have to be performed to sift the UGMO related sequences from the genomic sequences, especially for partially known genomes. Both multiplex and parallel sequencing can be achieved in an NGS approach, but so far, NGS has not yet been often applied in the field of GMO detection and identification. Selecting the appropriate platform for NGS sequencing will be based on the specific requirements and is not part of the current study, but in general where amplification-based NGS platforms (Illumina MiSeq and Ion Torrent) are well known and have been applied in prokaryotic and eukaryotic genome sequencing studies [14,33]. Single molecule based approaches (PacBio) might be advantageous as they do not require initial amplification and might be able to generate longer reads for better identification of GMOs and UGMOs, without the need of assembling fragmented reads associated with the risk of creating chimeric sequences that mimic UGMO sequences.

5.6 Discussion and Conclusion

GMO enforcement laboratories may benefit from the recent developments in sequencing technologies and bioinformatics. This is especially true as the current standard qPCR is becoming increasingly expensive in the light of the growing number of GMO events and elements that need to be included in informative screening assays. NGS may prove most effective in identification of UGMOs in food and feed products, provided it will be feasible to enrich DNA samples for not only known sequences of GMO elements, but also for the adjacent sequences. As these issues are different from those that have so far been at the basis of NGS approaches, it is necessary to develop innovative concepts that may serve the purpose of identifying GMOs and UGMOs in a single analysis. Alternatively, whole genome sequencing may be applied, but in many cases, this is still relatively expensive, and generally not applicable in the case of (complex) mixtures.

This review provides an overview of currently available approaches for the enrichment and characterization of sequences adjacent to known sequence motifs, with potential application for detection and identification of GMOs. These enrichment approaches have found application in various fields, and some have already been applied for detection and identification of, primarily unauthorised, GMOs. In the most recent years some have even been coupled with NGS analysis, as reviewed above. Coupling enrichment and NGS is a means to improve cost- and time-efficiency for broad-targeted GMO analysis of multiple samples.

It is important to assess whether these enrichment-NGS coupled approaches can meet the minimum criteria for the detection of GMOs, including UGMOs. In practice, the choice of DNA pre-treatment will strongly influence the efficiency of the enrichment approaches. Different DNA pre-treatments were applied in the described enrichment approaches. The inserted sequence can be known fully, partly or not at all. In this perspective, the enrichment approaches that imply restriction enzyme based digestion pre-treatment (Classical restriction enzyme digestion, A-T linker [232] and Loop-linker PCR [232]) or physical fragmentation pre-treatment (HITS [224] and RBF-PCR [231]) are not ideal options. In restriction enzyme based digestion pre-treatment approaches, choosing the restriction enzyme may compromise the broad detection and identification of UGMOs in those cases where the enzyme happens to digest an unknown genetic construct at crucial positions.

With physical, ultrasonic, fragmentation pre-treatment (HITS [224] and RBF-PCR [231]), similar problems might occur. Too short fragments or fragmentation within the insert hamper the enrichment for long and informative fragments. In both these approaches, the length of the enriched fragments depends on the parameters involved in the ultrasonication. Controlling for the size of the fragmentation was shown to be challenging and failure leads to creation of multiple unevenly sized bands [231]. This will compromise the applicability of both HITS [224] and RBF-PCR [231] for both GMO and UGMO detection similar to the restriction enzyme approaches.

The last approach, i.e. no DNA pre-treatment, as applied in LT-RADE [228], LAM-PCR [229], nrLAM-PCR [225], SiteFinding-PCR [227] and LF PCR [230] seems the best option, leaving the whole genome intact for the initial target specific primer extension. As a result, a maximum enrichment length can be obtained starting from the known element of GMOs into the unknown adjacent sequences. This is crucial for successful identification of UGMOs.

The different enrichment approaches reviewed here were compared with relation to different aspects that were deemed relevant for their success when applied in routine screening approaches for GMO detection and identification (Table 5.2).

Table 5.2 Overview on the reported enrichment approaches, based on the length of enrichment, tested on GMOs, number of steps involved, number of PCR steps, estimated time, coupled with NGS and DNA pre-treatment [219,223-225,227,228,230-232,253]

Enrichment approaches	Length of enrichment	Tested on GMOs	Number of steps involved	No of PCR steps involved	Estimated time	Coupled with NGS	DNA pre-treatment
SiteFinding-PCR	max ~4.5 kb	Yes	5	3	~11h 30min	Yes	No pre-treatment
Locus-finding PCR	max ~500 bp	Yes	3	2	~5h 30min	No	No pre-treatment
LT-RADE	max ~1 kb	Yes	5	2	~7h min	No	No pre-treatment
nrLAM-PCR	max ~250 bp	No	5	2	~29 h	Yes	No pre-treatment
HITS	max ~ 350 bp	No	4	1	~3h 30min	Yes	Ultra-sonication
RBF-PCR	max ~ 2 kb	Yes	6	2	~10 h 30min	No	Ultra-sonication
A-T linker	max ~1 kb	No	6	2	~36h 30min	No	Restriction digestion
Loop-linker PCR	max ~1.8 kb	Yes	5	2	~16h 30min	No	Restriction digestion
Restriction digestion- NGS	max ~500 bp	No	4	1	~19h 30min	Yes	Restriction digestion

In approaches where DNA is not pre-treated, initial single strand elongation can be obtained by using either of two types of primers: a) a semi-random primer (SiteFinding PCR [227] and LF PCR [230]), or b) a target-specific primer (LT-RADE [228], LAM-PCR [229], and nrLAM-PCR [225]). For the detection of GMOs and UGMOs it is important to initiate the enrichment with the target-specific primer rather than initiating with semi-random primer. This will ensure that from the initial step onwards, the desired fragment is amplified. In the case of SiteFinding-PCR [227] and LF PCR [230], multiple amplicons are amplified in the initial step due to the use of semi-random primers largely unrelated to the sequence of the GMO element of interest. In these cases, it is essential to include an additional step to select for the desired GMO-related amplicons. In SiteFinding-PCR [227] this was achieved by performing additional nested PCRs with target specific primer and SiteFinding primer. This, however, increases the risk of contamination. In LF PCR [230], a biotinylated capture primer step was included to select for the desired amplicons, but a nested PCR was also still performed. Based on these observations, SiteFinding PCR [227] and LF PCR [230] do not meet the criteria for the enrichment for UGMO detection and identification.

The target-specific primer approaches either use biotinylated target-specific primers (LAM-PCR [229] and nrLAM-PCR [225]) or non-biotinylated target-specific primers (LT-RADE) [228] for single strand enrichment. To reduce the presence of genomic DNA and non-specific amplicons, a magnetic bead clean-up step was introduced in the LAM-PCR [229] and nrLAM-PCR [225] procedures. For that purpose biotin was added to the 5′ end of the target-specific primer. The reduction of the background genomic DNA will increase the sensitivity and reduce the necessity for subsequent nested PCRs that are prone to contamination. In the LT-RADE [228] approach, column purification was introduced to remove the shorter fragments (< 100 bp). This purification step is important to facilitate the poly-dC tail synthesis in the longer fragments. Subsequently, dsDNA can be obtained in different ways. In LAM-PCR [229], a mixture of random hexanucleotide hybridises randomly to the ssDNA to allow synthesis of dsDNA that will lead to different amplicons. The subsequent restriction digestion followed by ligation may be prone to contamination and is time consuming. In nrLAM-PCR [225], single-stranded linkers are ligated to the ssDNA and a PCR is performed using the gene-specific primer and a linker primer. In nrLAM-PCR [225], single-stranded linker ligation to the ssDNA is not as efficient as dsDNA ligation. In the LT-RADE [228] approach, the synthesized ssDNA is poly-dC tailed and dsDNA is synthesized by performing a PCR on the basis of the target-specific and AAP primers. It was shown that the blend polymerase used in the LT-RADE approach results in longer enrichment DNA stretches when compared to the RADE approach, which uses regular *Taq* polymerase [228]. A further evaluation study between the different polymerases is recommended to choose an optimal polymerase (blend) that can enrich longer fragments for UGMO identification. From this overview, it seems clear the most suitable approaches to synthesize the dsDNA use amplification based on either random hexanucleotide primers (LAM-PCR) [229] or primers based on initial poly-dC tailing (LT-RADE) [228]. Additional research is necessary to evaluate the efficiency of the dsDNA synthesis in both cases. It is necessary to evaluate the quantity of the enriched fragments after the clean-up step in a single PCR. If the resulting amount of DNA proves to be sufficient for direct sequencing, this can be both time-efficient, and contribute to reducing the contamination risk. To further increase the cost efficiency of such analyses, it would be beneficial to perform the amplification steps of target-specific primer approaches in a multiplex fashion. With newly emerging techniques, such as droplet digital PCR (ddPCR), this could be done quite effectively. As already shown, at least ten different targets can be simultaneously amplified in one reaction of ddPCR [268]. Thus the option of fusing the target-specific primer approaches into droplet (emulsion) PCR format would be of interest.

For UGMO detection and identification, it is necessary that enrichment strategies will primarily enrich the longest possible fragments based on selected, targeted elements in GMOs in a sensitive way, performing equally well in complex mixtures. None of the discussed enrichment approaches have been shown to fully meet these requirements, and none have been tested in more processed samples that are commonly encountered in GMO detection in food and feed. Even approaches that were designed for obtaining long fragments may fail to yield enough information in the case of samples with highly fragmented DNA. The enrichment approaches that hold the best perspective for UGMO detection and identification are those that start the initial extension with target-specific primers, especially nrLAM-PCR and LT-RADE. It can be envisaged that the further comparison and development of this class of enrichment approaches can lead to efficient approaches for UGMO detection and identification, especially when coupled to NGS, as this will allow for simultaneous detection and identification of all GMOs, including UGMOs, in a single analysis. Thus, the present review will provide the basis for the development of effective methodologies to screen food and feed products for GMOs that have not yet been tested for their safety for the human and animal consumer and the environment.

Chapter 6

NGS-based amplicon sequencing approach; towards a new era in GMO screening and detection

This chapter was published as: **Arulandhu AJ**, van Dijk JP, Staats M, Hagelaar R, Voorhuijzen M, Molenaar B, van Hoof R, Li R, Yang L, Shi J, Scholtens I, Kok EJ. "NGS-based amplicon sequencing approach; towards a new era in GMO screening and detection" *Food Control* 2018; 93: 201-210

Abstract

The development and commercialization of Genetically Modified Organisms (GMOs) and its related products have been increasing in the last two decades. This challenges the currently applied time-consuming and expensive qPCR screening procedure from a practical perspective, due to the necessity to develop and validate additional targets at a regular pace and the increasing number of targets included in a single screening. In this study we developed a next generation sequencing (NGS)-based GMO screening approach covering 96 GMO targets and compared it to the two-step qPCR GMO screening approach; the two approaches were evaluated with five feed samples known to contain GMOs. The amplicons obtained from the feed samples were analysed using 150-bp Paired-End sequencing, Illumina HiSeq 4000 platform. A dedicated data analysis pipeline was developed, which allows automated identification of GMOs and associated genetic elements and constructs. The result of the NGS-based screening were compared with the qPCR approach, indicating that 92% of the targets were commonly identified between the qPCR and NGS-based screening. The remaining 8% of the targets had discrepancies in detection between the two methods, this was mainly observed for targets that were detected in qPCR with high Cq values (above 36), which could not be detected in NGS-based screening. Additionally, due to the more extensive screening in the NGS-based strategy, in total 43 additional GMOs and related targets were identified compared to the standard qPCR screening. From the commonly identified targets in both approaches, 8 targets could not be associated with the detected GMOs. These targets had late Cq values (above 36) and could indicate traces of unknown GMOs in the samples. The current study shows the applicability of NGS as a novel, broad and reliable screening strategy for GMOs and its potential to improve current screening methods.

Keywords: Genetically modified products, qPCR, NGS, amplicon sequencing, Illumina HiSeq, bioinformatics, Unauthorised GMOs.

6.1 Introduction

Genetically modified organisms (GMOs) are nowadays produced and marketed globally [269]. The regulations for the use of genetically modified (GM) crops in food and feed samples vary among countries, but in most countries GMOs need a pre-market authorization, and unauthorised GMOs (UGMOs) are not allowed to be present in food and feed products. For enforcement of GMO regulations, a two-step quantitative real-time PCR (qPCR) screening approach is the most commonly used strategy to detect and identify approved GMOs as well as potential UGMOs in food/feed samples [2-8]. In the qPCR screening approach, samples are analysed for common crop species (endogenous genes) as well as GMO targets (GMO-related elements and genetic constructs). Depending on the outcome of this initial screening step, additional event-specific qPCRs may be performed to confirm the presence of specific GMO events. In some cases, the detected elements, or combinations of elements, cannot be explained by the presence of an authorised GMO in the same sample, which may indicate the presence of a UGMO [22]. UGMOs are GMOs that have not yet been assessed for their food, feed and environmental safety in the country where these are marketed. Currently, the growing number and increasingly diverse characteristics of GMOs and UGMOs on the global market urge enforcement laboratories to screen for a growing number of GMO related targets. Moreover, new GMOs will continue to enter the market, requiring regularly expanding of the number of targets in the initial screening step. Incorporating this broad screening into the current qPCR strategy makes the procedure increasingly time-consuming. Furthermore, the addition of every new target to the screening requires the development and validation of a new method, a time and budget consuming process. In recent years, a number of multiplexing detection approaches have been developed to facilitate the broad screening of targets in a cost-efficient way. Generally, these approaches are probe and hybridization-based, such as microchip-PCR [221], microarrays [222], 384 well-plate qPCR, microfluidics dynamic array [205] and droplet digital PCR (ddPCR) [270]. Except for ddPCR, all of these newly developed assays were designed to facilitate simultaneous screening of a high number of targets, allowing the detection of GMOs and UGMOs in complex samples. While these screening methods allow for a high number of targets, they still require development and validation of GMO specific probes, similar as in qPCR analysis. In recent years, next generation sequencing (NGS) has been applied in the field of GMO detection and identification, focusing on the identification of UGMOs by using genome walking approaches [219,271,272] and whole genome sequencing (WGS) of GMOs crops [40,45,46]. However, NGS-based detection has not yet been developed and applied in protocols for routine GMO analysis. An NGS-based screening approach would have the following advantages over a qPCR-based approach: a) the reported amplicon sequence will be a direct proof of detection of the target, which circumvents the process of developing and validating target-specific probe methods and the need for the appropriate reference materials. b) Reference sequence information of amplicons can be generated and any variation from previously published sequences can be identified and may be used in future studies, thus increasing the knowledge basis.

Here as a comparative study, an NGS-based GMO screening approach was developed, and the applicability of the approach was evaluated by comparing it with the standard qPCR screening approach. To this end, five complex feed products were selected from a routine GMO feed monitoring programme. The feed products were all known to contain multiple GMOs in different quantities, as had been determined in the routine two-step GMO qPCR screening. The developed NGS-based GMO screening approach was based on 96 targets PCR assay [22] combined with Illumina paired-end (PE) technology. The Illumina technology was preferred over other NGS sequencing technologies based on its performance characteristics, such as the output read length, number of output reads, read quality, runtime, type of reads and lowest cost per sequenced base pair [23]. The obtained amplicon NGS datasets were analysed using a bioinformatics pipeline, amplicon-sequencing (AM-SEQ), developed for

this purpose. The results of the developed NGS based screening and qPCR screening were compared and the applicability of the NGS-based approach for routine GMO screening was evaluated.

6.2. Materials and methods

6.2.1 GMO feed samples

Five feed samples from GMO feed monitoring programme were used for analysis with both the NGS and the qPCR-based approach (Table 1). Based on the product description sample 1 (S1) and sample 3 (S3) contained maize GMOs and sample 4 (S4) contained soy GMOs. The label description for sample 2 (S2) was supplementary feeding for chickens and for sample (S5) calves growing mix, in both cases no specific GMO-related information was provided. These samples originated from Brazil, Canada and The Netherlands.

Table 6.1 GMO feed samples used for analysis.

Sample name	Product description	Sample origin
Sample 1 (S1)	Maize GMO	Brazil
Sample 2 (S2)	Supplementary feeding, chickens	Netherlands
Sample 3 (S3)	Maize GMO	Canada
Sample 4 (S4)	Soybean scrap GMO	Brazil
Sample 5 (S5)	Calves growing mix	Netherlands

6.2.2 DNA isolation from feed samples

Per sample, 100 mg of ground, homogenised dry weight material was used. DNA was isolated using the following procedures; the Maxwell 16 Tissue DNA purification kit was used in combination with the Maxwell 16 Instrument (Promega, USA) to isolate DNA from samples 1,4 and 5, whereas the modified Qiagen DNeasy Plant mini kit (Qiagen) with CTAB extraction [186] was used for samples 2 and 3. DNA for the 96 GMO targets of the positive control sample was obtained from several certified reference materials (Additional file 6A: Table A.1) by following either Maxwell 16 Tissue DNA purification kit for soybean, rice and sugar beet materials or a modified Qiagen DNeasy Plant mini kit (Qiagen) with CTAB extraction for maize, potato, cotton, canola and wheat materials. The purity and quantity of the extracted DNA was assessed using Nanodrop absorbance measurements (Nanodrop 1000 instrument, Thermo Fisher Scientific).

6.2.3 GMO qPCR screening and detection

The DNA from the samples was diluted to ~10 ng/µl prior to qPCR amplification. Routine two-step GMO qPCR screening was performed for the five feed samples. Initially, qPCR for endogenous reference genes, GMO specific elements, constructs (Table 2) were performed, and based on the outcome additional event-specific qPCR were performed to confirm the presence of the GMOs in the samples. All qPCRs for each target were performed in two replicates including the positive (reference material for each target) and negative control (reaction without DNA, DNA volume was replaced by PCR grade

water). The total volume per reaction was 25 µl, which contains 12.5 µl of the Diagenode master mix (DMMM2XA300), 50 ng of DNA, the used concentration of forward and reverse primers, and the probe for each target are specified in Appendix A: Table A.1. All qPCRs were performed using the following protocol: decontamination UNG (uracil-DNA glycosylase) 120 s at 50 °C, initial denaturation 600 s at 95 °C, amplification 45 cycles of 15 s at 95 °C and 60 s at 60 °C using a BioRad CFX96 thermocycler. The obtained qPCR data were analysed using the Bio-Rad CFX Manager 3.0 software.

Table 6.2 GMO specific targets (endogenous, elements, constructs and events) analyzed in the qPCR and NGS-based approach.

Endogenous PCR (n=9)	**Element PCR (n=31)**	**Construct PCR (n=9)**	**Event PCR (n=47)**	
Actin-Plant DNA [1 2 3 4 5]	*P-35S* [1 2 3 4 5]	*Adh1-cry1Ab*	A2704-12 (Soybean) [2 3 5]	MS8 (Canola)
Potato	*P-FMV* [3]	*Cry1Ab*-intron	A5547-127 (Soybean) [2 3 5]	RF3 (Canola)
Cotton	*P-FMV2* [1 2 4 5]	*Ctp2-cp4epsps* [1 2 3 4 5]	CV127 (Soybean) [2 3]	GT73 (Canola) [5]
Canola [1 2 3 4 5]	*P-NOS* [1 2 3 4 5]	*Ctp4-cp4epsps* [1 2 3 4 5]	DAS44406 (Soybean)	T45 (Canola)
Maize [1 2 3 4 5]	*P-Riceactin* [1]	*Hsp70-cry1Ab*	DAS68416 (Soybean) [2]	281-24-236 (Cotton)
Rice [1 2 3 4 5]	*P-SSuAra* [1 2 4 5]	*OTP-mepsps*	DP305423 (Soybean) [2 3 4 5]	3006-210-23 (Cotton)
Soybean [1 2 3 4 5]	*P-TA29*	*P-35S-bar*	DP356043 (Soybean) [2 3 5]	GHB119 (Cotton)
Sugar beet [1 2 3 4 5]	*P-ubi* [3]	*Pat-T-35S*	FG72 (Soybean) [5]	GHB614 (Cotton)
Wheat [1 2 3 4 5]	*T-35S* [1 2 3 4 5]	*P-ubi-cry*	MON87701 (Soybean) [2 4 5]	LL25 (Cotton)
	T-E9 [1 2 3 4 5]		MON87705 (Soybean) [3 4]	MON1445 (Cotton)
	T-G7 [1 2 4 5]		MON87708 (Soybean) [3]	MON15985 (Cotton)
	T-nos [1 2 3 4 5]		DAS81419 (Soybean)	MON531 (Cotton)
	T-OCS		GTS 40-3-2 (Soybean)	
	cp4-epsps(1) [1 2 3 4 5]		MON87769 (Soybean)	
	cp4-epsps(2)		MON89788 (Soybean) [2 3 4 5]	
	Cry1Ab [1 2 3 4 5]		Event 3272 (Maize) [3]	
	Cry1A.105 [1 2 3 4 5]		BT11 (Maize) [1 2 3]	
	Cry1Ab/Ac [1 2 3 4 5]		DAS40278 (Maize) [2 3]	
	Cry1Ac [3]		DAS59122 (Maize) [3]	
	Cry1F [1 2 3 4 5]		DP98140 (Maize) [2 3]	
	Cry2Ab2 [1 2 3 4 5]		GA21 (Maize) [3]	
	Cry3A		MIR162 (Maize) [1 2 3]	
	Cry3Bb1 [1 2 3 4 5]		MIR604 (Maize) [3]	
	Vip3a [1 2 3 4 5]		MON810 (Maize) [3]	
	Bar [1 2 3 4 5]		MON863 (Maize) [3]	
	Pat [1 2 3 4 5]		MON87460 (Maize) [3]	
	nptII [1 2 3 4 5]		MON88017 (Maize) [1 2 3]	
	I-rActin1 [1 2 3 4 5]		MON89034 (Maize) [1 3]	
	barnase		NK603 (Maize)	
	barstar [1 2 3 4 5]		T25 (Maize)	
	CaMV [1 2 3 4 5]		DAS1507 (Maize) [1 2]	
			5307 (Maize)	
			MON87427 (Maize)	
			DP073496 (Canola)	
			MON88302 (Canola)	

The numbers in superscript indicate the sample (name) analyzed for the specified qPCR target (e.g. 1 refers to qPCR targets in S1).

6.2.4 Comparison of master mixes for NGS-based GMO detection

The comparison was performed on four GMO related targets in three different samples; *hmg*, *P35S*, *T-nos* and *Cry1Ab/Ac.* The reactions were carried out in 25 µl volumes containing 12.5 µl of PCR master mix (Diagenode master mix (DMMM2XA300) or HotStar *Taq* master mix kit (Qiagen; 203443)), sense and antisense primers (Appendix A: Table A.1), RNase-Free water, and 5 µl of template DNA (~10 ng/µl). The following PCR protocol was used: 95° C for 10 min, 45 cycles of 95° C for 15 sec, 60° C for 1 min using a BioRad CFX96 thermocycler. The 5 µl of the PCR products and 3 µl of 10 bp DNA ladder were loaded on 1% agarose gel containing ethidium bromide, Gel Doc XR+ System (BIO-RAD) was used to visualize the amplified product.

6.2.5 PCR and Illumina sequencing

The positive control DNA (~10 ng/µl) and primer concentration used for each PCR target is specified in Additional file 6A: Table A.1. Using HotStar *Taq* master mix kit, the PCRs were performed in a 96-well plate (Bio-Rad; 1 plate per sample), with each reaction well containing the primers of a different PCR target. After PCR amplification, 5 µl of the amplified product from the positive control plate (96 targets) was visualised on 1% agarose gel. The 96-well plates were sent for sequencing to the BGI genome sequencing centre (Shenzhen, China), where the 96 PCR targets per plate were pooled, and purified using the QIAquick PCR purification kit (Qiagen). Next, 3 µl of each DNA sample was quantified using the Qubit dsDNA BR Assay Kit (ThermoFisher Scientific). Indexed paired-end adapters were then ligated to 1 ug of DNA per sample using the TruSeq DNA Sample Preparation kit (Illumina Inc.) according to manufacturers' protocol. The index barcodes in the adapters, used to label different libraries in a HiSeq lane, were generated by the sequencing provider. The six DNA libraries used in this study were pooled with other libraries that were not part of this study and sequenced together on a HiSeq 4000 lane using paired-end 150-bp mode.

6.2.6 Bioinformatics analysis

Identification of GMO related reference sequences with the AM-SEQ pipeline involved the following steps: 1) Illumina adapters were removed from reads with a 10% error tolerance using Cutadapt v1.8.1 [273]. 2) The forward and reverse reads were merged to create a pseudo-read using PEAR [274]. 3) A quality filtering step was performed with the criteria of a base quality ≥ 20, using the FASTX-toolkit v0.014 (http://hannonlab.cshi.edu/fastx_toolkit). 4) The pseudo-reads that contained both assay-specific primer sequences were selected and sorted using Cutadapt v1.8.1, in sequential order. The reads that did not contain the first assay-specific primer sequences on the list (eg. Actin) were searched and sorted for the next target assay-specific primer sequences, this process was performed for all the 96 targets included in this study. 5) All primer selected reads of the positive control were aligned against the local database, which contained 17 reference sequences obtained from NCBI database. 6) Unmapped reads were clustered using USEARCH v8.0.1632 [189] with a minimum cluster size of 50. The representative read of each cluster was manually verified for the presence of a target-related probe sequence and, if positive, the representative read was added to the local database as a new reference sequence. After the local database of GMO related reference sequences was established, the NGS datasets generated for the positive control and five feed samples were analyzed. Amplicon targets that were supported by at least 0.01% of the total mapped reads were considered to be detected. As a final step in the data analysis, detected targets (endogenous, elements and constructs) were associated with the detected GMOs in the sample and presented in a graphical overview.

6.3 Results

6.3.1 NGS-based GMO screening

The applicability and performance of an NGS-based GMO screening approach was determined by evaluating five feed samples (S1-S5), confirmed to contain GMOs. Furthermore, a positive control sample was included in the analysis containing DNA of all relevant GMO targets, obtained from several certified reference materials (Additional file 6A: Table A.1). As an initial step to optimize the NGS-based GMO screening procedure, the Diagenode master mix and HotStar *Taq* master mix kit (Qiagen) were tested for their ability to amplify four GMO related targets in three different samples; *hmg*, *P35S*, *T-nos* and *Cry1Ab/Ac* in two GMO feed samples and a positive control. The comparison results indicated that the amplification efficiency of HotStar *Taq* master mix was consistently high for all four amplicons under the selected PCR thermocycling conditions, whereas fainter PCR bands were observed for *hmg* and *Cry1Ab/Ac* using the Diagenode master mix (Additional file 6A: Figure A.1). In addition, no primer dimers or non-specific amplification products were observed with HotStar *Taq* master mix. For these reasons, the HotStar *Taq* master mix was selected for further PCR analyses using the DNA from the five feed samples and the positive control to amplify 96 targets per sample: 9 crop-specific, 31 elements, 9 constructs and 47 events (Table 2). The 96 positive control reactions were visualized on gel, confirming that all PCR targets were successfully amplified, except for the event-specific sequences of DP073496, DAS59122-7 and 3006-210-23 (Additional file 6A: Figure A.2). The pooled 96 reactions per sample were purified and sequenced using paired-end (PE) 150 HiSeq Illumina technology. In total, 51.6 Gb of NGS data was yielded form the five feed and positive control samples

6.3.2 Creating a local database and processing the NGS data with AM-SEQ

A dedicated bioinformatics pipeline named AM-SEQ was developed and used to process the raw Illumina data in 1.8+ FASTQ format (Figure 6.1). For the correct identification of GMOs and related targets a local reference sequence database for all 96 GMO targets was built using the following two approaches; a) mining the published literature and in-house or public databases (EUginius, NCBI nucleotide) for reference sequences and b) identifying reference sequences from the positive control NGS dataset. With approach "a", 17 of the 96 reference sequences could be identified and were added to the local database; *Actin*, *SPS*, *P35S*, *PFMV*, *cp4-epsps*(1), *cry1A(b)*, *Cry1A.105*, *Cry1F*, *Cry2Ab2*, *Cry3A*, *Bar*, *nptII*, *I-rActin1*, *Barnase*, CaMV, DAS59122-7 and 3006-210-23 (Additional file 6A: Table A.3). To identify the reference sequences of the remaining 79 targets the NGS dataset of the positive control was processed with the AM-SEQ pipeline (Figure 6.1). During the initial quality filtering, the 5′ and 3′ Illumina adapter sequences were removed, and reads shorter than 50 nucleotides were discarded. Reads with a minimum base quality of 20 were selected for further analysis. The base quality selected forward and reverse reads were merged into pseudo-reads, a primer selection based on presence of both the assay-specific forward and reverse primers was performed on these base quality selected pseudo-reads. Furthermore, during the selection procedure pseudo-reads were sorted according to their PCR assay. The remaining unmapped pseudo-reads were clustered with a minimum cluster size of 50. Subsequently, the representative pseudo-reads of a cluster were manually analyzed for the presence of assay-specific probe sequences to verify the biological relevance, if the probe sequence was present, then the representative pseudo-read was added to the local database. The representative pseudo-reads of the unmapped clusters that did not contain a probe sequence were not added to the local database.

The positive control contained reads for 93 targets based on the assay-specific primer selection. For three targets no reads were found (DP073496, DAS59122-7 and 3006-210-23), as specified in Additional file 6C: Table C.6. This was due to no amplification of the target during PCR, as observed on gel (Additional file 6A: Figure A.2), however, the reference sequences for DAS59122-7 and 3006-210-23 were already available, only for DP073496 the reference sequence could not be obtained. For the 17 targets with sequence information previously added to the local database, reads were mapped using Bowtie v2.2.6 with default setting to their respective reference sequences. For the remaining 79 GMO targets, reference sequences were selected based on the assay-specific primer and were manually verified for the presence of the specific probe sequence and added to the local database and EUginius (http://www.euginius.eu). For *Actin*, two sequence variants were included in the local database; one from *Zea mays*, and one from *Brassica napus*. (Additional file B: Table B.1).

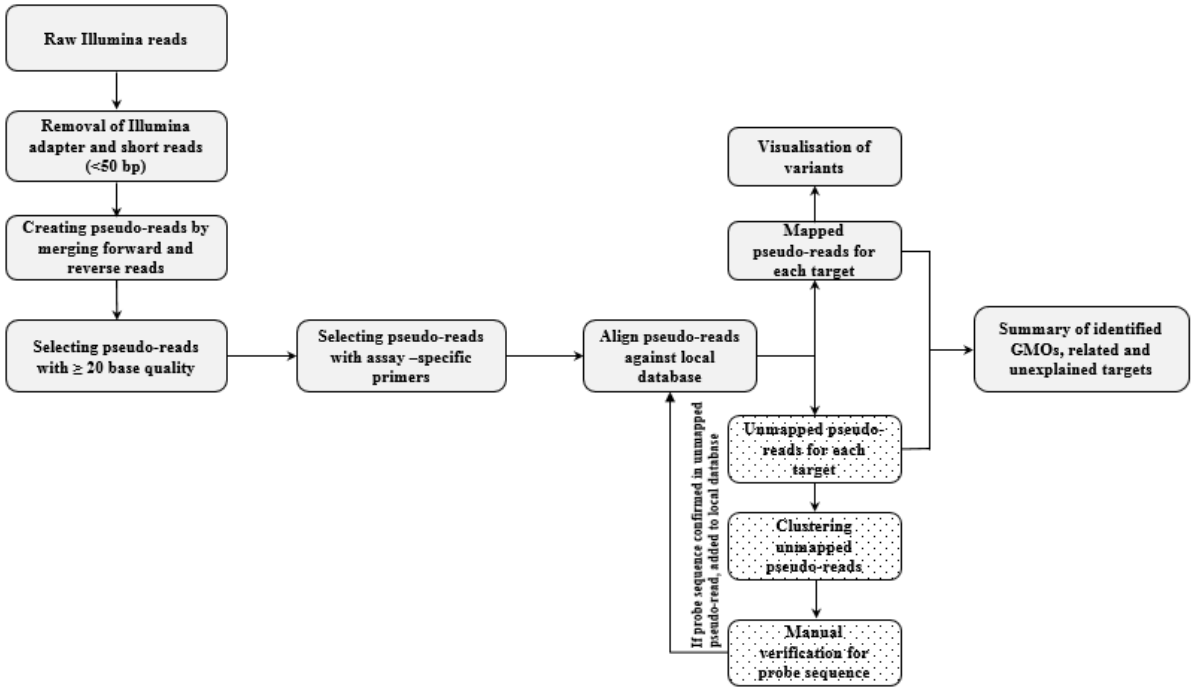


Figure 6.1 Schematic representation of the amplicon resequencing pipeline (AM-SEQ) to process the Illumina data. The dotted boxes indicate the steps in obtaining the reference sequences from the NGS data for the local database.

With the complete local database in place, the NGS data yielded from the five feed and the positive control samples were analysed using the AM-SEQ pipeline. On average, 24 million raw forward and reverse reads per sample were obtained (Additional file 6A: Table A.2). On average 93.84% (min = 91.28%; max = 96.83%) of the reads passed base quality filtering, indicating a high overall quality of the NGS datasets (Additional file 6A: Table A.2). During primer selection, on average 91.83% of the reads (min = 90.18%; max = 92.91%) were selected based on the presence of both the assay-specific forward and reverse primers. Next, the pseudo-reads were sorted according to their PCR assay and aligned using Bowtie v2.2.6 with default setting to their respective reference sequences, on average 87.61% (min = 71.10%; max = 98.71%) of pseudo-reads were mapped (Additional file 6A: Table A.2). None of the representative pseudo-reads of the unmapped clusters from the feed sample contained a probe sequence, and hence these reads were discarded. An overview of the number of primer-selected pseudo-

reads, the number of mapped and unmapped pseudo-reads, and the number of clusters for unmapped pseudo-reads are presented in Additional file 6C: Table C.1-C.6.

6.3.3 Setting the threshold of detection by comparing qPCR and NGS datasets

A threshold was established to distinguish between amplified targets (targets with high read counts) and background noise (targets with low read counts) in the data. Reads were observed for nearly all PCR targets in the feed samples. On average 59 PCR targets were observed with a relatively low (<0.01%) percentage of mapped reads (Additional file 6C: Table C.1-C.6). To determine a threshold for detection and to avoid false positive results, the percentage of mapped reads per target of the positive control and S4 were plotted in ascending order (Figure 6.2). Additionally, the qPCR results of S4 were compared to the percentage of mapped reads per target, showing that nearly all GMO targets with a low amount of mapped reads (<0.01%) were not detected in the qPCR analysis, while GMO targets with high read counts (>1%) generally had low Cq values (≤ 35). Hence, to avoid the false positive identification in the samples, targets with less than 0.01% mapped reads were considered background noise and were scored as not detected.

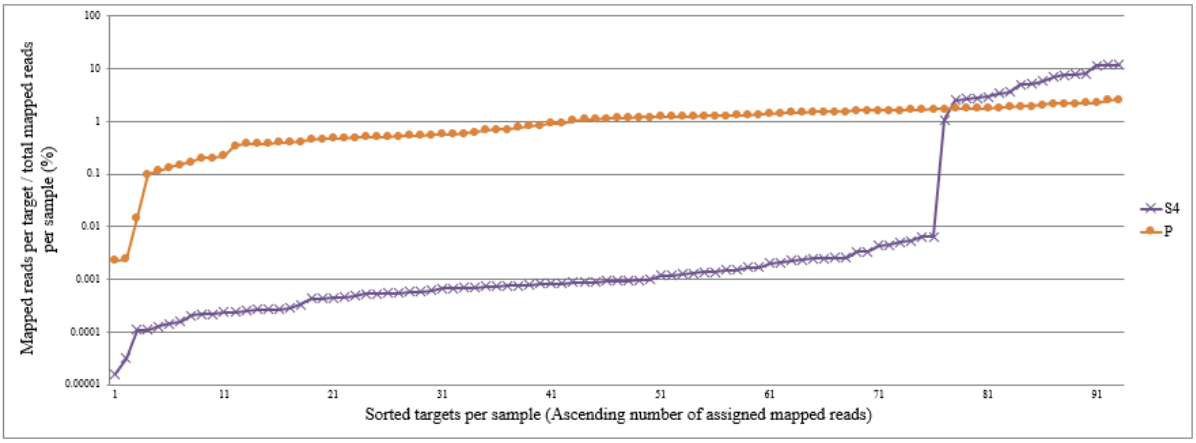


Figure 6.2 Positive control (P) and sample 4 (S4) were plotted to determine the background noise in the data. The points in the lines represent the percentage of mapped reads per target in a sample relative to the total number of mapped reads in a sample (y-axis). Per sample the targets were sorted in ascending order of assigned mapped reads (x-axis).

6.3.4 Data interpretation of the feed samples

The NGS data obtained from the feed samples were reanalysed with the established threshold. The detected targets of each feed sample were analysed using an event-related element matrix, which is based on available elements, construct and event data for 47 GMOs [186]. The graphical output, in HTML format, displaying the GMOs and related targets identified in each feed sample can be found in Additional file 6A: Figure A.3. As an example, the visual output for S2 is presented (Figure 6.3), which summarizes the identified GMOs and related targets in the sample, for the other four samples visual outputs are presented in the Appendices A: Figure A.3. In each feed sample, crop-specific targets (*FatA, UGP, Wx-1* etc.) were detected that could not be linked to GMOs indicating the presence of non-GMO crops.

Between 17 and 42 targets (endogenous, elements, construct and GMO events) were detected per sample (Table 6.3). Maize GMOs were detected in S1 and S3, whereas S4 contained only soy GMOs, and S2 and S5 contained both maize and soy GMOs. A total of 12 GMO targets were unexplained in S2 to S5 meaning that these GMO targets could not be associated to any of the GMOs identified in these samples (Table 6.3). To determine whether these unexplained targets are present in EU authorised and known unauthorised GMOs that were also present in the samples, but may have been masked by the

events found in the initial analysis, a further analysis was performed through www.euginius.eu (Figure 6.4). Based on the outcome, additional GMO specific qPCR experiments were performed.

Table 6.3 Targets identified in the five real-life feed samples using NGS-based GMO screening approach.

Sample no	Detected GMOs	Detected endogens	Detected elements	Detected constructs	Unexplained targets
S1 (33)	**GA21**, MIR162, **MON810**, MON88017, MON89034, **NK603**, DAS1507, BT11 *(8)*	*Actin, hmg and Le1(3)*	*P35S, **PFMV**, PFMV2, P-Riceactin, **P-ubi**, T-35S, T-nos, cp4-epsps1, Cry1Ab, Cry1A.105, Cry1Ab-Ac, Cry1F, Cry2Ab2, Cry3Bb1, Pat and I-rActin1(16)*	***Adh1_cry1Ab,** ctp2-cp4epsps, **hsp70-cry1Ab**, OTP-mepsps, **Pat-T-35**, **P-ubi-cry** (6)*	
S2 (36)	A2704-12, MON87701, **GTS40-3-2**, MON89788, **GA21**, MON88017, **NK603**, DAS1507 (8)	*Actin, **UGP**, FatA, hmg, Le1,Wx-1 (6)*	*P35S, **PFMV**, PFMV2, **P-Riceactin**, P-SSuAra, P-ubi, T-35S, T-E9, T-nos, cp4-epsps1, **cp4-epsps2**, Cry1Ab, Cry1Ab-Ac, **Cry1Ac**, Cry2Ab2, **Cry3Bb1**, Pat, I-rActin1,barstar (19)*	*ctp2-cp4epsps, ctp4-cp4epsps, **Pat-T-35S** (3)*	*Cry1Ab, Cry2Ab2, barstar*
S3 (42)	GA21, MIR162, MON810, MIR604, MON863, MON88017, MON89034, **NK603**, **T25**, DAS1507, BT11 (11)	*Actin, hmg, Le1,Wx-1 (4)*	*P35S, PFMV, **PFMV2**, **P-Riceactin**, P-SSuAra, P-ubi, T-35S, T-nos, cp4-epsps1**, cp4-epsps2**, Cry1Ab, Cry1A.105, Cry1Ab-Ac, Cry1F, Cry2Ab2, **Cry3A**, Cry3Bb1, Pat, nptII, I-rActin1(20)*	***Adh1_cry1Ab**, ctp2-cp4epsps, ctp4-cp4epsps, **hsp70-cry1Ab**, **OTP-mepsps**, **Pat-T-35S, P-ubi-cry (7)***	*Ctp4_cp4epsps and P-SSuAra*
S4 (17)	MON87701, GTS40-3-2 MON89788 (3)	*Actin, Le1(2)*	*P35S, PFMV2, P-SSuAra, **P-ubi**, T-E9, T-nos, cp4-epsps1, **cp4-epsps2**, Cry1Ab-Ac, **Cry1Ac (10)***	*ctp2-cp4epsps,ctp4-cp4epsps(2)*	*P-ubi*
S5 (34)	A2704-12, MON87701, GTS40-3-2, MON89788, **MON810**, **NK603 (6)**	*Actin, **UGP**, FatA, hmg, Le1,Wx-1(4)*	*P35S, **PFMV**, PFMV2, **P-Riceactin**, P-SSuAra, P-TA29, **P-ubi**, T-E9, T-g7, T-nos, cp4-epsps1, **cp4-epsps2**, Cry1A.105, Cry1Ab-Ac, **Cry1Ac**, Cry2Ab2, Bar, Pat, I-rActin1, barstar (20)*	*ctp2-cp4epsp, ctp4-cp4epsps, **hsp70-cry1Ab**, **Pat-T-35S (4)***	*Tg7, bar, Cry1A.105, Cry2Ab2, P-TA29,barstar*

Targets that were additionally detected in NGS-based screening compared to the qPCR screening are highlighted in bold. (n) Indicates the total number of targets identified.

Detected GMO Events

UIDs	Events	Endogenes	Constructs		Elements							
ACS-GMØØ5-3	A2704-12 SOY EVENT	Endo soy Le1	Pat_T-35S		P-35S	T-35S	pat					
MON-Ø4Ø32-6	GTS 40-3-2 SOY EVENT	Endo soy Le1	Ctp4_cp4epsps RRS		P-35S	T-nos	Cp4-epsps(1)	Cp4-epsps(2)				
MON-877Ø1-2	MON87701 SOY EVENT	Endo soy Le1			P-SSuAra	Cry1Ab/Ac	Cry1Ac					
MON-89788-1	MON89788 SOY EVENT	Endo soy Le1	CTP2_CP4EPSPS		P-FMV	P-FMV(2)	T-rbcS-E9 (pea)	Cp4-epsps(2)				
MON-ØØØ21-9	GA21 MAIZE EVENT	Endo maize HMG	OTP_mepsps GA21 MAIZE		P-Rice actin	P-Ubi	T-nos	I-rAct1				
MON-88Ø17-3	MON88017 MAIZE EVENT	Endo maize HMG	CTP2_CP4EPSPS		P-35S	P-Rice actin	P-Ubi	T-nos	Cp4-epsps(1)	Cp4-epsps(2)	Cry3Bb1	I-rAct1
MON-ØØ6Ø3-6	NK603 MAIZE EVENT	Endo maize HMG	CTP2_CP4EPSPS		P-35S	P-Rice actin	P-Ubi	T-nos	Cp4-epsps(1)	Cp4-epsps(2)	I-rAct1	
DAS-Ø15Ø7-1	TC1507 MAIZE EVENT	Endo maize HMG	P-Ubi_CRY	Pat_T-35S	P-35S	P-Ubi	T-35S	Cry1F	pat			

Legend:

Unique GMO identifiers
Detected
Not Detected
Not Tested

REMARKS

Endogenes not linked to GMOs: Endo canola FatA, Endo potato UGP, Endo wheat Wx-1,

Unexplained Constructs:

Unexplained Elements: Cry1A(b), Cry2Ab2, Barstar,

Figure 6.3 Visual output of S2 generated using the AM-SEQ pipeline. Column one and two indicate the identified GMOs and its unique GMO name. The detected targets (endogenous, elements and constructs) are marked in green and are associated with detected GMOs, targets in red represent not detected targets. Unexplained targets are presented in the remarks.

In S1, all the detected elements were explained by the identified GMOs. In S2, *Cry1Ab*, *Cry2Ab2*, *barstar* were unexplained with 1.55%, 1.27% and 0.29% of total mapped reads, respectively, which was above the established threshold. These targets were detected in qPCR with average Cq values of 39, 37 and 40. As a follow-up, methods for the EU authorised GMOs knowing to contain *Cry1Ab* and available in our laboratory were applied, i.e. methods for MON810, MON89034, BT176, T304, BT63, and BT11. None of these GMOs were detected in the sample. Similarly, *Cry2Ab2* containing GMOs were tested with qPCR and not detected, i.e. MON89034, MON15985, and MON7751. For *barstar*, RF1, RF2, and RF3 were tested negative in the additional qPCR screening.

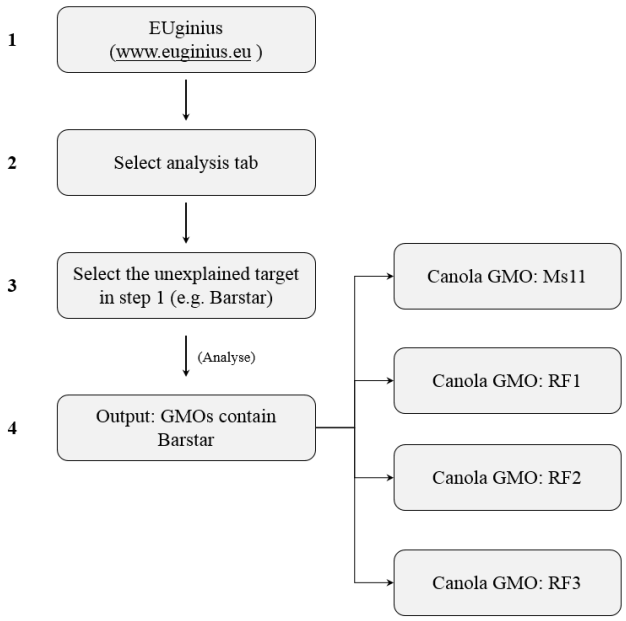


Figure 6.4 Schematic representation of the EUginius database (step 1 to 4) analysis process to identify all EU authorised and known unauthorised GMOs that contain the unexplained targets, which are found in the samples. The summary of EUginius output for GMOs containing Barstar is used as an example.

In S3, *Ctp4_cp4epsps and P-SSuAra* elements could not be associated to any of the detected GMOs in the sample and remains unexplained. The presence of the construct *ctp4-cp4epsps* was confirmed with qPCR. The associated GMO, GTS 40-3-2, for this construct was detected with an event specific qPCR with a Cq value of 35. However, event GTS 40-3-2 was not detected with NGS (0.0035% mapped reads, below the established threshold). The unexplained element *P-SSuAra* was detected above the established threshold, 0.065% of total mapped reads, and in the conformational qPCR analysis it was identified in one out of three separate reactions with a late Cq value of 38. The GMOs known to contain *P-SSuAra* were tested negative in the additional qPCR screening, i.e. MON87701, MON87751, MS1, MS8, RF1, RF2, and RF3.

In S4, *P-ubi* was the only unexplained element, which was detected in one out of two qPCR reactions with a high Cq value of 38. The element *P-ubi* is derived from *Zea mays*, and low-level presence of *Zea mays* in this sample was confirmed with *hmg* qPCR, while it was not detected with NGS.

In S5, six unexplained elements were identified, *Tg7, Bar, Cry1A.105, Cry2Ab2, P-TA29*, and *barstar*. The qPCR analysis of *Tg7*, *Bar*, *P-TA29* and *barstar* showed that these were detected in one out of two reactions with late Cq values, i.e. 37, 41, 36, and 39, respectively. As a follow-up, the in-house available event methods for GMOs containing these four targets were tested with the event qPCR method but none of these GMOs were detected in the sample, i.e. MS1, MS8, RF1, RF2, RF3, MON87460, DAS44406-6, and BT176. The elements *Cry2Ab2* and *Cry1A.105* had 1.7% and 2.8% of the total mapped reads with late Cq values of above 37 in qPCR. These two elements are present in MON89034, which was detected in the qPCR analysis with a high Cq value of 36 while it was not detected in NGS.

6.3.5 Comparison of the NGS-based GMO screening results with qPCR approach

The standard two-step qPCR screening results for the five feed samples showed that 16 to 34 targets out of 36 to 52 targets were detected across the feed samples (Additional file 6A: Table A.4). In general, most of the detected targets could be associated with identified GMOs. However, 6 elements and 1 construct could not be associated with any of the identified GMOs. These targets were identified with either a Cq value above 37 or were detected in only one of the replications of the qPCR assays (Additional file 6A: Table A.4). All these targets were considered to be low copy number targets from traces and no follow up was performed.

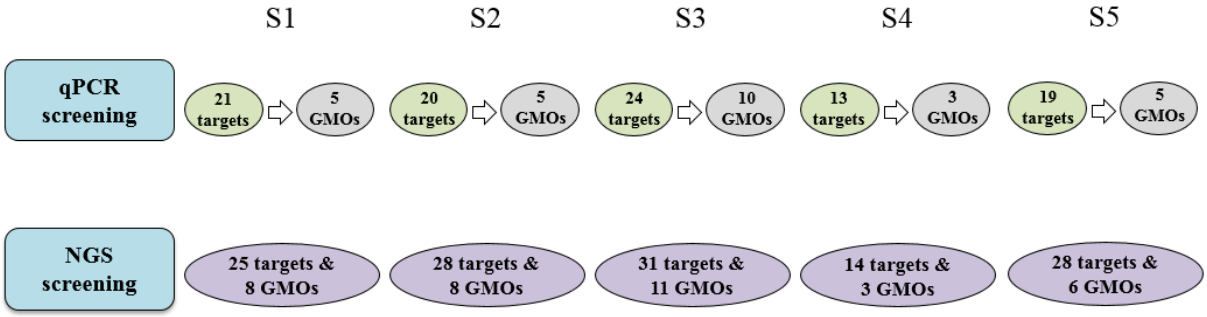


Figure 6.5 Comparing the number of targets identified in the two-step qPCR approach and wider screening as applied in the NGS-based screening approach. The number in the circles indicates the total number of targets detected in the samples for the respective screening strategies. The light green circle ◯ indicates step one in the qPCR approach (endogenous, elements and construct) and the light grey circle ◯ indicates step two in qPCR approach (events). The light purple circle ◯ indicates the one step wider screening approach.

The two-step qPCR screening approach could be considered as a less broad screening, with fewer GMO-related targets included compared to the NGS based screening. The qPCR screening results from the five feed samples were compared to the wider screening strategy used in the NGS-based approach (52 vs 96 targets). (Additional file 6D: Table D.1). This comparison showed that the same GMOs and related targets were identified to a large extent, however, in total 43 targets (endogenous, elements, construct and events) were additionally identified across the five feed samples in NGS due to broader screening strategy (Table 3). In all samples, a larger number of targets and GMOs were identified using wider screening approach in NGS compared to qPCR, except for S4 (Figure 6.5). To confirm the presence of the additional identified targets in the NGS-based screening, target-specific qPCRs were performed. All the additionally detected elements and constructs in the NGS-based screening, over all the samples, could be confirmed with qPCR assay. In total 10 GMOs were additionally detected across the samples between the qPCR screening strategy and the wider NGS screening strategy. In conformational experiments MON810 and NK603 were detected with qPCR in S1, GT3 40-3-2 in S2, NK603, and T25 in S3, all with a Cq value < 35 (Additional file 6C: Table C.1-C.3). The total mapped reads for these GMOs varied from 0.68 to 3.6%. (Additional file 6C: Table C.1-C.3). The other additionally detected GMOs, had Cq values > 38 or were not detected in all the reactions, i.e. GA21 in S1, GA21 and NK603 in S2 and MON810 and NK603 in S5. The total mapped reads for these low abundance GMOs varied between 0.02 to 0.07 %, except for MON810 in S5, which corresponded to 1.2 % of the mapped reads (Additional file 6C: Table C.1-C.5).

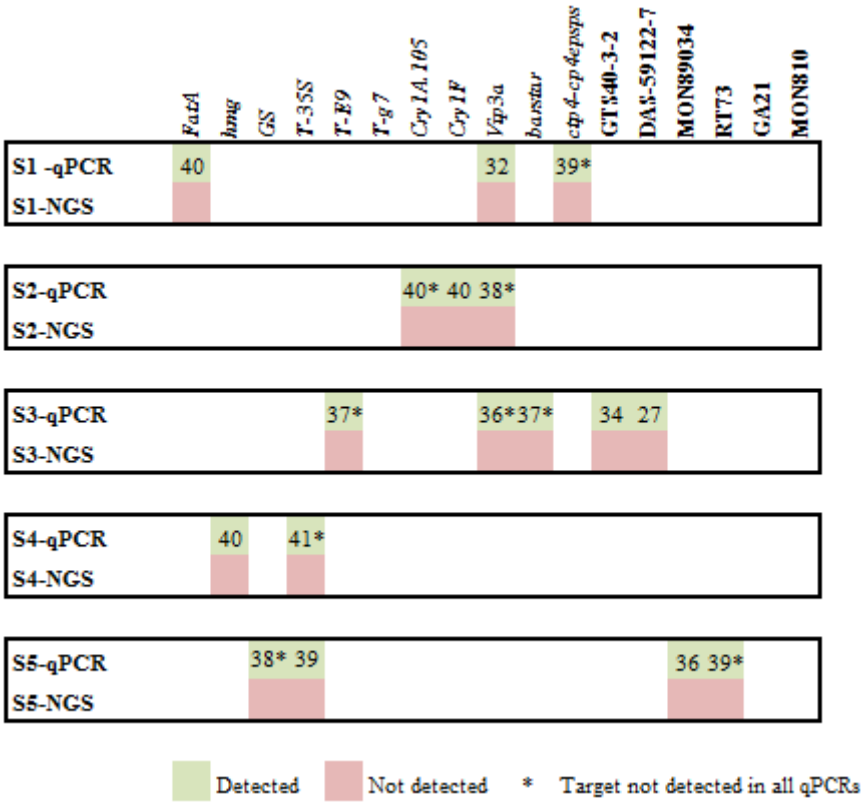


Figure 6.6 Discrepancies occurred between qPCR and NGS screening results. The number indicate the obtained Cq value

Reverse discrepancies were also observed between NGS-based and qPCR screening. In total 17 targets were detected in qPCR, but not in NGS (Figure 6.6). From these, 13 targets had Cq values above 36 and were not detected in all the qPCR reactions. The other four targets (S1: *Vip3a*, S2:GTS40-3-2 and DAS59122-7 and S5: MON89034) had low Cq values (≤ 36) in all four qPCR reactions, but were not detected in NGS. In NGS, no reads were identified for *Vip3a* and DAS59122-7. GTS40-3-2 and MON89034 had low numbers of mapped reads (220 and 407). DAS59122-7 event failed to amplify in the positive control and the five feed samples. However, in an independent experiment, the amplification of DAS59122-7 with HotStar *Taq* master mix was confirmed in the positive control and in S2.

6.4 Discussion

In the current study we present an NGS-based method for GMO detection and identification and compared the newly developed NGS approach with the two-step qPCR GMO screening method. To facilitate the NGS-based amplicon approach, an optimized NGS-based amplicon sequencing protocol was developed, as well as a bioinformatics pipeline (AM-SEQ) to analyse paired-end Illumina data. This comparison showed that the NGS-based screening approach has a similar sensitivity level as the qPCR screening. Additionally, the wider screening strategy applied in the NGS-based screening provided a complete overview of the content of the samples in a single analysis. The amount of additional targets identified in the NGS screening compared to the qPCR screening, indicates the advantage of a wider screening strategy in a situation where the variety in GMOs is increasing.

In the initial PCR master mix comparison, Qiagen HotStar *Taq* (PCR) master mix outperformed the routinely applied Diagenode (qPCR) master mix in efficient amplification of the target. Using

Qiagen Hotstar *Taq* master mix almost all targets of the positive control sample were successfully amplified, as was visualized by gel-electrophoresis. Subsequently, Qiagen Hotstar *Taq* master mix was used to amplify all 96 targets with DNA obtained from the five feed samples. The obtained amplicons from positive control and feed samples were successfully sequenced with PE HiSeq 150 Illumina technology. In the present study, PE 150 HiSeq Illumina technology was the preferred NGS platform of choice, because of very low error rates (< 0.1%). Alternatively the currently available Illumina MiSeq technology could also be applied, which can cover amplicon lengths > 250 bp [275]. The AM-SEQ pipeline for the NGS data analysis as presented in this paper will be applicable for MiSeq NGS data as well.

The obtained NGS data from the positive control was used as input for the AM-SEQ pipeline to generate reference sequences. In total, 79 new reference sequences for targets (endogenous, elements, construct and events) were generated and made publicly available in the EUginius database (www.euginius.eu). For the feed samples, on average, more than 80% of the obtained data were of good quality and matched with a reference sequence in the local database. In the NGS data from the samples we also identified primer selected reads that did not map against the reference sequence containing the probe sequence. These unmapped reads clusters have been manually checked, but biological relevance could not be verified due to the lack of probe sequences. It is possible, however, that these unmapped clusters indicate genuine variations in the genetic constructs of the different varieties that have not been reported as yet, or indeed may represent sequences of unknown GMOs that have not yet been documented. However, the results may also be explained by non-specific amplification in the PCR.

To avoid false positive results, a threshold for detection (0.01% of mapped reads) was established for the NGS datasets based on a comparison of qPCR and NGS results. As a consequence, two GMO targets (*cp4-epsps*(2) and MON89788) from the positive control with a low percentage of mapped reads (0.002 %) were scored as negative. For these two targets, a faint band was observed in the gel image of the positive control PCR (Appendices A: Figure. A.2) and only 281 (*cp4-epsps*(2) and 261 (MON89788) pseudo-reads were primer selected (Appendices C: Table A.6). When comparing mapped read counts in the NGS data from 0.01% to 1% of reads, as depicted in figure 6.2, suggested a different nature of amplification for targets with these read counts, i.e. noise *vs* true amplification. The high occurrence of low assigned read counts to targets in the NGS data indicates that this may result from either neighbouring cluster overlapping or cross-contamination of the indexed libraries. The latter kind of cross-contamination may be overcome by using double indexing on the Illumina platform [276]. However, with the currently widely applied single indexing in Illumina, these contaminations cannot be prevented and may result in false-positive identification. By establishing a cut-off for the NGS dataset like reported in other studies [181,198], true amplification (targets with high reads counts) and cross-contamination (targets with low reads counts) can be distinguished. Application of the threshold may lead to false negatives, but in these cases it seems likely that the amounts will be below values that would have been identified by current qPCR strategies, and would therefore not lead to decreased sensitivity in the screening as such.

An alternative screening approach should at least perform as well as established methods. For that reason, the results of the NGS-based strategy were compared with the qPCR method, which is the current golden standard for GMO screening and identification. This comparison showed that 92% of the targets were commonly identified in the qPCR and NGS-based screening. For 17 targets a discrepancy was observed in that these targets were detected in qPCR but not in NGS. Of these, 13 cases may likely be explained by subsampling effects i.e. the weak signal in the qPCR strategy (Cq value above 36 and not detected in all the replication reactions) suggests that the levels were low and therefore a chance effect may determine a positive signal. Correspondingly, the number of mapped reads for these 13 targets were in the range of 0-450, which is below the threshold for detection (< 0.01%). Of the 17 targets not detected in NGS, four targets had a Cq value of ≤ 36 and were not detected in the NGS

analysis. The event DAS59122-7 was one of the targets that was not detected in NGS and also failed to amplify in the positive control PCR. However, to confirm that amplification of DAS59122-7 event is possible, the PCR was repeated with HotStar *Taq* master mix for the positive control and S2, which resulted in a clear amplification in both cases. These observations indicate that in the initial experiment probably the master mix preparation was not according to protocol. Another target that was not detected in the NGS data in case of S1 was the GMO element *Vip3a.* This was the only target for which failure could not be explained, a problem may have occurred during the library preparation The last two targets (GTS40-3-2 and MON89034) had a low number of mapped reads (407 and 220) and a Cq value of 34 and 36 respectively, which is considered to be at the border of detection. The 0.1% positive control of these GMOs had a Cq value of 34 and 35, respectively. EU GMO regulations stipulate that food/feed products containing > 0.9%, per ingredient, need to be labelled as such (https://ec.europa.eu/food/plant/gmo/traceability_labelling_en). Moreover, a limited set of GMO events that are in the approval pipeline, but not yet authorised are allowed in feed products up to the level of 0.1% [277]. Considering these GMO labelling stipulations, the high Cq value combined with the low read count in the NGS analysis indicated that these targets were most likely below the given percentages. While in some cases NGS thus failed to detect low abundant targets, on the other hand, across the samples 23 targets were detected in NGS-based screening (above 0.01% threshold), but all these targets had a high Cq value above 36, or were not detected in all of the reactions in the qPCR analysis. These results highlight that a target with a low copy number can be either detected or not-detected in the NGS-based screening, as in the qPCR screening strategy, which fits the occurrence of a subsampling effect. To reduce a subsampling effect in NGS-based screening it is therefore advocated to perform at least two NGS screenings on an individual sample as is the standard procedure in the qPCR approach.

In the NGS-based screening, 12 unexplained targets were identified. Three targets were explained by the presence of GTS 40-3-2 and MON89034 in samples S3 and S5. These two GMOs targets were not detected in the NGS-based screening; in qPCR these two GMOs were detected with a Cq value above 35, which is at the border of detection. Another unexplained element *P-ubi* in S4 could be explained by the presence of *Zea mays* in this sample, which was confirmed in an additional *hmg* qPCR, also here, the HMG target was a false negative in the NGS screening. The remaining eight unexplained targets (S2: *Cry1A(b)*, *Cry2Ab2* and *barstar*, S3: *P-SSuAra*, S5: *Tg7*, *Bar*, *P-TA29*, and *barstar*) had read counts above the threshold. However, the corresponding Cq value for these targets in the qPCR screening was above 36 and the targets were not detected in all qPCR assays. This indicates a low presence of these targets in the samples and a related reduced chance of identification. This illustrates that NGS is not a quantitative method and while a positive identification in NGS is a reflection of the presence of a target, the read counts are probably not a reliable reflection of the abundance level of this target. An explanation for the presence of these unexplained targets could be traces of GMOs, or perhaps more likely, these targets may indicate the presence of the donor organism in trace amounts, below the limit of detection. For example, *P-SSuAra, P-TA29*, and *barstar* are targets obtained from the donor organism *Arabidopsis thaliana*, tobacco, and *Bacillus amyloliquefaciens* respectively. Any detected unexplained GMO-related target with an Cq value ≤ 35 may indicate the presence of an unknown or unauthorised GMO that was not included in the analysis, or for which no method is available yet. In cases were the Cq value ≤ 35, further analysis needs to be performed to explain the respective targets. Genome walking approaches coupled to NGS can be applied to identify the adjacent sequences in these cases [271,272]. The genome walking approach could help to determine the flanking region of the identified unexplained targets, which could eventually lead to identification of the UGMOs or donor

organism. In this study, the wider screening including 96 targets, which was applied in the NGS approach, and basic screening, applied in two-step qPCR approach, were compared side by side in figure 6.5. Applying the wider screening strategy, 43 GMO related targets and GMOs (2 endogenous, 19 elements 12 constructs and 10 GMOs) were additionally identified compared to the qPCR screening as highlighted in table 3. Especially, the identification of additional GMOs indicates the necessity of a more extensive screening in an NGS-based approach or in qPCR screening, as it is clear that the presence of some GMOs may mask the presence of others, and this may include also UGMOs.

Currently, 411 GM crops from 28 plant species are authorised for their use as either food/feed or for cultivation (http://www.isaaa.org/gmapprovaldatabase/), which is a fourfold increase in the last decade [278] and more GMOs are in the pipeline for approval [45]. The steady increase in GMOs entering the consumer market raises traceability issues, as it will be necessary to continuously increase the number of screening elements, otherwise, the presence of (some) authorised GMOs may increasingly mask the presence of additional GMOs, that may include UGMOs. The present study underlined this effect as the basic screening resulted in the identification of numerous GMOs, yet the broader NGS screening showed that others were present as well, which had not been identified based on the current qPCR strategy. The NGS based screening approach has the potential to be more versatile and flexible: as the sequence information of amplicons will be the direct proof for the detection, there is no need for the validation of target-specific methods, and thus not for the related reference materials, as the respective amplicons can be confirmed by subsequent DNA sequence analysis. This is a major asset in the case of the detection of UGMOs, where generally no reference material will be available. Furthermore, the identified UGMO-related sequences can be used as starting point in a genome walking approach to obtain the adjacent sequence information necessary to identify the UGMO. The present study focused on the application in the field of the enforcement of current GMO legislation, but it is clear that the strategy may find wider application to screening complex products for the presence of ingredients or elements that may not be regarded as safe for human or animal consumers or the environment.

6.5 Conclusions

In this research, we developed a next generation sequencing (NGS)-based GMO screening approach, covering 96 GMO targets, and a data analysis pipeline to detect and identify GMOs in complex food or feed samples. Initially, we were able to generate 79 new GMO related amplicon reference sequences and added these to the public database EUginius. Then we compared the developed NGS-based GMO screening approach with the qPCR-based GMO screening, currently the golden standard in the area of GMO analysis. The comparison indicated that for highly abundant GMOs and targets both approaches had a similar level of sensitivity. However, for GMOs and targets present in low concentrations, the detection between the NGS-based screening and qPCR approach showed discrepancies. Furthermore, due to the extensive screening in the NGS-based strategy, more GMOs and related targets were identified compared to the standard qPCR screening. Additional targets identified urges for the importance of a wider screening strategy for GMO identification and detection. The current study proves the applicability of NGS as an accurate and reliable screening method for GMOs and its potential to improve current screening methods for, especially, the presence of unauthorised new plant varieties in complex food or feed samples.

6.6 Additional files (available with the publication)

Additional file 6A: Table A.1 GMO specific targets (endogenous, elements, constructs and events) that are analysed in the current approach. A detail information about the source of the reference material, primer name, sequence information of the forward and reverse primer and concentration of the primer and references are specified. **Table A2** Individual and average number of Illumina Hiseq reads, Number and percentage of QC pseudo-reads, primer selected pseudo-reads and mapped pseudo-reads to references are generated per sample. **Table A3** Reference sequences obtained from public databases. **Table A4** Targets tested and identified in the two-step qPCR GMO screening. **Figure A1** Comparison of amplification efficiency between two different master mixes. PCR amplicons that were generated from 3 samples (2 feed samples and positive control) with four targets (*hmg*, *p35S*, *tNOS* and *Cry1AB/Ac*) using two different master mix (Diagenode and HotStar *Taq*). The feed sample and positive control is represented by lane 1 2 and 3. Lane 4 and M represent the negative control and 10 bp DNA ladder. **Figure A2** Image of the PCR amplicons that were generated from positive control with 96 targets. P-MFV(2) amplicon was not loaded in the gel. Lane 1-96 represent an unique target and name of the targets are presented in the table below, 10 bp and 100 bp DNA ladder are represented as M1 and M. **Figure A3** The identified targets (endogenous, constructs and elements) in a samples that are associated with detected GMOs. The targets in green are the identified targets that can be associated with detected GMOs, the target in red represent that the target is not detected. The target detected in NGS and cannot be associated with the detected GMOs are presented in the remarks.

Additional file 6B: Table B1 Analysed GMO endogenous, GMO source, method name and reference sequence. **Table B2** Analysed GMO elements, GMO source, and method name and reference sequence. **Table B3** Analysed GMO constructs, GMO source, method name and reference sequence. **Table B4** Analysed GMO events, GMO source, and method name and reference sequence.

Additional file 6C: Table C1 Targets identified in sample 1 with qPCR and NGS analysis. **Table C2** Targets identified in sample 2 with qPCR and NGS analysis. **Table C3** Targets identified in sample 3 with qPCR and NGS analysis. **Table C4** Targets identified in sample 4 with qPCR and NGS analysis. **Table C5** Targets identified in sample 5 with qPCR and NGS analysis. **Table C6** Targets identified in positive control with NGS analysis.

Additional file 6D: Table D1 Targets identified in the NGS and qPCR across the samples

Chapter 7

ALF: a strategy for identification of unauthorised GMOs in complex mixtures by a GW-NGS method and dedicated bioinformatics analysis

This chapter was published as: Košir AB, **Arulandhu AJ**, Voorhuijzen MM, Xiao H, Hagelaar R, Staats M, Costessi A, Žel J, Kok EJ, van Dijk JP. "ALF: a strategy for identification of unauthorised GMOs in complex mixtures by a GW-NGS method and dedicated bioinformatics analysis". *Scientific Report* 2017; 7(1): 14155.

Abstract

The majority of feed products in industrialised countries contains materials derived from genetically modified organisms (GMOs). In parallel, the number of reports of unauthorised GMOs (UGMOs) is gradually increasing. There is a lack of specific detection methods for UGMOs, due to the absence of detailed sequence information and reference materials. In this research, an adapted genome walking approach was developed, called ALF: Amplification of Linearly-enriched Fragments. Coupling of ALF to NGS aims for simultaneous detection and identification of all GMOs, including UGMOs, in one sample, in a single analysis. The ALF approach was assessed on a mixture made of DNA extracts from four reference materials, in an uneven distribution, mimicking a real life situation. The complete insert and genomic flanking regions were known for three of the included GMO events, while for MON15985 only partial sequence information was available. Combined with a known organisation of elements, this GMO served as a model for a UGMO. We successfully identified sequences matching with this organisation of elements serving as proof of principle for ALF as new UGMO detection strategy. Additionally, this study provides a first outline of an automated, web-based analysis pipeline for identification of UGMOs containing known GM elements.

Keywords: GMOs, UGMOs, genome walking approach, NGS

7.1 Introduction

Nowadays, the vast majority of feed products in industrialised countries contain materials that are derived from genetically modified organisms (GMOs). For food products the situation is still very different, primarily due to the lack of public acceptance in some countries and amongst specific groups of consumers, but here also a slow trend can be observed of increased use of products derived from GMOs. From the Food and Agriculture Organization of the United Nations (FAO) report of 2014 [279], it can furthermore be seen that the number of incidents with unauthorised GMOs (UGMOs) is gradually increasing. The reported incidents related primarily to identified UGMOs, that had received market approval in other countries. However, in a growing number of cases the incidents related to unknown UGMOs where the mere combination of the crop at hand and the detected GMO elements were deemed sufficient to take action. It should be stressed that so far no genetically modified (GM) crops have been identified to have adverse effects on humans, animals or the environment [280,281]. Nevertheless, in the light of the rapidly expanding diversity of GMOs in experimental settings, it seems prudent to have the methodologies in place to detect and identify UGMOs, including the as yet unknown ones. There is a general lack of specific detection methods for UGMOs, usually due to the absence of detailed sequence information and reference materials.

In recent years, a number of strategies have been developed that focus on the detection of UGMOs. The strategy generally used is based on the screening of samples for a range of GMO elements. Subsequently, the presence of the observed elements is compared with the confirmed presence of authorised GMOs or known UGMOs in the same sample. A mismatch between observed elements and observed GMOs indicates the potential presence of a UGMO. Clearly, only GMOs for which adequate methods for identification are available [19,186,217,233,282-289] can be taken into account here. Additional analyses are necessary to determine whether the identified GMO element is indeed linked to an unknown, unauthorised GMO, or rather to, for instance, the native organism of the element. Therefore, in these cases, additional experiments will be required, that are currently usually based on variants of gene walking (GW) strategies where the unexplained elements can be used as a starting point for GW to obtain adjacent sequence information. This sequence information could lead to the identification of a specific UGMO, especially when the sequence information stretches into the flanking genomic region of the GMO insert. This information can ultimately, if deemed necessary, be used to develop a specific method for the identified UGMO.

In the last two decades several GW approaches have been developed and modified for application in the GMO field. Examples are Long template-Rapid Amplification of Genomic DNA Ends (LT-RADE) [228,251], SiteFinding-PCR [219,290], APAgene GOLD Genome Walking Kit [254,271,291], A-T linker adapter PCR [232], Randomly broken fragment PCR (RBF-PCR) [231], Locus-finding PCR (LF PCR) [230] and Loop-linker PCR [292] recently reviewed in Arulandhu et al. [23]. Most of the approaches have been initially evaluated in a pure GMO sample. The APAgene GOLD Genome Walking Kit [254,271,291] was also tested in low GM percentages in combination with fragment cloning and Sanger sequencing [254,291]. Only very recently this method was also succesfully applied in combination with NGS and more complex GMO mixtures [271].

Any enrichment approach applied to a mixed sample may yield a mixture of sequences, because even a single common element may have several adjacent sequences, from different GMOs in the mixture. Next Generation Sequencing (NGS) is ideally suited for sequencing all amplified fragments in a mixture. In the development of a new GW approach we chose the PacBio RSII as NGS platform, as long reads of single molecules minimize the chance of artefacts, frequently referred to as chimeric

sequences, due to incorrect assembly of short sequence reads [219]. Target molecules can be sequenced multiple times or 'passes', as hairpin adaptors are ligated to both ends, creating a circular DNA molecule, serving as a polymerase template without an end. The initial output is a polymerase read with multiple copies of the target sequence interspaced by adaptor sequences. When such a subread is present at least four times, a consensus sequence is made after removing the adaptor sequences. This is the circular consensus sequence (CCS) output of the PacBio RS II [293] (Additional file 7A; Figure S1). With 80-85% accuracy [294], subreads have a relatively high error rate[294-297]. As the errors are random, with the error pattern of 10% insertions and 5% deletions[296], the accuracy of the CCS read is higher than that of the subread, and it increases with the number of subreads. SiteFinding-PCR [219,290], and APAgene GOLD Genome Walking Kit [271], are the only enrichment approaches that have been published in combination with NGS analysis to detect and identify GMOs/UGMOs. However, given the random start and two nested PCRs of SiteFinding-PCR, and two semi nested PCRs and the use of DRT primers in APAgene GOLD Genome Walking Kit, these approaches might be sensitive to contamination. Therefore, we adapted the LT-RADE method with a biotinylation primer mediated clean-up prior to PCR and only one round of PCR. Coupling this approach to NGS should allow for simultaneous detection and identification of all GMOs in a sample, including UGMOs, in a single analysis.

In this research, an adapted GW approach was developed, hence referred to as Amplification of Linearly-enriched Fragments (ALF), encompassing the advantages of the available methods, particularly the rapid amplification of cDNA ends (RACE)[298] approach and the use of biotinylated primers. The efficacy of the ALF approach was assessed on a complex mixture consisting of four GMOs: maize MON810, MON89034, MON88017 and cotton MON15985, using the GMO elements p35S promoter and tNOS terminator as starting-points for the elongation. After amplification of the elongated fragments, all obtained fragments were sequenced using the PacBio RSII platform. The complete insert and genomic flanking regions were known for the first three GM maize events, while for the MON15985 only partial sequence information was available; combined with a known organisation of elements, this GMO served as a model for a UGMO in this set-up. In the present study, the ALF approach is presented and applied on the mixture of all four GMOs. The first outline of an automated, web-based analysis pipeline for identification of UGMOs containing known GM elements was set, and the results of the new strategy are evaluated in the light of the necessity to have adequate methodology in place to identify GMOs for which limited information is available.

7.2 Results

7.2.1 Amplification of Linearly-enriched Fragments (ALF)

A protocol was developed for the identification of unknown GMO-related sequences starting from known GMO elements, called ALF: Amplification of Linearly-enriched Fragments (Figure 7.1).The success of the ALF protocol was determined both in terms of quantity and length of the molecules of interest. In the reference materials used, upstream and downstream GMO elements were known. By performing qPCR analysis prior to, and at different steps of the procedure, relative increase in target fragments was estimated.

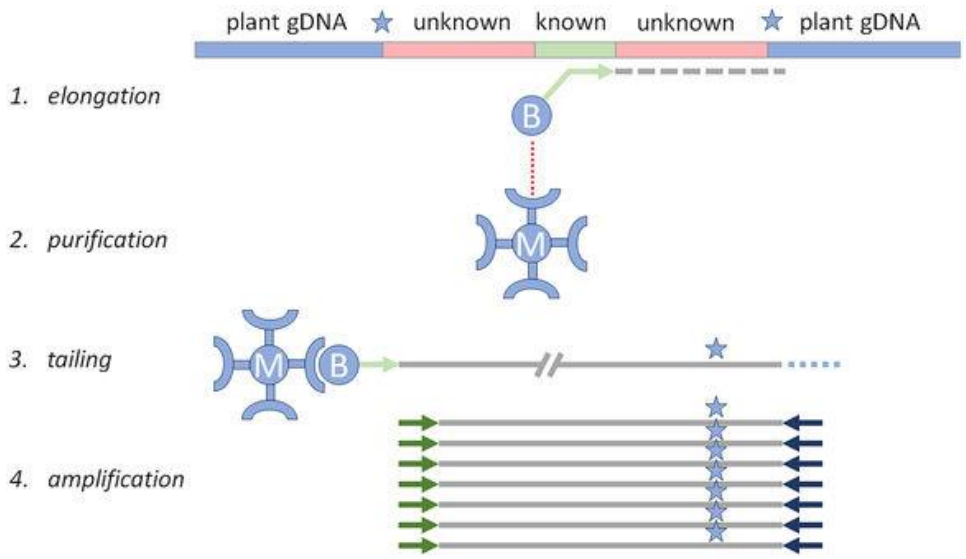


Figure 7.1 Schematic overview of the ALF procedure. The procedure yields dsDNA molecules after initial primer elongation, of various length, the longest of which will pass a construct-genome boundary (star), leading to GMO identification.

Targets for the qPCR analysis were chosen at various distances from the start of the LE (Figure 7.2). To increase the reliability of the quantification outcomes, Joint Research Centre (JRC)-validated qPCR methods were used in combination with certified reference material for >99.05% MON88107. qPCRs were performed before the procedure in the starting material (SM), after LE, and after snPCR. Apart from the qPCRs targeting the GM elements of interest, also the abundance of the maize endogenous high mobility group (*hmg*) gene was evaluated with qPCR, to monitor the removal of genomic DNA (Figure 7.3). After the LE step, a slight increase was observed for targets of interest, more prominently for targets closer to the starting point of LE. As expected, the genomic DNA amount was very similar before and after LE. After snPCR, a large increase in targets of interest was observed, again more prominent for target sequences close to the LE start, indicating a size-dependent enrichment. Loss of gDNA beyond detection was observed for all targets after the snPCR step (Figure 7.3).

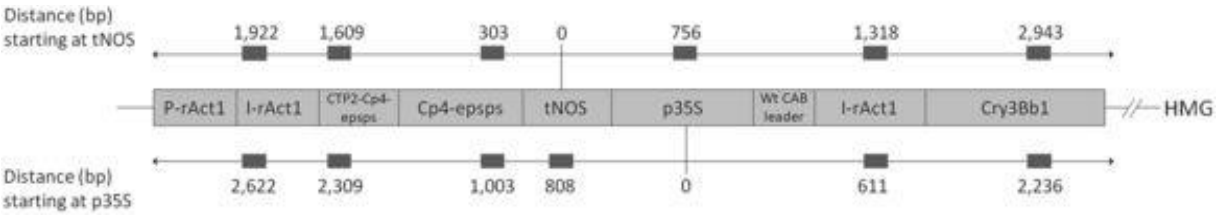


Figure 7.2 Distances of several qPCR targets from the enrichment starting points in the MON88107 GMO, used for evaluation of the ALF protocol. The light grey boxes indicate the different elements in the GMO. The upper and lower line indicate the enrichment for tNOS (upper) and p35S (lower), 0 indicates the starting points of enrichment and arrows indicate the direction of enrichment (upstream or downstream). The dark grey boxes indicate qPCR targets used for evaluation of the ALF protocol and distance of the qPCR targets from the starting points.

The relative amount of the different targets was expressed as the relative Cq value and calculated by subtracting the Cq after LE or snPCR from the Cq of the starting material. The largest increase in relative Cq was observed for GW starting from tNOS in the upstream direction, for the element closest to the LE start, Cp4-epsps (Figure 7.3, panel B). The fold change for the different targets

upstream of tNOS was estimated using a ΔCq approach, assuming a qPCR efficiency of 2. The Cq for the Cp4-epsps element, 303 nt upstream of the tNOS LE primer, was 30.3 in the starting material. During development of the method, we found it to be necessary to dilute the snPCR material 100-fold prior to qPCR evaluation, in order to obtain Cq values higher than five. Even in the diluted snPCR material the Cq value for the Cp4-epsps element was as low as 9.6. This was converted to the theoretical value of 2.96, by subtracting the 6.64 cycles from this Cq value, corresponding to a 100-fold dilution. The ALF related ΔCq for the Cp4-epsps element was therefore calculated to be 27.3 cycles, indicating an estimated increase of 165 million-fold of fragments at least 303 nt long. A fraction of these molecules was at least 1609 nt long, evidenced by the increase in relative Cq value for the Ctp2-cp4epsps element, located 1609 nt upstream of the tNOS LE primer. This fraction was estimated to be enriched 30 thousand fold. Likewise, a subfraction of these fragments was at least 1922 nt long, based on relative rAct1 Cq increase. This fraction was estimated to be enriched 3.5 thousand fold.

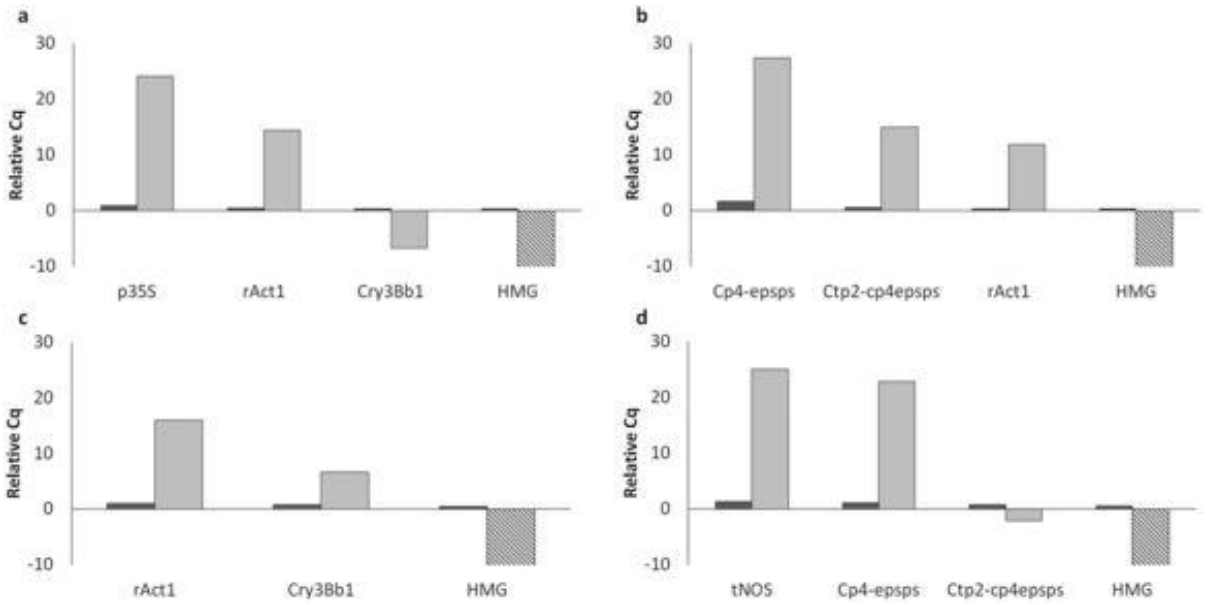


Figure 7.3 Length-dependent increase of specific targets and loss of genomic background shown with qPCR. In all cases the amplicon closest to the element targeted for linear enrichment showed the highest increase in signal, expressed as the relative Cq value and calculated by subtracting the Cq after LE (dark bars) or snPCR (light bars) from the Cq of the starting material. This means the relative Cq value for each starting point is zero. Panel A: Enrichment for tNOS downstream. B. Enrichment for tNOS upstream. C. Enrichment for p35S downstream. D. Enrichment for p35S upstream. In all panels, the dashed bars for HMG indicate a reduction beyond detection.

7.2.2 PacBio sequencing of a GMO mixture after ALF

A DNA-based mixture was made using DNA extracts from four reference materials, in an uneven distribution, mimicking a real life situation, where one or several GMOs may be present in a much lower concentration than others in the same mixed sample. DNA of MON810 from a co-existence field trial [299] was used as the most abundant GMO and was present in 97%. Of the other three GMO DNA isolations, 1% relative weight was added for each (Table 7.1). Two maize lines were used, MON89034 and MON88017, and one cotton line, MON15985. This mixture was subjected to the ALF protocol in four separate reactions for the different elements and directions: the p35S promoter and the tNOS terminator, both in upstream and downstream direction. The reactions were pooled and sequenced on the PacBio RSII platform. The results of sequencing were 411 CCS reads (length distribution is shown in Additional file 7A; Figure S2).

Table 7.1 Details of reference materials [300].

GMO	Content in mixture (%)	Estimated copy number	Supplier	Code	Description	Element order*	Donor organism
MON810	97	~40.000	Field trial [299]	In-house	50% MON810, ground corn	P-e35S I-hsp70 CS-cry1Ab	Cauliflower mosaic virus *Zea mays* *Bacillus thuringiensis*
MON89034	1	~400	AOCS	0906-E	>99.42% MON89034, ground corn	V-LB P-e35S L-cab I-1_act1 CS-cry1A_105 T-hsp17_3 P-FMV I-hsp70 I-1_rbcS CS-cry2Ab2 T-nos V-RB	*Agrobacterium tumefaciens* Cauliflower mosaic virus *Triticum aestivum* *Oryza sativa* Synthetic *Triticum aestivum* Figwort mosaic virus *Zea mays* *Zea mays* *Bacillus thuringiensis ssp. Kurstaki* *Agrobacterium tumefaciens* *Agrobacterium tumefaciens*
MON88017	1	~400	AOCS	0406-D	>99.05% MON88017, ground corn	P-act1 I-1_act1 TP-ctp CS-CP4epsps T-nos P-e35S L-cab I-1_act1 CS-cry3Bb1 T-hsp17_3	*Oryza sativa* *Oryza sativa* *Arabidopsis thaliana* *Agrobacterium tumefaciens ssp. CP4* *Agrobacterium tumefaciens* Cauliflower mosaic virus *Triticum aestivum* *Oryza sativa* *Bacillus thuringiensis ssp. Kumamotoensis* *Triticum aestivum*
MON15985	1	~400	AOCS	0804-D	>98.45 % Bollgard II cotton, ground cotton seed	P-e35S L-hsp70 TP-ctp CS-cry2Ab2 T-nos P-e35S CS-cry1Ac T-7Salpha P-35S CS-nptII T-nos P-e35S CS-uidA T-nos CS-aadA	Cauliflower mosaic virus Petunia hybrida *Arabidopsis thaliana* *Bacillus thuringiensis ssp. Kurstaki* *Agrobacterium tumefaciens* Cauliflower mosaic virus *Bacillus thuringiensis ssp. Kurstaki* *Glycine max* *Cauliflower mosaic virus* *Escherichia coli* *Agrobacterium tumefaciens* Cauliflower mosaic virus *Escherichia coli* *Agrobacterium tumefaciens* *Escherichia coli*

**Initial capital characters: P = promoter, I = intron, CS =coding sequence, V = vector, LB = left border, L = leader, T = terminator, RB = right border, TP = transit peptide.*

7.2.3 Sequence analysis: database construction

To analyse the CCS reads gained after PacBio RSII sequencing three databases were constructed: an event, a Constructs-And-Flanks (CAF) and an element database. The event database (Additional file 7A; Data S1) consisted of the most likely amplicon sequences of the event specific qPCR methods used for

quantification of events (MON810, MON88017, MON89034 and MON15985) in the sample. Sequences of primers and probes, together with the number of unknown nucleotides between them, were taken from the EU Database of Reference Methods for GMO Analysis (GMOMETHOD) (http://gmo-crl.jrc.ec.europa.eu/gmomethods/) and from the method validation reports [301-304]. These sequences were queried against the NCBI patent database (Table 7.2). After inspection, the top hit for each of them was added to the event database.

Table 7.2 Event database sequences and corresponding NCBI accession numbers.

GM event	Event sequences	Best hit accession number
MON810	TCGAAGGACGAAGGACTCTAACGT***TT***AACATCCTTTGCCATTGCCCAGC***TATCTGTCACTTTATTGTG***AAGATAGTGGAAAAGGAAGGTGGC*	AR490568
MON89034	TTCTCCATATTGACCATCATACTCATT***GC***ATCCCCGGAAATTATGTT***T***TTTAAAAACCACGGTATTATAGATACCG	FV532179
MON88017	GAGCAGGACCTGCAGAAGCT***AGCTTGATGGGGATCAGATTGTCGTT***TCCCGCCTTCAGTTTAAACAGAGTCGGGT***T***TGGATGGTCAACTCCGGCA	DJ058152 / DJ058151
MON15985	GTTACTAGATCGGGGATATCC***CCGGGGCGG***CCGCTCTAGAACTAGTGGATCTGCACTGAAA***A***TCCCATCCATTTAGCAACCTT	EA135634

** nucleotides in italics denote the string of unknown nucleotides between primer and probe sequences in the search template.*

The CAF database consisted of the available construct sequences and flanking plant genomic regions of individual events in the sample (Additional file 7A; Data S2). In case of MON15985 only partial 3′ and 5′ insert sequences with corresponding flanking genomic sequences were available. For MON810 two reference sequences were available: sequence JQ406879, covering the 5′ flank and the p35S promoter, and AY326434, covering the insert from the p35S promoter onwards to the 3′ flank. These sequences showed a 75-nucleotide overlap. In the initial workflow both sequences were in the CAF database, but the merged MON810 sequence of JQ406879 and AY326434, named RIKILT20151130, was used in the final workflow (Additional file 7A; Data S2).

The third, element database, consisted of element sequences present in the experimental mixture. All element sequences from individual lines were gathered. For MON810, MON89034 and MON88017 complete reference sequences were already known and annotated. These reference sequences were divided in elements based on their annotation, for which a General Feature Format (GFF) file was constructed (Additional file 7A; Data S3). For MON15985, for which only the order of elements was known, generic sequences for these elements were taken from the NCBI patent database. Initially, this database contained all element sequences (Additional file 7A; Data S4). To reduce redundancy, the final version contained only the longest sequence of elements with several entries in the first version (Additional file 7A; Data S5). All designed databases were transformed into BLAST+ databases [305] and imported into the open, web-based analysis platform Galaxy [167].

7.2.4 Building the workflow for GMO sequence identification

The workflow for GMO sequence identification in the samples was constructed in Galaxy to answer the following questions: (1) is there any potential evidence for UGMOs, (2) can known GMOs be identified and (3) can a list of potential GMOs be prepared. Considering these questions, a workflow with six main steps was constructed (Figure 7.4 and Additional file 7A; Figure S3).

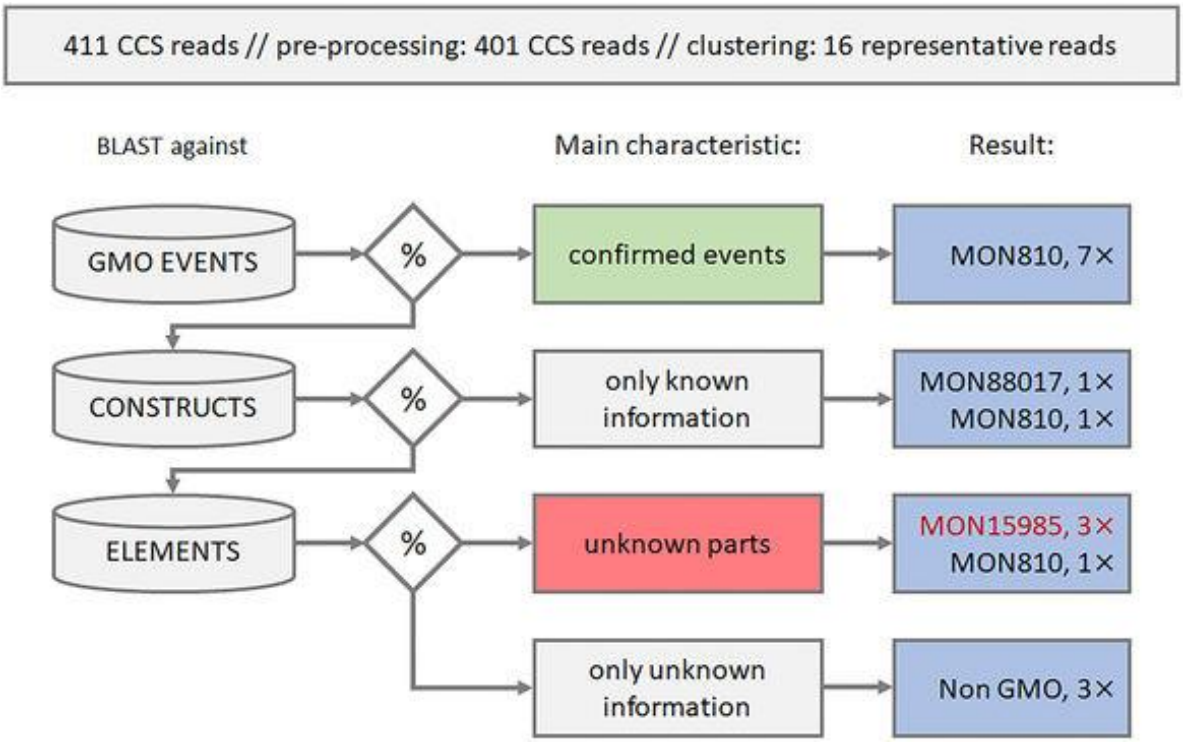


Figure 7.4 Schematic overview of the data analysis workflow. CCS reads are first processed to reduce noise and redundancy. Representative reads are then grouped in relevant bins based on homology with sequences in different databases, using blast. In the results, reads likely to be related to MON15985 are marked in red as this authorised GMO with incomplete sequence information served as a model UGMO in this study.

Step 1 consisted of a trimming and length selection; removing artificial sequences from the CCS reads and subsequently discarding short reads. The poly-dC tail and AAP adaptor sequences were removed and a length filtering was applied at >50 nt. All 411 CCS reads were used as the workflow input, a total of 401 reads proceeded to step 2.

Step 2 was selection of reads containing an enrichment primer, i.e. the nested tNOS and p35S LE primers, to make sure only specific sequences were kept. A total of 383 sequences passed the LE primer filtering; 368 were enriched for the p35S and 15 for the tNOS.

Step 3 was a clustering, using UPARSE [306], of CCS reads to further reduce the number of sequences. A three-step clustering approach was used. In this approach, sequences were dereplicated, sorted by abundance and clustered. Due to the nature of PacBio sequencing a CCS read either starts or ends with the enrichment target element, in this case the tNOS or p35S. An additional step of reverse complementing was applied to all the reads containing the enrichment target element on the 3′ end before dereplication, since the clustering algorithm aligned sequences only in plus/plus manner. With a minimum cluster identity, also termed cluster radius, of 0.97, 383 CCS reads were clustered in 16 clusters. For each cluster, the longest CCS read was used as the representative sequence, and termed cluster representative CCS read (crCCS read).

Step 4 was a Megablast of the crCCS read against the event database. Seven of the 16 crCCS reads showed a match to a sequence in the database with an identity of 95% or higher and a minimum

coverage of 97%. The main characteristic of these reads was that they included a *confirmed event*, and were not analysed further. The nine remaining crCCS reads were further processed in step 5.

Step 5 was a Megablast of the remaining crCCS reads against the CAF database. Two out of nine crCCS reads aligned completely to a CAF database sequence, i.e. full query coverage, and were annotated using the intersect interval tool (part of BEDtools [307]). The main characteristic of these annotated crCCS reads was the establishment that they contained *only known information*.

Step 6 was a Megablast of the remaining seven crCCS reads against the element database, aiming at identification of reads that contain both known and unknown sequence, as is expected for UGMOs. The crCCS reads were divided into two bins, also based on their main characteristic, either containing *unknown parts*, or containing *only unknown information*. The reads containing unknown parts also contained homologies to known elements, by the definition of this workflow. The Blast output of these reads against the element database was sorted in a way that the top hit was the one with the longest sequence alignment. The alignments with different elements were ordered according to their position in the CCS read, from 5′ to 3′. All four output bins, *confirmed events, only known information, unknown parts,* and *only unknown information* were imported in an excel template, constructed to give a user-friendly output (Additional file 7B; spreadsheet 1).

7.2.5 Connecting sequences with the experimental set-up

Relating to the three main questions (1) is there any potential evidence for UGMOs, (2) can known GMOs be identified and (3) can a list of potential GMOs be prepared, the crCCS reads in the *unknown parts* bin were the most informative for the first question (Figure 7.4 and Additional file 7A; Figure S3). The four crCCS reads in that bin showed three different element orders. One was *tNOS – nnn - NPT II – nnn - p35S*, present in two crCCS reads in both orientations, where *nnn* denotes an unknown sequence. This element order was consistent with that of the MON15985 parental line MON531. The second order was *p35S – nnn – tNOS – nnn – UidA*, consistent with the order in the retransformation construct of MON15985 (Figure 7.5 and Additional file 7A; Data S6). Sequences with these two element orders were previously not linked to any MON15985 designated sequence in a public database. The third element order was *p35S – nnn - hsp70 - cry1ab*, corresponding to MON810. The gap between the p35S and HSP70 was 16 nucleotides, and was the result of an incomplete alignment with the p35S element reference sequence.

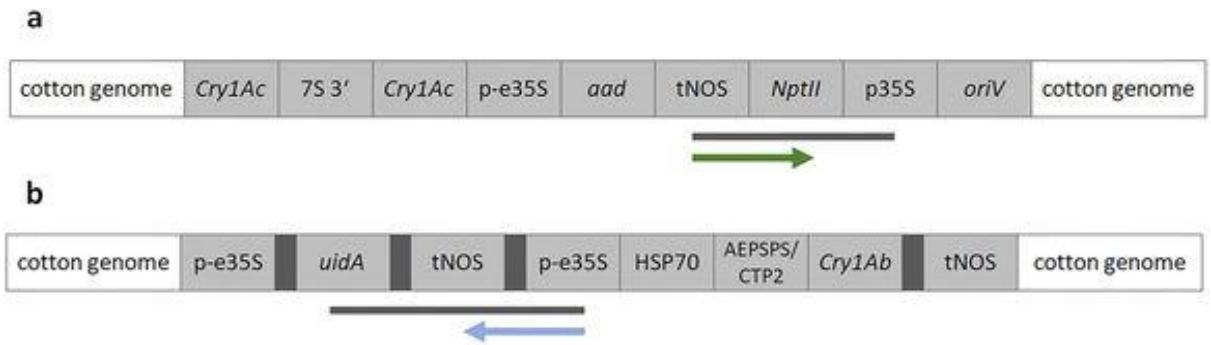


Figure 7.5 Two partial sequences of the MON531 and MON15985 inserts. Panel A shows the alignment of crCCS reads 133951 and 45207 to the insert and flanking region of MON531, with the position of enrichment primer (downstream NOS terminator). Panel B shows the alignment of and crCCS read 156962 to the insert and flanking region of MON15985, and the enrichment primer (upstream 35S promotor). Both sequences cover a previously unknown insert sequence.

In addition to the four crCCS reads in the *unknown parts* bin, three crCCS reads (*120091, 43434,* and *106520*) ended up in the *only unknown information* bin (Figure 7.4). To find the origin of these reads an NCBI Blastn, against nucleotide database, and a Megablast and a Blastn against NCBI patent database were run. For read *120091* a 99% identity and a 98% CCS read coverage against *maize genotype CMS-S mitochondrion* (DQ490951) was observed. For read *106520* top hits were *nucleotide sequences and polypeptides encoded thereby useful for modifying plant characteristics* (GP689587) and *long intergenic noncoding RNAs in maize* (JC761506). A total of 76% of the read was covered. For read *43434* the top hits were described as a *composition and method for therapy and diagnosis of ovarian cancer* (DL059073), with 90% identity and query coverage of 60% (Additional file 7A; Table S4).

The *confirmed events* bin, as the result of step 4, provided the answer to question (2) on identification of known GMOs. In order for a crCCS read to be acknowledged as containing a confirmed event, a 97% minimal coverage of the event sequence in the database was required, with at least 95% identity (Table 7.3). All seven reads that fulfilled the criteria of step 4 originated from MON810 (Figure 7.4). In one of these, the read aligned to the same database event sequence twice, both with a 100% identity match. After closer investigation, the CCS read turned out to contain a complete repetition of one sequence; an AAP primer with a poly G tail prior to a 5' MON810 flanking region followed by a p35S promoter sequence. The sequence was kept in the *identified GMOs*.

In step 5, a bin was made for the fully annotated CCS reads, containing *only known information*. In this step the answers to the question (3) on the list of potential GMOs were found. An indication of MON88017 and MON810 was observed (Figure 7.4). Two crCCS reads completely aligned with a database sequence, the first CCS read had the element order of *p35S - hsp70 - cry1ab*, present only in MON88017 and the second *cp4epsps – tNOS – 35S* present only in MON810, considering the GMOs used in this experiment.

Table 7.3 Results of NCBI BLAST+ against the event database.

crCCS read	Identified GMO	% of identical match	Alignment length	Length of db sequence	Alignment length in %
33879	MON810	100.00	92	92	100
40876	MON810	95.65	92	92	100
78816	MON810	98.92	93	92	101
119658	MON810	96.81	94	92	102
128024	MON810	100.00	92	92	100
132414	MON810	96.74	92	92	100
158734*	MON810	100.00	92	92	100
158734*	MON810	100.00	92	92	100

**CCS read 158734 showed two perfect alignments with the database sequence.*

7.2.6 The merged MON810 reference sequence

The merged MON810 read RIKILT20151130 (deposited at The European GMO reference database - Euginious http://www.euginius.eu/euginius/pages/sequence_string.jsf?sequence=4165286415439512999) that comprised of sequences JQ406879 and AY326434 was compared to the CCS reads. Bowtie2 [308] mapping with default settings of the CCS reads was performed using very sensitive local alignment (Additional file 7A; Figure S4). The CCS reads were first filtered by quality, only those that had a quality score of 33 or more over at least 90% of the sequence were accepted. Out of the initial 411 CCS reads,

165 filtered sequences aligned to this reference. A total of 133 CCS reads covered the overlap between sequences JQ406879 and AY326434, confirming the trueness of the new, merged MON810 reference sequence.

7.3 Discussion

The main goal of this study was to develop a protocol for identification of UGMOs in a complex mixture. The new approach was successful in its aim, and a proof of principle was found by identifying previously unknown sequences corresponding to the element order in a GMO.

The adapted GW approach, ALF, aimed to enrich the adjacent regions of the target prior to sequencing. Enrichment decreased the required sequencing depth and cost, as well as the data analysis time and cost. A new combination of GW steps was conceived and tested. Two common approaches were combined: gDNA reduction and the RACE/LT-RADE principle of tailing and semi-universal PCR. Biotinylation of the enrichment primer was used for background reduction. Since genomic DNA background was already reduced beyond detection, we decided against the use of a nested PCR approach. The semi-nested PCR in this paper is actually a first round of PCR. It is called semi-nested because one of the primers is nested, i.e. slightly downstream of the primer used in the LE step. PCR in general, and nested PCR in particular, is prone to contamination. For most diagnostic laboratories, the use of nested PCR is something to be avoided, especially in case of many repetitive tests. A contamination of a GMO testing laboratory with especially common GM elements would severely compromise an efficient GMO screening using the matrix approach. Target-specific LE primes as well as the nested primers of the snPCR were designed on the basis of the available qPCR primer-probe sequences for the targeted GMO elements, in our case p35S promoter and tNOS terminator. Primer and probe sequences of an unexplained GMO element test might be the only basis for finding related sequences, in case of looking for unknown GMOs as the result of a GMO matrix approach outcome. Therefore, the primer and probe sequences are a logical starting point for enrichment.

The high quality of the CCS reads warranted the omission of an extra quality check at the beginning of the analysis pipeline. This is important, as the library preparation requires a high amount of DNA, with low CCS read counts resulting from a lower DNA input. In this experiment, 411 CCS reads were generated despite a high amplification of relevant molecules in the ALF protocol. A clustering step was added to the workflow to further reduce the number of redundant CCS reads. To make sure no information is lost during clustering, a workflow omitting this step was tested, and it showed no extra information. The largest decrease in redundancy was observed in the *confirmed event* bin, where the decrease of sequences was more than 97% (295 *vs* 7 sequences). However, the main difference influencing the time of further analysis was observed in the *only known information* and *unknown parts* bins, where the number of sequences was decreased more than 92%; 28 *vs* 2 sequences after clustering in the *only known information* bin and 53 *vs* 3 sequences in the *unknown parts* bin. To reduce the time for clustering, a length filtering set at 50 nucleotides was applied in the very first step of the workflow. Reads of this length are not informative, as the shortest transgenic element in our element database is 60 base pairs.

The experimental sample was composed of known GMOs in different quantities: MON810 maize– 97%, MON89034 maize– 1%, MON88017 maize– 1%, and MON15985 cotton– 1%, or expressed in estimated copy numbers: the MON810: ~40.000 estimated copies, and the other three: ~400 copy numbers. This approach is somewhat different from the approach taken by Fraiture et al 2017 [21]. They used lower copy numbers (20 in their lowest mixtures), but in equal amounts, for three GMOs. Both situations may actually occur in real samples that may be contaminated by UGMOs. For three GMOs

the complete insertion sequence including the corresponding flanking genomic sequence was known. For one GMO, MON15985 cotton, only the element order and partial 5′ and 3′ insert sequences with corresponding flanking genomic sequence were available. Furthermore, MON15985 is a retransformation of another GMO cotton, MON531. The MON531 sequence was also unknown, with the exception of the element order. Therefore, MON15895 mimicked a UGMO in this experimental set-up.

The results can be looked at from two different perspectives. The first is whether or not the workflow enabled identification of UGMOs. The second is whether or not all the input sequences were detected. The transgenic cotton mimicked a UGMO, as explained above. Element orders corresponding to MON15985, to both constructs, were found in the *unknown parts* bin. By finding these two sequences, it was shown that a UGMO containing a known element can be identified. The sequences as such were not before specifically annotated to be part of the MON19585 and MON531 constructs. Event-specific sequences for either event were not found. The MON15985 event-specific test is at 3′ end of the insert. We therefore expected MON15985 in the *confirmed events* bin, through reads starting from tNOS in downstream direction. However, the tNOS enrichment was less successful than p35S enrichment. Out of 383 sequences, 368 were enriched starting from p35S and only 15 from tNOS, and upstream enrichment was in both cases seven times more successful than downstream. This is indeed a striking difference, and seemingly in conflict with the observation in the qPCR evaluation. Potential explanations would be variability in the procedure and or the samples, as the NGS sample (mixture of GMOs) was different from the qPCR evaluated sample (single GMO), or, some kind of bias in the library preparation/sequencing run. Noteworthy, Fraiture et al also found a large variation in numbers of reads related to the different starting points and directions they used for GW[21].Of the 16 CCS reads, two were tNOS enriched, one upstream and one downstream. The one downstream was indicative of the MON15985 GMO, via the tNOS-nnn-NTPII-p35S sequence present in the MON531 construct of MON15985 (Figure 7.5).

From the other perspective, finding back the input transgenes was partially successful. Out of the 16 crCCS reads, 13 could be unequivocally linked to one of the four input sequences, MON810, through reads starting from p35S in upstream direction (Figure 7.4). We could not find back event-specific sequences for MON89034 and MON19585 since the event-specific tests are at 3′ end of insert, and tNOS reads were much less abundant than p35S ones. Still, there was sequence evidence linked to the presence of MON88017 and MON15985, the latter one is discussed above. For MON88017, a partial insert sequence covering cp4epsps, the tNOS and the p35S was found. This clearly indicated the presence of MON88017, as no other transgenic line contained this sequence in this experiment. The MON88017 event-specific sequence was actually not expected to be found in the dataset, as both p35S and tNOS are present in the middle of the inserted sequence and in order to reach the point of insertion the upstream or downstream enrichments should cover ~3.5 kb. The longest CCS read was approximately 2.5 kb. Without detection of the point of insertion, the presence of MON88017 and MON15985 cannot be asserted with the same certainty as the presence of MON810. Some other transgenic lines containing the same element order could have been present in the reference material used for sample preparation. As a matter of fact, the *nptII* cassette with the CDS flanked by a p35S and tNOS is found in several GMOs, e.g. MON87460 and MON863 maize, and MON757 and MON1698 cotton.

Finding back specific sequences depends on the databases used. Elements are the building blocks of a transgenic sequences. Elements with the same function, the p35S for instance, do not always have the same sequence length in all transgenic lines. This makes building of an element database more

difficult and the database larger than desired. For this reason, a local database was built containing only the longest sequence for each element. The drawback of such a database is that although element sequences are very similar they are not identical. This needs to be considered when interpreting the results. A 75% coverage of a p35S does not necessarily mean that there is an incompletely explained read with a gap between elements: it could be that this certain promoter sequence is not completely identical to the one in the transgenic line. Such was the case for the crCCS read with the event sequence *p35S – nnn - hsp70 - cry1ab*, where a 16 nucleotide gap was found between 35S promoter and HSP70. This sequence was placed in the *unknown parts* bin, where potential UGMOs are expected, although further investigation of this read showed that it belonged to MON810. Such a bias might be resolved by addition of all versions of a particular sequence to the database and considering the bit score of the hits after BLAST analysis. Implementing the 'filter and sort' tool in Galaxy would allow this, and this addition is recommended for future use of this pipeline. Another database challenge lies in element sequences not always being annotated in the same manner in all database entries. For instance, there is a possibility that a part of a plasmid used in the process of transformation can be included into the element annotation in some database entries and not in others.

The identification of three crCCS reads without any known GMO homology was not expected. All three reads contained at least partial enrichment primer sequences at 5' or 3' end, and therefore passed through the enrichment primer filter, pointing to an unspecific binding of enrichment primers to a sequence other than p35S or tNOS. In case of crCCS reads 120091 and 43434 the sequences can be explained as maize mitochondrion and a non-plant sequence, respectively. For crCCS read 106520, however, NCBI patent database Blast identified this sequence as part of patents US756989 and WO2014036048A1, sequences accession number GP69587 and JC61506, respectively, with the field of invention for both patents described as *related to methods altering gene expression* (Additional file 7A; Table S4). While the crCCS read contained only the p35S enrichment primer and not the p35S promoter sequence, there is still a possibility that this sequences could be left over from the genomic transformation. In the current version of the element database we only included the known elements of the GMOs in the study, and actually left out the elements known to be present in the mimic-UGMO, such as nptII. Obviously, *nptII* should be added to the database in a real life situation. Likewise, any homology leading to identification of previously unknown elements should be added to this database, when verified. Within the current three hits, besides the titles, there were no real solid clues for a GMO related sequence. In Table S4, the query and subject starts and ends are given, plus the description of this region in the subject. None of the three reads showed any known GMO related sequence besides the presence of the primer. Therefore, the identification of a previously unknown GMO element, harbouring the 35S primer sequence, is a possible explanation. Another, perhaps more likely explanation would be off-target hybridisation of the 35S primer. In either case, if proven repeatable, this sequence might perhaps be added to one of the dedicated databases in the pipeline, probably the element database, with an annotation such as: previously found potential GMO-related sequence.

A new MON810 reference sequence was combined from previously known sequences JQ406879 and AY326434. There is overlap between sequences JQ406879 and AY326434 in the p35S promoter region. With 133 CCS reads mapping to this overlap, sufficient evidence was given for merging these two reference sequences in one reference sequence named RIKILT20151130. RIKILT20151130 replaced the JQ406879 and AY326434 sequence in the final workflow (Additional file 7A; Data S2).

The main objective of the analysis pipeline was to guide the end-user as quickly as possible to those sequences that require further investigation. For identification of UGMOs, known authorised GMO events are merely nice to know, and reads completely covered by a known construct sequence

are of even lower priority. Artefacts should be recognized and filtered, independent of where in the procedure they might occur. For this reason, the reads were set apart that contained only known primer sequences and not any further GM element homology. The placement of reads in separate bins instead of removing reads from the output, enables the end-user to further detail the analysis at any point in the procedure. The 18 CCS reads without a primer were not investigated further. The reasoning behind this was that all sequences should be related to the experiment that was performed. If none of the primers used in the experiment could be found, the link of such a sequence to the experiment itself is weaker. On top of that, if some of these did actually contain a primer, but with too many errors, the quality of the rest of the sequence would also be less reliable, increasing the risk of having to manually explore false positives. The contents of all other bins were analysed and a likely artefact was found in the *confirmed events* bin. One read showed homology to two event sequences instead of one. The raw CCS read, before trimming and primer filtering, turned out to consist of an AAP primer with a poly G tail, 5' MON810 flanking region followed by p35S sequence, with this motive being repeated. The repetition in itself is already rather strange, and less likely to be a true molecule because of the simple reason that if so, chances are that it was published already, given the well sequenced nature of the MON810 GMO. The presence of primers used in the experiment at the very borders of the repeated sequence makes it near impossible that this sequence is not an experimental artefact. A very likely explanation is a ligation artefact, i.e. the same molecule was ligated twice in a row between smartbell adapters. In the potential UGMO output, an element order consistent with MON810 was identified, besides the relevant sequences for the partially unknown GMO. This read should have been placed in the *only known information* bin, but was not, due to threshold settings. We did choose to keep the threshold as it was. A lowering might easily solve the problem in this dataset, yet it might cause false negative discoveries in others. In this setting, it serves to illustrate the point that UGMO discovery may never be free of false positives. Therefore, all these potential discoveries should be confirmed independently, preferably by design of a novel PCR, based on the newly found sequence. Within the current dataset 2372 polymerase reads containing 11.538 subreads were present. Of those, 411 polymerase reads contained more than four subreads of high enough quality to be merged into a CCS read. Those 411 could be further clustered, based on sequence homology, into 16 distinct sequences. Only four of those needed hands-on evaluation, of which three proved 'unknown'. All this was automated, meaning that out of a potential 11.538 reads, only 4 had to be manually checked, and three of those rightly so. In case this is a true new, potential UGMO sequence, this sequence should then be confirmed through Sanger sequencing of the PCR amplicon, from an independent DNA isolation of the suspect sample.

In summary, we conceived and tested an integral approach for a lab-based target enrichment based on a suspect GM element, followed by NGS and data analysis by design. We successfully showed the identification of partial sequences of a model UGMO. Future experiments will be aimed at further testing the approach in more complex mixtures, and the current settings of the pipeline design. This study provides a first outline of an automated, web-based analysis pipeline for identification of UGMOs containing known GM elements.

7.4 Methods

7.4.1 Description of certified reference materials

Certified reference materials (CRMs) of MON89034, MON88017, and MON15985 and the well-characterized reference material MON810 from a field trial [299] were used for the preparation of complex mixtures. For detailed information on the reference materials see Table 7.1.

7.4.2 DNA isolation and preparation of the mixture

Per CRM DNA was isolated using a CTAB extraction followed by the Qiagen DNeasy plant mini kit (Qiagen) according to Scholtens et al.[5]. 100 ± 10 mg of dry material was weighed and extraction was performed by adding 700 µl of CTAB buffer (20 g/L CTAB, 1.4 M NaCl, 0.1 M Tris, 20 mM Na2EDTA), 200 µl of nuclease-free water (Life Technologies) and 5 µl of Rnase A (Qiagen, 100 mg/µl) and incubated for 15 min at 65 ˚C in a thermo shaker at 250 rpm. Subsequently, 20 µl of proteinase K solution (Fermentas; 20 ng/µl) was added and the mixture was incubated in the thermo shaker for another 30 min at 65 ˚C, 250 rpm. To precipitate detergent, proteins, and polysaccharides 200 µl of Buffer P3 (Qiagen, DNeasy plant mini kit) was added to the lysate, the mixture was mixed and cooled on ice for 5 min. After cooling on ice, the manufacturer's protocol (Qiagen, DNeasy Plant Handbook 10/2012) was followed starting from step 10. The quantity and purity of the isolated DNA was assessed from Nanodrop absorbance measurements (Nanodrop 1000 instrument, Thermo Fisher Scientific). The mixture was prepared by combining 2.68 µl of MON15985 (18.68 ng/µl), 0.91 µl of MON88017 (54.92 ng/µl), 1.09 µl of MON89034 (45.79 ng/µl), 38.4 µl of MON810 (126.4 ng/µl) and 57 µl of nuclease-free water (Life Technologies).

7.4.3 Linear enrichment

Linear enrichment (LE) was performed in four separate reactions: p35S up, p35S down, tNOS up and tNOS down. Each 20 µl reaction contained: 1× Buffer 1 (17.5 mM MgCl2, Expand Long Template PCR System, Roche), 200 µM dNTPs (10 mM dNTP mix each, Invitrogen), 3.75 U polymerase blend Taq + Tgo (Expand Long Template PCR System, Roche), 125 nM biotinylated enrichment primer and 200 ng genomic DNA. Copy numbers were estimated to be the ~40.000 for MON810 and ~400 for the other three GMOs. A 1 C value was used of 2.725 for maize and 2.33 for cotton [309] and a convension factor for 50% for hetero/hemizygous GMOs as recommended by the JRC [310] to calculate in the 200 ng mixture: 35596 MON810 copies, 367 MON88017 and 367 MON89034 copies, and 429 MON15985 copies, which we rounded to 1 significant number. The following program was performed in a thermal cycler (iCycler, Bio-Rad): 2 min at 95 °C, 20 cycles of 1 min at 95 °C, 5 sec at 60 °C, ramp to 72 °C, over 1 min and 5 min at 72 °C.

7.4.3.1 Column & bead purification

Column purification was performed (Qiaquick PCR Purification Kit, Qiagen) according to the manufacturer's instructions, for removal of surplus primers and primer dimers. The linearly enriched fragments were eluted using 30 µl elution buffer provided with the kit. Streptavidin coated magnetic beads were used to select for the biotinylated enriched fragments from the genomic background. 15 µl Dynabeads MyOne Streptavidin C1 (Invitrogen) per sample were washed according to the manufacturer's instructions and resuspended in 30 µl 2× B&W buffer (10 mM Tris-HCl (pH 7.5), 1 mM

EDTA, 2 M NaCl). To immobilize the fragments, 30 µl column purified biotinylated enriched fragments were added to 30 µl washed beads and incubated for 30 min at 20 °C using a thermoshaker at 600 rpm. After immobilization, the DNA-bead complexes were washed for three times with 1× B&W buffer. Finally, samples were resuspended in 10 µl 10 mM Tris.

7.4.3.2 Tailing

The tailing reaction consisted of 5× tailing buffer (5' RACE System for Rapid Amplification of cDNA Ends, Invitrogen), 200 µM dCTPs and 10 µl purified sample. Water was added to a final volume of 24 µl. After mixing well, the mixture was incubated for 3 min at 94 °C and chilled for 1 min on ice. 1 µl of TdT (15 U/µl, Life Technologies) was added to each sample and incubated at 37 °C for 10 min. To heat inactivate the enzyme a last step of 10 min at 65 °C was performed then the samples were put on ice.

7.4.3.3 Semi nested PCR

Expand Long Template PCR Buffer 1 (1×), dNTPs (200 µM; each, Invitrogen), semi-nested primer (0.4 µM) and Abridged Anchor Primer (0.4 µM, 5' RACE System for Rapid Amplification of cDNA Ends, Invitrogen) (Additional file 7A; Table S1) and *Taq*+*Tgo* polymerase blend (2.5 U, Expand Long Template PCR System, Roche) were combined with 10 µl of tailed product in a total volume of 50 µl. Amplification was achieved using the following cycling program: 94 °C for 2 min, 45 cycles of 95 °C for 10 sec and 59 °C for 6 min and a final extension of 72 °C for 5 min.

7.4.4 Real-time polymerase chain reaction (qPCR)

Prior to q-PCR, reactions were diluted based on the amount of DNA put in the LE reaction. The following concentrations were used: 0.8 ng/µl for starting material (SM) and, 0.008 ng/µl for LE. qPCR reactions consisted of Diagenode (GMO-MM2X-A300), primers and probe (Additional file 7A; Table S2) and 5 µl template in a total volume of 50 µl. Amplification was performed as follows: 2 min at 50 °C for Uracil N-Glycosylase (UNG) decontamination, denaturation for 10 min at 95 °C, 45 cycles of 15 s at 95 °C and 1 min at 60 °C using a MyiQ or CFX real-time PCR machine (Bio-Rad). Data-analysis was performed using CFX Manager Software Version 3.1 (Bio-Rad).

7.4.5 PacBio analysis

A sequencing library for the PacBio platform was generated and sequenced at BaseClear BV (Leiden, The Netherlands). The ALF reactions were pooled, concentrated using 1.8 volume AMPureXP beads (Beckman Coulter Life Sciences) and subject to size-selection on agarose gel. The DNA fraction larger than 1 kb was isolated from gel, with a yield of 9.3 ng according to a measurement with the dsDNA high sensitivity Qubit assay (Life Technologies). The standard library preparation kits and protocols of the manufacturer (Pacific Biosciences) were used for library preparation, without DNA fragmentation, and despite the low input. The final library had a very low concentration of 0.1 ng/µl and was sequenced on one PacBio RSII SMRT cell. The run generated 2372 polymerase reads with an average subread length of 1031 bp. PacBio SMRT Portal's CCS workflow was used to generate 411 CCS reads (deposited at NCBI's Biosample database; accession number SAMN06461360).

7.4.6 NGS data processing

CCS reads gained by PacBio RS II NGS sequencing were processed with Galaxy platform [311-313]. A pipeline, i.e. workflow (Additional file 7C; pipeline, https://github.com/RIKILT/ALF), was built using tools available in Galaxy (Additional file 7A; Table S3). Three databases were built: event, CAF (Construct and Flanks) and element database (Additional file 7A; Data S1, S2, S4 and S5). CAF database comprises of annotated reference sequences of MON810 (RIKILT20151130 joined JQ406879 and AY326434), MON89034 (DI362404), MON88017 (HV702026) and two partial MON15985 (EA135634 3' flank, EA135633 5'flank). Annotations were written in a General Feature Format (GFF) file (Additional file 7A; Data S3).

7.5 Additional files (available with the publication)

Additional file 7A: Table S1 Function and sequence of the primers used. **Table S2** Primer sequences and final concentrations used in the qPCR reaction used to determine size-dependent increase of specific target and loss of genomic background. **Table S3** Tools used in building of the Galaxy workflow for UGMO detection. **Table S4** BLAST output (first 5 hits) of the crCCS reads in the 'only unknown information bin'. **Data S1** Event database sequences. **Data S2** Reference sequences, CAF database. **Data S3** Annotations for reference sequences in CAF database. **Data S4** Element database - redundant, containing element sequences from all input GMOs. **Data S5** Element database , an un-redundant element database, containing only the longest of the element sequences. **Data S6** CCS reads with element orders tNOS – gap- *Npt II* – gap- p35S (133951, 45207), and p35S – gap– tNOS – gap– *uidA* (156962). p35S promoter is coloured yellow, tNOS terminator green *Npt II* gene red and *uidA* gene purple. **Figure S1** Formation of the circular consensus sequence (CCS) read. **Figure S2** Length distribution of the 411 CCS reads as generated by FastQC. **Figure S3** Building the workflow: in depth description. Figure S4. Bowtie2 mapping of CCS reads to RIKILT20151130 sequence.

Additional file 7B: Table S1 Element order list 1_raw. **Table S2** Element order list 2_raw. **Table S3** Identified GMO. **Table S4** Element order list 1. **Table S5** Element order list 2. **Table S6** Artefacts sequences.

Chapter 8

General discussion

8. 1 Introduction

The genetic code of every living organism on this planet is written in deoxyribonucleic acids (DNA). Every species, as well as varieties within a species, has a unique combination of conserved and variable, unique DNA regions. These species/varieties -specific variable regions can form the basis for DNA-based identification methods. For species identification DNA barcoding is an often applied approach to identify species based on variable regions in the mitochondrial DNA. For GMO varieties, identification is based on specific junctions between the GM construct and plant genome. Traditionally, a specific PCR followed by Sanger sequencing or a specific TaqMan PCR are the frequently applied molecular biological techniques to identify species or GMOs, respectively. However, in the last two decades, sequencing technologies aiming at determining multiple DNA sequences have evolved to achieve simultaneous sequencing of different DNA strands from one to many species in a single analysis. These new sequencing techniques are often called next generation sequencing (NGS). Examples of such NGS technologies and platforms are HiSeq and MiSeq (Illumina), PacBio RSII (Pacific Bioscience), Ion Torrent (ThermoFisher) and MinION (Oxford Nanopore). Each of these technologies has its own advantages and disadvantages. Unravelling the genetic composition of a complex product may identify the presence of specific ingredients, such as endangered species or authorised or unauthorised GMOs, in this way evaluating the authenticity of food/feed products of interest.

The objective of this thesis was to use detailed genetic differences to identify species/varieties in complex products based on the application of advanced analytical NGS based strategies, with a specific focus on the identification of two target groups: endangered species and GMOs. An initial enrichment step, coupled to an apt NGS-based strategy, were applied to detect and identify endangered species or GMOs, including UGMOs, in a product of interest. Two identification strategies were considered that were either based on known sequences (endangered species and GMO), as available in dedicated databases, or on the elucidation of unknown sequences, adjacent to identified known GMO-related sequences (in the case of UGMOs).

The current chapter discusses the research findings of this thesis and the interconnection between the research objectives as proposed in *chapter 1* for endangered species and the broader GMO detection. With respect to endangered species, the aim was to explore whether NGS- based strategies allow the simultaneous identification of all species, including endangered species, present in a sample, even in complex samples that may be heavily processed (*chapter 2, 3 and 4*). With respect to GMOs, the aim was to reliable identify all GMOs and UGMOs present in a given sample, regardless of their relative abundance, based on enrichment of known or, in the case of UGMOs, additional adjacent unknown sequences (*chapter 5, 6 and 7*). Additionally, the implications of applying the newly developed NGS based methods for food authenticity are discussed and the main conclusions are presented.

8. 2 Overall discussion of the findings

8.2.1. Endangered species

Around 35,000 species worldwide, belonging to a wide range of plant and animal taxa, are classified as endangered by CITES [53]. These endangered species are sometimes used as illegal ingredients in food samples. Customs authorities enforcing the CITES convention screen food products for the illegal presence of endangered species. However, identifying the genetic composition of food samples, in terms of species present in the sample, can be a challenging task due to the complexity of some of the samples of interest. Seized food samples might consist of difficult matrices, such as, powders, tablets, wet-balls etc [8,9,27]. Additionally, these food samples often contain many different species, belonging to both the plant and animal kingdom [8]. Therefore, the identification of all species present in a product using a single analysis is a challenge [9], due to the lack of adequate DNA extraction protocols, the fragmentation of the DNA, and the complexity of the species composition. In *chapter 2, 3* and *4* these

issues have been addressed, with the aim to develop improved strategies to identify (endangered) species in a multiplex setting on the basis of NGS-based strategies, that make use of enrichment of well-established species-specific sequences.

8.2.1.1 Enrichment strategy

For successful enrichment of species-specific sequences in complex samples, several preconditions need to be met. First of all, the DNA extraction needs to be adequate to obtain sufficient quality DNA to allow efficient amplification. Secondly, the complexity of the sample composition, in terms of numbers and variety of species, requires a broad range of markers yet with enough resolution to enable species level identification. Additionally, high resolution for identification should be obtained both in cases were non-degraded, high quality DNA is available as well as in cases were the DNA is degraded and only fragments are available for enrichment.

With respect to the DNA extraction, DNA isolation methods for the different types of products of interest can be very difficult to standardise and optimise because of the complexity and diversity of samples that may contain wild life forensic materials [8]. Traditional medicines (TMs), one group of products that is of interest with relation to the potential presence of endangered species ingredients, often consist of difficult matrices making it challenging to extract good quality DNA. These difficulties in DNA extraction arise from the significant processing of the raw materials present in the products [8,9,27] As a consequences the DNA of the ingredients may be heavily fragmented, or there may be multiple (polymerase) inhibiting factors present that will affect the quality of the resulting DNA. Currently, no universal DNA isolation method is available for the various TMs matrices. Studies have shown that CTAB extraction buffer combined with an additional silica or resin-based DNA purification step is most efficient for a wide range of plants and plant-derived products, in particular for the separation of polysaccharides from the DNA [66,89,90]. In line with these results, in *chapter 3* it was shown that the CTAB isolation method was the optimal method to extract good quality, amplifiable DNA from both plant and animal reference materials, multiple non-processed complex experimental mixtures and from a limited number of TMs. Nonetheless, in *chapter 4* when multiple TMs from various matrices were analysed, CTAB failed to obtain amplifiable DNA from all TMs included in the assessment. This was likely due to the presence of potential inhibitory components and interfering substances derived from the samples (e.g. protein, lipids, polyphenols, polysaccharides). To address this issue, eight commercially available and commonly-used DNA extraction methods (organic extraction, silica-based, and magnetic beads based) were compared for their ability to obtain sufficient DNA of good quality from TMs to allow adequate PCR amplification. This comparison showed that CTAB in combination with the Wizard DNA clean-up system was the most effective strategy, since it gave a positive amplification for all the analysed TM samples. DNA isolation is a very crucial step in a metabarcoding analysis, the combination of CTAB and clean-up system efficiently removed potential inhibitory components and interfering substances. These results are in accordance with another study, that showed that the combination of CTAB and clean-up system aid to achieve good quality DNA from herbarium specimens [66].

With respect to the choice of markers, in order to be able to identify a wide range of both plant and animal species in a single analysis, universal markers are needed, yet these markers should provide enough resolution to distinguish between species. In *chapter 2*, we explored the scientific literature with the aim to select the most informative universal DNA barcode and mini-barcode markers and their related primer sets. In general, the conclusion from the literature review was that all universal barcode and mini-barcode markers and related primer sets had been shown to amplify the target region efficiently in single species analyses. However, their performance in a meta-barcoding study had yet to be evaluated. In *chapter 3*, the selected universal DNA barcode and mini-barcode markers and their related primer sets from *chapter 2* were used to develop a metabarcoding panel that potentially would

be able to identify both plant and animal species in complex matrices. However, all selected universal barcode and mini-barcode primer sets had been optimised to amplify targets under different PCR conditions. Therefore, as a first step a single optimal PCR condition was determined for twelve plant and animal barcode and mini-barcodes, based on the amplification of the barcode region in 19 plant and 29 animal species, belonging to different families. The results of *chapter 3* and *4* showed that no single barcode marker (plant or animal) used in the developed barcode panel could amplify all the species-specific sequences present in the experimental mixtures. This was especially the case for species present in low percentages, where multiple markers were necessary to identify the presence of the species. With a combination of multiple barcode markers nearly all of the species in the experimental mixtures could be identified, including the endangered species.

8.2.1.2 Next generation sequencing

Current approaches for DNA-based species identification include PCR amplification of the specific target, followed by Sanger sequencing [18]. A main problem in Sanger sequencing is that it is not applicable to products containing more than one species, unless additional time-consuming work (cloning of PCR products) is performed [198]. Although barcoding and mini-barcoding have been used to identify animal and plant species separately in a simple to complex mixture [198], so far no methodology had been described combining informative plant and animal barcodes and mini-barcodes in a single analytical strategy. In this thesis the use of metabarcoding, a rapid method that combines two technologies: DNA based identification and high-throughput DNA sequencing, was studied to identify a mixture of species in one analysis. The obtained DNA barcoding sequences can be compared with the database to pinpoint the species based on the nucleotides' variations observed between the DNA barcode sequences.

To allow high confidence in identification, here Illumina sequencing was selected over 454 and Ion torrent sequencing because of its low error rates (< 0.1%) in the generated data [18]. An high error rate in sequences could lead to misidentification of a species, as a comparison showed in the *chapter 2*. Additionally, for species identification in TMs, DNA might be severely damaged and a major advantage of using Illumina Miseq PE300 is that the generated paired-end reads can be merged to obtain pseudo reads with lengths of ~550 bp, which could provide a relatively high discrimination power even from fragmented DNA.

The amount of data generated by Illumina Miseq PE300 (2-5 Gb per sample) requires an adequate bioinformatics workflow to process these DNA metabarcoding data. To facilitate this requirement, CITESdetectpipeline was developed in *chapter 3*. The pipeline processes the NGS data by trimming the Illumina adapters from the reads, and merging the reads, subsequently performing a quality trimming, primer selection and trimming. In the next step the resulting reads are clustered and chimeras are filter out. Finally, a BLAST analysis is performed on the consensus sequences from the clusters to identify the species, including potential endangered species [198]. The robustness of the pipeline was assessed by varying multiple parameters, such as base quality, error tolerance for primer selection, Operational Taxonomic Units (OTU) radius, query coverage, E-value, OTU abundance and identity threshold. From this parameter evaluation optimal settings for the samples were determined, and a 98% identity threshold was found to give the best cluster separation between closely related species and also result in low false-positive identification. Subsequently this 98% identity threshold was used in the analysis of the experimental mixtures and the TMs (*chapter 3 and 4*). An OTU abundance threshold of 0.2% of the mapped reads was found to be most optimal in reducing background noise and potential false-positive. However, in *chapter 4* it was determined that the applied OTU abundance of 0.2% was not stringent enough to separate the low abundant species from background noise and that

this applied threshold may actually lead to false-positive identification. Therefore, to reduce the false positive identification, in *chapter 4* a static OTU threshold of 100 on top of the threshold of 0.2% of the mapped reads was implemented, similar to the threshold settings applied in another recent metabarcoding study with TMs[181].

A metabarcoding approach results in large NGS datasets, the analysis of these data in practice requires bioinformatics knowledge. However, end users, in the case of endangered species these are often customs authorities enforcing the law at the borders, might not be skilled in bioinformatics. Therefore, the web-based CITESdetectpipeline was developed and made freely available. The interpretation of the results was done in a conservative approach to avoid any false identification of the species. Generally, the top hit in the BLAST analysis is considered to be the identified species [198] However, to avoid any false-positive identification resulting from low quality sequence data in the database, a conservative approach was used requiring at least three similar top hits in the BLAST output for a positive identification. The down side of such an approach is that species with relatively few entries in databases are less likely to be positively identified at species level, especially considering the fact that endangered species are still underrepresented in public databases [18,198].

8.2.1.3 Identification of endangered species

In *chapter 3* and *4* the objective was to study whether the use of an informative set of barcoding and mini-barcoding markers combined with an NGS approach will allow the identification of multiple species in individual complex samples. With the final aim to setup a method that will allow the identification of multiple CITES species in a single analysis of real-life samples as provided by European customs laboratories. An important aspect of the study was the resolution at which the species could be identified. The results of this thesis show that the developed multi-locus barcode approach could achieve a high resolution of identification compared to other studies [35,143]. In general, we observed that with respect to animal identification, species level resolution could generally be achieved with a combination of barcodes. For plants, however, this was not always possible. In 50% of the cases a maximum resolution at family level could be achieved. The ITS2 barcode accounted for most species level identifications in plants, followed by mini-*rbcL* and *trn*L (P6-loop). Achieving species level resolution to identify plants with a single barcode marker is generally not feasible, especially in case of the mini-barcodes that have an inherent lower resolution compared to the full-length barcodes. Here and in other studies, ITS2 barcode showed to be a reliable barcode for the identification of taxa at species level in complex samples (herbal products) [199,200]. This barcode has been proposed as a universal barcode for plant species identification, especially for the medicinal plants [123] and endangered plant species [197]. In animal identification, mini-16S mainly accounted for the cases of species level identification, followed by 16S and cyt *b* barcode markers.

The emphasis in the first part of the thesis was on the identification of multiple species, including endangered species, in a single analysis. Therefore, in *chapter 3* also endangered species were used in composing the experimental mixtures. The results showed that all endangered species in the experimental mixtures could be successfully identified with the newly developed multi-locus metabarcoding method. Nonetheless, when analysing real-life TMs, only in two cases the presence of endangered species (*Ursus arctos* and *Cibotium barometz*), declared on the label, could be confirmed with the barcode panel (*chapter 3* and *4*). It could be that the barcode panel failed to identify all the species. This could be due to multiple reasons, such as, processing of the ingredients in such a way that the DNA was either degraded or effectively removed [183,198], or primer-template mismatches. Alternatively, it is possible that the claimed endangered species were not identified because the label information was

not correct, for example due to a lack of proper taxonomical knowledge with the producers, or because of deliberately but unjustly specifying a rare species on the label to increase the value [8]. Incorrect label information was commonly observed when the positively identified species in a TM sample were compared to the respective label information. It was found that only in two TMs all ingredients could be confirmed. Many undeclared taxa, however, were identified across the TMs, which were even higher in number compared to the confirmed taxa. One undeclared species *Saccharum hybrid cultivar* (sugar cane) was found in one-third of the analysed TMs, and is a commonly used species to achieve a sweet taste in food products or supplements. The presence of undeclared species is a common result in studies analysing TMs. In line with the results of chapter 7, recent studies found in the analysis of TMs that up to 98% of the species identified in TMs were not declared on the label [200,314-316].

8.2.2 GMOs

The second part of the thesis addresses the general aim to determine the presence of GMOs using advanced analytical NGS based strategies. In the EU, GMOs that have received market approval, are allowed to be present in food and feed products. It is furthermore stipulated that labelling of GMOs is mandatory in the EU for products that contain more than 0.9% of authorised GMO per ingredient. Initial detection of GMO varieties is usually based on the finding of GMO-specific elements and constructs in the DNA. In case of UGMOs, that are not allowed on the European market, generally limited sequence information will be available, and genome walking strategies are used to identify the unknown adjacent sequences of the known GMO element.

Currently, GMO-specific (elements, constructs and events) TaqMan PCRs have been developed to screen for the presence of GMOs in a sample) [19,22]. In a TaqMan PCR, target-specific primer sequences are used to amplify the target DNA of interest, in the presence of a probe sequence that will allow relative quantification of the generated amplicons. In recent years, an increasing number of GMOs with new GMO elements have been approved for use in food/feed products [23]. In order to have an informative GMO screening, an increasing number of new GMOs and related targets should be part of the analysis, requiring the development and validation of the related single methods [317]. As a consequence, the current GMO screening approach becomes increasingly more time-consuming and costly. To overcome these issues, in the second part of this thesis the aim was develop an efficient NGS based GMO screening approach, which can circumvent the elaborate validation of each new GMO specific assay.

Additional to the increase in authorised GMOs, an increasing number of UGMOs have been entering the world market. Since generally limited sequence information is available for UGMOs, GMO screening approaches can usually only lead to indications for , but not confirmation of, the presence of a UGMO [23]. To confirm the presence of a UGMO, generally a genome walking approach combined with Sanger sequencing or NGS is applied. To this end, a number of genome walking strategies were reported to identify the adjacent sequence of an identified GMO element in reference materials (LAM PCR, LT-RADE, SiteFinding-PCR, A-T linker and LF PCR) [227-230,232]. However, those enrichment approaches have their setbacks. Bottlenecks are, for instance, the sensitivity, or the limited level of multiplexing that is allowed, or the flexibility of the system regarding the targets to be detected. Currently there is no single system that can meet the criteria for detection of low abundant UGMOs in complex samples consisting of a mixture of GMOs. Therefore, for UGMO identification the aim was to develop an adequate genome walking approach combined with an NGS strategy that will not be influenced by the relative abundance of the UGMO.

8.2.2.1 Enrichment strategies

With respect to enrichment of GMO related targets, the aim was to set up a methodology on the basis of informative combinations of well-characterised GMOs reference materials and investigate the applicability of such methodology in real-life samples. In *chapter 6*, such a GMO screening method coupled to NGS was presented. For this method it was shown that the Qiagen HotStar *Taq* (PCR) master mix was more efficient in amplification of the DNA targets compared to the routinely applied Diagnode TaqMan (TaqMan PCR) master mix. Furthermore, it was shown that the developed broad NGS-based screening (96 GMO-related targets) could identify the authorised GMOs present in the sample. This identification was achieved by screening for the presence or absence of candidate GMO elements common to multiple GMOs, as well as for the presence of GMO event-specific sequences. Some of the GMO elements present in the sample could not be explained by authorised GMOs and may be indicative for the presence of UGMOs. However, to confirm the presence of a UGMO, an additional genome walking strategy will be required, provided no known UGMO event-specific sequences match, to identify the unknown adjacent sequences of the detected GMO element. For this, the identified unexplained targets in the broader screening can serve as a starting point to 'read' into unknown regions using a genome walking approach.

In *chapter 5* a literature review on available genome walking approaches was presented in relation to possible UGMO detection. The available GW approaches were reviewed and evaluated with respect to their advantages and disadvantages for UGMO identification. The conclusion of the review was that in order to meet the specific demands of UGMO detection, it is necessary to develop a new gene/genome walking approach by combining the advantages of available approaches to have an optimised enrichment strategy. Subsequently, the approach can be coupled with NGS for sensitive UGMO detection in complex samples where the UGMOs may be present at low percentages. In *chapter 7* the development of such gene/genome walking approach, the Amplification of Linearly-enriched Fragments (ALF) approach was presented. Because the ALF approach aimed to elucidate unknown regions in the DNA, the length of the enriched product was of crucial importance, since enriching longer fragments will make the procedure less dependent on correct assembly of shorter sequences, will therefore provide more direct information about the adjacent elements and may include the event-specific region bridging the inserted construct and the endogenous plant genome. Previous approaches were found to enrich long DNA fragments, for example, LT-RADE [228] is a non-restrictive enrichment method, starting with gene specific primers for initial single strand enrichment, followed by tailing and nested PCR to synthesise double stranded DNA (dsDNA) of the desired amplicons [228,318]. Additionally, Locus-finding PCR (LF PCR) is a combination of random-priming PCR and nested PCR along with affinity-based purification to obtain the unknown sequence [230]. The ALF approach combines the advantages of the LT-RADE and LF PCR by using blend polymers to enrich longer DNA fragments, starting from the identified element in the GMO screening, using a linear enrichment step followed by a magnetic bead clean-up system to select only the desired linear fragments for further enrichment in semi nested PCRs [230].

Initially the efficiency of the ALF approach was evaluated on the basis of >99.05% MON88107 maize reference material. The efficiency was evaluated based on qPCR results performed on the starting material (SM), after the linear enrichment, and finally after the semi-nested PCR. The qPCR results after each step in the procedure were compared to determine the size of the obtained (amplified) fragments. The results of this analysis showed that the longest fragment in the ALF approach was at least ~2 kb, and a 3.5 thousand fold enrichment compared to the initial quantification of the basic template DNA was observed. Additionally, the genomic DNA quantity was compared before the enrichment and after the column and magnetic beads clean-up steps, and results show that the genomic DNA (gDNA) was reduced to beyond the detection level. The reduction of gDNA offers the opportunity to circumvent the use of a nested PCR to amplify the selected sequences, since nested PCRs require the re-opening of the

initially enriched DNA fractions, thus considerably increasing the chances of contaminations in the laboratory [272].

In general, the findings showed that the ALF approach is a more efficient method for enriching long, informative fragments compared to LT-RADE and LF-PCR GW approaches [228,230], and will allow easier identification of potential present UGMOs in samples of interest.

8.2.2.2 Next generation sequencing

As mentioned, one of the bottlenecks with relation to the identification of UGMOs is that, contrary to approved GMOs, for UGMOs there will usually be only limited sequence information available as a basis for the development of methods for detection and identification [15]. A matrix approach, such as presented in *chapter 6*, can be used to detect UGMOs that contain identical or allelic variants of elements present in authorised GMOs. Subsequent further elucidation of those unexplained GMO targets and their flanking regions by sequencing can be used to experimentally identify UGMOs in the sample as was shown in *chapter 7*. Both issues require different sequencing strategies. For sequencing of known, short length sequences from the GMOs (elements, construct and events), Illumina Miseq PE150 was found to be an apt technology due to the large amount of output reads, read quality, short runtime, paired-end reads and lowest cost per sequenced base pair (*chapter 6*). For the sequencing of long enriched DNA fragments, as generated in the ALF approach, the PacBio technology was chosen as the preferred NGS strategy, since this sequencing technology has the ability to sequence longer fragments.

Additional to an apt NGS technology, reliable identification of GMOs and UGMOs present in a sample can only be achieved when the obtained data can be efficiently analysed. In *chapter 6*, to analysis the generated NGS data, a bioinformatics pipeline was developed. This pipeline aimed to analyse the generated GMO related sequences, or amplicons, and was named amplicon sequencing pipeline (AM-SEQ). The developed pipeline consists of eight steps: removal of Illumina adapter, merging reads, quality selection, selection of primer containing reads, aligning against a local database, verification of new refence sequence and visualise of the output to efficiently process NGS data. Since the developed GMO screening method aimed to identify both GMOs and UGMOs even at low abundant levels, it was of crucial importance that background noise or cross contamination (targets with low reads counts) could be distinguished from true amplification (targets with high reads counts). Therefore, a threshold for detection (0.01% of mapped reads) was established for the NGS datasets of the real-life feed samples, this was based on a comparison of TaqMan PCR results to the analysed NGS data of a sample. Application of a threshold could potentially lead to false negative identification of GMO, nonetheless, it seems likely that targets that will not pass the threshold are present at such low abundance that they would not have been identified by the current TaqMan PCR strategies, and therefore the application of a threshold will not lead to decreased sensitivity of the screening.

Data analysis of UGMO-related sequences is rather difficult due to the same fact that in many cases only partial sequence information is available. To overcome this problem a step-by-step data process approach was followed in *chapter 7*, were only good quality reads were selected and the obtained reads were mapped to different databases with known GMO sequences (event- , construct- and element-specific databases) in a sequential order, in order to segregate the known sequences from the (partially) unknown sequences and to determine the GMO element order in unknown sequences of interest. The elements of the unknown sequence can be mapped against a public GMO database to retrieve the order of the elements and subsequently to determine whether it belongs to a known GMO. In this way, by using the element order of the unknown sequence of interest, in *chapter 7* the GMO MON15985 could be identified. Since in *chapter 7* a GMO served as a model for a UGMO, as a proof of principle of the ALF-approach, the analysis ended with the successful identification of an unknown sequence of an authorised GMO (MON15985). Nonetheless, in cases were the element order of the identified unknown sequence does not belong to a known GMO the event sequence (junction between

the GMO construct and plant) need to be identified to show the presence of a UGMO. In all cases where a putative UGMO has been identified, this will require confirmation by a dedicated PCR assay.

8.2.2.3 Identification of GMO and UGMO

For correct interpretation of sequence information it is important to have dedicated databases with high quality data at hand. In the field of GMOs, the EUginius database is a database that combines information on both authorised GMOs as well as on UGMOs (www.euginius.eu). It was shown in *chapter 6*, in preparation of the AM-SEQ approach, 79 new reference sequences for GMO-related element targets were obtained from analysing the positive control sample. Along with the positive control, 5 feed samples knowing to contain GMOs were analysed. The identified GMO-related targets from the feed samples were associated to the GMOs as identified in the same sample, and the results were compared with the two-step TaqMan GMO screening approach to understand the strengths and limitations of an NGS based approach. In this comparison, it was found that with the use of a wider NGS-based screening strategy 10 low abundant GMOs could be additionally identified in the same samples, which were not tested in the initial, routine GMO screening. It is clear that the presence of some GMOs in a sample may mask the presence of other, perhaps low abundant GMOs, and this may include also UGMOs. The results from the data analysis of the five feed samples showed that 9 of the identified targets were unexplained, these were all observed with a late Cq value of ≥ 37 and were not detected in all 3 replicates of the GMO screening. This confirms that identification of targets present at a low level is subject to stochasticity and that an NGS-based approach should preferably be performed in replicates. The stochasticity is also a reflection of the non-quantitative nature of NGS and the read counts are probably not a reliable reflection of the abundance level of the target due to exponential processes. The samples that contained the 9 unexplained targets were additionally tested in a TaqMan PCR analysis for known and in-house available GMOs containing these targets and were found negative, possibly indicating the presence of UGMOs at low levels. Alternatively, the presence of these unexplained targets could be traces of authorised GMOs, or in specific cases, donor organisms below the limit of detection by PCR [319]. We suggested that any detected unexplained GMO-related target with a Cq value ≤35 may indicate the presence of an unknown or unauthorised GMO that was not included in the analysis, or for which no method is available yet. In these cases, the NGS-based screening method may be followed by a genome walking strategy to identify the flanking region of the identified unexplained targets. Eventually this may lead to the identification of the UGMOs and thus provide sequence information for new methods for detection and identification [271,272].

In *chapter 7* such a genome walking approach, the ALF approach, was successfully developed and evaluated on the in-house made complex mixtures containing four GMO crops in which MON810 maize was present at 97% and three other GMO crops (MON88017 maize , MON15985 cotton and MON89034 maize) were present at 1% each. For the GM cotton (MON15985) only partial sequence information was available, essentially mimicking a UGMO from which limited sequence information is available. It was found that only the high abundant GMO (MON810) could be confirmed with event-specific sequence information. The lower abundant (1%) GMOs (MON88017 and MON15985) could only be partially confirmed using the construct and flanks and element databases, while MON89034 sequences could not be retrieved. After mapping to the databases some unmapped sequences having the partial element order of MON15985 were found. To find the origin of these unmapped reads, a manual BLAST analysis was performed, identifying the nature of the unknown sequence. This illustrates that the ALF approach may be used to identify UGMOs in a similar process.

8.3 Implications, limitations of the findings and recommendations for future research.

8.3.1 Endangered species

8.3.1.1 Enrichment strategies

Identification of a species in a product of interest is initially dependent on the availability of amplifiable DNA, and therefore the selection of an adequate DNA isolation methods is a vital step. In *chapter 3* and *4*, it was determined that standard CTAB isolation is sufficient to obtained amplifiable DNA from non-processed samples containing both plant and animal materials. However, the samples seized by custom laboratories are most often highly processed samples (TMs). It was shown that for such products a combination of CTAB with a clean-up system was necessary to extract amplifiable DNA for the DNA barcode panel analysis. Although amplifiable DNA of sufficient quality was obtained from these difficult samples, many species listed on the ingredient list could not be identified. It might be the case that the non-identified species were not present in the sample, alternatively, the clean-up process might have caused loss of DNA of low abundant species, as was shown by another study [66] leading to the false-negative identification of low abundant species. In future studies, it should be determined what the effect of the clean-up methods is in practice for the identification of low abundant species in metabarcoding approaches. Alternatively, matrix-dependent solutions may be considered, similar to what is currently done for complex GMO samples [320,321]. Ideally, the quality of the DNA should be adequately determined after the DNA extraction, to estimate chances for efficient subsequent amplification and related species identification. In *chapter 3* and *4* it was, however, found that it was not possible to determine the quality of the obtained DNA from the real-life samples in an adequate way using standard DNA assessment and quantification methods (e.g. NanoDrop or Fluorometer). Alternatively, a DNA fragmentation analyser could be used, but this is relatively expensive and customs laboratories are generally not equipped with these more advanced methods for DNA quality assessment.

Secondly, the identification of a species is based on the enrichment of the DNA sample for the target region. Amplification can be hampered by the complexity of a sample, for example, when DNA is heavily degraded and only short regions are available for enrichment or when multiple species are present in one sample and high abundant species may mask the presence of low abundant species. In *chapter 3* and *4* it was shown that with a combination of barcode and mini-barcode markers species identification can be achieved even when DNA integrity is low and species composition is complex. Based on the sample type, or with advance knowledge of the DNA integrity, the current barcode panel could be used in a more flexible manner, by using only full-length or mini barcode markers depending on the fragmentation of the DNA and only plant or animal markers depending on the sample type. In this thesis the aim was to cover a wide range of species and types of samples with the 12 selected barcode markers. It is, however, not certain that all the endangered species can be covered with these primer sets, and in the future newly developed universal primer sets could be added to the panel .

8.3.1.2 Next generation sequencing

Regarding the choice of the NGS technology to be used for metabarcoding, for species identification the Illumina technology was used to allow sequencing of full-length barcodes. However, the recently developed MinION system from Oxford Nanopore Technologies is able to sequence a range of barcode lengths with a good resolution. Also, the technology is portable and can theoretically be used on-site to determine the composition of samples. Before applying this new technology in a routine setup, however, a full validation study will need be performed.

The use of an alternative sequencing technology, such as PacBio or MinION, that is able to sequence longer DNA fragments may provide a higher resolution, especially for plant species identification. For example , it was found in *chapter 5* that the full-length barcode markers *matk* and *rbcL* were amplified in the case of highly processed TM samples, probably primarily containing fragmented DNA. Plant full-length barcode markers can provide species or genus resolution [122,188]. However, in the study we used paired-end (PE) 300 Illumina technology where the maximum read length is limited to ~300 bp, which is often shorter than a full-length barcode. Therefore, the resolution to identify species in practice depended on the forward or reverse reads of the full-length barcode (i.e. *matK, rbcL*).

8.3.1.3 Identification of endangered species

In *chapter 3* and *4* is was underlined that the identification of endangered species is limited by sequence information available in databases. Currently, CITES-listed species are underrepresented in public databases [183] and future studies should address the availability of these sequences in order to make identification possible of all endangered species. Another disadvantage of a species identification method based on sequence information is the requirement that samples actually contain DNA, it will not work in samples where no DNA is present, or if it is too heavily degraded. In such samples, identification of species might require additional steps in the analysis and studies have shown that a more holistic approach, including proteomics and metabolomics, in specific cases can be of help in efficient species identification [322,323]. Nonetheless, adding additional steps to the analysis will considerably increase the cost and time of the analysis.

Nonetheless, a major advantage of an NGS based detection method is the possibility for the simultaneous, broad screening for many targets. Raclariu et al, (2017) showed that a metabarcoding method was more suitable to determine the composition of complex products and yielded better identification compared to TLC and HPCL-MS [324,323]. The metabarcoding panel may aid the customs authorities to identify the illegal use of endangered species in products such as TMs. The high resolution achieved with the metabarcoding panel may be applied in future studies in other, related fields, such as food fraud issues [48] or the environmental monitoring of species [25].

8.3.2 GMOs and UGMOs

8.3.2.1 Enrichment strategies

In *chapter 6*, broad enrichment using the master mix HotStar *Taq* was more efficient in amplifying all targets included in the screening compared to the routinely used TaqMan Diagenode master mix. Although HotStar *Taq* outperformed TaqMan Diagenode in the positive control, still three out of 96 could not be amplified using HotStar *Taq*. This indicates failure of HotStar *Taq* to amplify these targets and requires an optimisation of the protocol before HotStar *Taq* can be applied in the further study of NGS-based GMO screening approach. In *chapter 7*, it was shown that the ALF approach, coupled to NGS and data analysis could identify UGMOs with partial unknown sequences in a sample. In the ALF approach, the longest enriched sequence was ~ 2kbp, but Fraiture *et al*. (2017) showed that with the APAgene GOLD Genome walking approach kit enriched sequences of ~15 kbp could be obtained [271]. However, the approach described by Fraiture et al. (2017) requires a more elaborate protocol with a higher number of semi-nested PCRs, increasing the risk of cross contamination in the laboratory. Future research could explore the enhanced enrichment of different polymerases to obtain longer fragments as was shown by Fraiture *et al*. (2017), but with a reduced risk of cross-contamination. However, in another study by Fraiture *et al*. (2018) the maximum enrichment was also around 2 kbp, implying that the longer enrichment may not always be possible on the basis of the APAgene genome walking kit [325]. Furthermore, for UGMO identification, it will be interesting to know whether long enriched fragments

can be obtained from real-life GMOs samples (food or feed products), which undergo varies treatments in the preparation of the product, where the DNA is fragmented. Such a method, aiming at elucidating unknown sequences from degraded DNA, could combine insights from *chapter 3* and *4*, for identification of sequences in samples from various matrices containing degraded DNA, and *chapter 6* and *7* on GM detection and identification.

8.3.2.2 Next generation sequencing

The developed AM-SEQ approach in *chapter 6* was found to be efficient in identifying multiple GMOs in a sample, as well as obtaining indication for the presence of UMGOs. Nonetheless, NGS-based approaches for GMO detection are a relatively new concept and an alternative screening approach needs to be sufficiently sensitive to allow effective enforcement of GMO regulatory requirements. Therefore, before application in a real-life set up can be considered, future studies need to address the sensitivity, repeatability and reproducibility of the NGS-based method by screening well-characterised samples that consist of a mixture of high and low abundant targets. An important issue in the application of an NGS based screening procedure is the data-analysis. To reliable identify GMO sequences targets at low abundance, it is important that background noise can be separated from true amplification. Background noise can potentially cause false-positive identification, and may be the result of either neighbouring cluster overlapping or cross-contamination of the indexed libraries. The latter kind of cross-contamination may be overcome by using double indexing on the Illumina platform [276] , instead of the single indexing used in *chapter 6*. However, single indexing in Illumina is widely applied and cut-offs are commonly applied in other studies to avoid false-positive identification [181,198].

In the ALF approach, long sequences were generated and PacBio is an apt technology to sequence these long reads. In a recent study, a new NGS platform from Oxford Nanopore Technologies, the MinION, was coupled to an enrichment approach for UGMO detection [326]. The time required to sequence the enriched products and data process was 2 hours, which is very fast. In the experiment 99% of the obtained reads could be mapped to the reference sequence [326]. In future studies the ALF approach could also be coupled to the MinION, for more rapid identification of UGMOs. Theoretically, the MioION technology may be applied on-site as the sequence device is portable and the run time is limited. An additional automated and web-based NGS data analysis workflow will further improve the approach and will aid laboratories to enforce GMO regulations.

8.3.2.3 Identification of GMOs and UGMOs

For identification of GMOs and UGMOs using NGS strategies, it is also of high importance that an adequate NGS data-analysis workflow is available, especially in case of UGMO identification were annotating an unknown sequence is part of the data analysis. In *chapter 6 and 7*, automated data analysis pipelines were presented for GMO and UGMO detection, respectively. However, using these pipelines requires basic bioinformatic knowledge, and not all the end users might have this expertise. Alternatively, a web based data analysing module could be developed for GMO and UGMO identification, similar to the CITESdetectionpipeline that was developed in *chapter 3* for endangered species identification.

Furthermore, using the data analysis pipeline for GMO detection in *chapter 6* all the high abundant targets could be identified. However, not all of the low abundant GMOs could be identified. Future studies have to further assess the sensitivity of the NGS-based GMO screening for the detection of low abundant GMOs in reference and real-life materials. In the present studies, the wider NGS-based GMO screening approach was shown to be as sensitive as the standard TaqMan PCR assay for GMO screening. Based on these results, the application of the NGS=based screening is not expected to result in a loss of sensitivity compared to the current routine screening strategy. Additionally, it was shown

that with the wider NGS-screening additional GMOs could be identified. It can be recommended to already implement the wider screening strategy instead of the two-step TaqMan for GMO screening to improve the resolution of the screening. In chapter 7, for UGMO detection it was also found that the detection and identification of target may be compromised when targets are present at low abundance; the ALF approach was sensitive enough to identify the GMOs at 1%, but at the same time it was shown that not all low abundant GMOs could be identified. Future studies may address the efficiency of the ALF-approach for the detection of low abundant targets (~0.1-1%) in known complex reference samples to allow optimal results in the different types of samples.

8.4 Conclusion

8.4.1 Endangered species

For the first part of the thesis the aim was to explore whether NGS-based strategies allow the simultaneous identification of all species, including endangered species, present in a sample, even in complex samples that may be heavily processed. In chapter 3 and 4 it was shown that the developed multi-locus metabarcoding method provides detailed information on the composition of highly complex experimental mixtures and different types of complex products.

- ***It can be concluded that a combination of universal plant and animal barcode and mini-barcode markers can provide high resolution for species detection, including endangered species, without being necessarily limited by matrix, DNA integrity or species richness of a sample.***

8.4.2 GMOs and UGMOs

In the second part of the thesis the aim was to develop a reliable NGS-based method for the detection and identification of GMOs and UGMOs present in a given sample, regardless of their relative abundance. The strategy should be, on the one hand, based on enrichment of known sequence for GMOs, and on the other hand based on the elucidation of unknown sequences adjacent to identified known sequences in the case of UGMOs. In *chapter 6* it was shown that the developed NGS-based broad GMO screening approach screening has a similar level of sensitivity compared to the currently routinely applied TaqMan PCR screening. Additionally, the more-targets NGS-based approach identified more GMOs and GMO-related sequences compared to the standard screening strategy.

- ***It can be concluded that NGS-based screening for known GMO targets can provide reliable identification of GMOs in feed samples. The broader NGS-based screening facilitated the identification of more GM elements with a similar sensitivity compared to the currently applied routine two-step TaqMan PCR strategy.***

In *chapter 7* it was shown that by using the ALF-NGS-based genome walking approach the partially known sequences of a model UGMO could be further elucidated. The method was shown to be effective in enriching for long fragments and in removing the genomic DNA. Furthermore, the strategy was found to be sensitive enough to identify some of the GMOs present at 1%.

- ***It can be concluded that the ALF NGS-based genome walking approach can more effectively identify previously unknown sequences adjacent to an identified but not explained GMO-related element.***

8.4.3 Final conclusion

The overall objective of this thesis was to use detailed genetic differences to identify species/varieties in complex products based on the application of advanced analytical NGS-based strategies, with a specific focus on the identification of two target groups: endangered species and GMOs. Currently applied screening methods provide limited resolution and flexibility in case of complex samples and are often limited by the coupling to Sanger sequencing that can only sequence single sequences [18]. In recent years the field of molecular biology is increasingly using NGS technologies to identify species, strains, varieties etc [24,25]. The application of NGS-based approaches in food authenticity for the detection and identification of endangered species as well as of GMOs and UGMOs is currently, however, still limited. In this thesis it was shown that NGS-based screening can contribute to an enhanced resolution and quality assurance, even in heavily processed samples, in case of the potential presence of endangered species in products of interest. Additionally, due to the more extensive screening in the NGS-based strategy, more GMOs and related targets could be identified compared to the standard TaqMan PCR screening. In case of the NGS-based genome walking approach, the advantage of using an NGS-based strategy was shown in the simultaneous amplification of multiple amplicons and the enrichment of long sequences, providing better data to also identify UGMOs in specific samples.

- ***It can be concluded that the use of NGS-based methods for screening and identification can provide accurate and reliable information on specific genetic differences related to species/varieties present in complex food or feed products. The screening methods for both endangered species as well as GMOs as have been developed within the frame of the present thesis, have improved the analytical repertoire in both fields of application.***

The results of this thesis highlight the potential of NGS-based strategies in species and GMO identification and show that NGS-based approaches have the potential to be effectively used for food compositional screening. The methods as have been developed in this thesis study will aid customs and regulatory agencies in monitoring food and feed samples for enforcement purposes.

References

1. Wandel M (1997) Food labelling from a consumer perspective. British food journal 99 (6):212-219
2. Bernués A, Olaizola A, Corcoran K (2003) Labelling information demanded by European consumers and relationships with purchasing motives, quality and safety of meat. Meat science 65 (3):1095-1106
3. Bansal S, Singh A, Mangal M, Mangal AK, Kumar S (2017) Food adulteration: Sources, health risks, and detection methods. Critical reviews in food science and nutrition 57 (6):1174-1189
4. Flores‐Munguia M, Bermudez‐Almada M, Vázquez‐Moreno L (2000) A research note: Detection of adulteration in processed traditional meat products. Journal of Muscle Foods 11 (4):319-325
5. Padovan G, De Jong D, Rodrigues L, Marchini J (2003) Detection of adulteration of commercial honey samples by the 13C/12C isotopic ratio. Food Chemistry 82 (4):633-636
6. Sadat A, Mustajab P, Khan IA (2006) Determining the adulteration of natural milk with synthetic milk using ac conductance measurement. Journal of Food Engineering 77 (3):472-477
7. Awad T, Moharram H, Shaltout O, Asker D, Youssef M (2012) Applications of ultrasound in analysis, processing and quality control of food: A review. Food research international 48 (2):410-427
8. Coghlan ML, Haile J, Houston J, Murray DC, White NE, Moolhuijzen P, Bellgard MI, Bunce M (2012) Deep sequencing of plant and animal DNA contained within traditional Chinese medicines reveals legality issues and health safety concerns. PLoS genetics 8 (4):e1002657
9. Coghlan ML, Maker G, Crighton E, Haile J, Murray DC, White NE, Byard RW, Bellgard MI, Mullaney I, Trengove R, Allcock RJ, Nash C, Hoban C, Jarrett K, Edwards R, Musgrave IF, Bunce M (2015) Combined DNA, toxicological and heavy metal analyses provides an auditing toolkit to improve pharmacovigilance of traditional Chinese medicine (TCM). Scientific reports 5
10. Chang C-H, Jang-Liaw N-H, Lin Y-S, Fang Y-C, Shao K-T (2013) Authenticating the use of dried seahorses in the traditional Chinese medicine market in Taiwan using molecular forensics. Journal of Food and Drug Analysis 21 (3):310-316
11. Rosen GE, Smith KF (2010) Summarizing the evidence on the international trade in illegal wildlife. EcoHealth 7 (1):24-32
12. South N, Wyatt T (2011) Comparing illicit trades in wildlife and drugs: an exploratory study. Deviant Behavior 32 (6):538-561
13. Bright SW, Greenland AJ, Halpin CM, Schuch WW, Dunwell JM (1996) Environmental impact from plant biotechnology. Annals of the New York Academy of Sciences 792 (1):99-105
14. Hails RS (2000) Genetically modified plants–the debate continues. Trends in Ecology & Evolution 15 (1):14-18
15. Kalaitzandonakes N, Kaufman J, Miller D (2014) Potential economic impacts of zero thresholds for unapproved GMOs: The EU case. Food Policy 45:146-157
16. Önal A (2007) A review: Current analytical methods for the determination of biogenic amines in foods. Food chemistry 103 (4):1475-1486
17. Herrero M, Simó C, García‐Cañas V, Ibáñez E, Cifuentes A (2012) Foodomics: MS‐based strategies in modern food science and nutrition. Mass spectrometry reviews 31 (1):49-69
18. Staats M, Arulandhu AJ, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E (2016) Advances in DNA metabarcoding for food and wildlife forensic species identification. Analytical and bioanalytical chemistry 408 (17):4615-4630
19. Holst-Jensen A, Rønning SB, Løvseth A, Berdal KG (2003) PCR technology for screening and quantification of genetically modified organisms (GMOs). Analytical and Bioanalytical Chemistry 375 (8):985-993
20. Eurlings M, Lens F, Pakusza C, Peelen T, Wieringa JJ, Gravendeel B (2013) Forensic identification of Indian snakeroot (*Rauvolfia serpentina* Benth. ex Kurz) using DNA barcoding. Journal of Forensic sciences 58 (3):822-830
21. Ramos-Gómez S, Busto MD, Albillos SM, Ortega N (2016) Novel qPCR systems for olive (Olea europaea L.) authentication in oils and food. Food chemistry 194:447-454
22. Scholtens IM, Molenaar B, van Hoof RA, Zaaijer S, Prins TW, Kok EJ (2017) Semiautomated TaqMan PCR screening of GMO labelled samples for (unauthorised) GMOs. Analytical and Bioanalytical Chemistry:1-13

23. Arulandhu AJ, Dijk JP, Dobnik D, Holst-Jensen A, Shi J, Zel J, Kok EJ (2016) DNA enrichment approaches to identify unauthorized genetically modified organisms (GMOs). Analytical and bioanalytical chemistry 408 (17):4575-4593
24. Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annu Rev Genet 38:525-552
25. Taberlet P, Coissac E, Hajibabaei M, Rieseberg LH (2012) Environmental DNA. Molecular Ecology 21 (8):1789-1793
26. Gathier G, van der Niet T, Peelen T, van Vugt RR, Eurlings MC, Gravendeel B (2013) Forensic identification of CITES protected slimming cactus (Hoodia) using DNA barcoding. Journal of Forensic Sciences 58 (6):1467-1471
27. Cheng X, Su X, Chen X, Zhao H, Bo C, Xu J, Bai H, Ning K (2014) Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan. Scientific reports 4:5147
28. Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. PLoS biology 2 (10):e312
29. Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences 74 (11):5088-5090
30. Gissi C, Iannelli F, Pesole G (2008) Evolution of the mitochondrial genome of Metazoa as exemplified by comparison of congeneric species. Heredity 101 (4):301-320
31. Ivanova NV, Zemlak TS, Hanner RH, Hebert PDN (2007) Universal primer cocktails for fish DNA barcoding. Molecular Ecology Notes 7 (4):544-548
32. Palumbi S, Martin A, Romano S, McMillan W, Stice L, Grabowski G (1991) The Simple Fool's Guide to PCR, Version 2.0, privately published document compiled by S. Palumbi. Dept. Zoology, Univ Hawaii, Honolulu, HI 96822
33. Sarri C, Stamatis C, Sarafidou T, Galara I, Godosopoulos V, Kolovos M, Liakou C, Tastsoglou S, Mamuris Z (2014) A new set of 16S rRNA universal primers for identification of animal species. Food Control 43:35-41
34. Meusnier I, Singer GA, Landry J-F, Hickey DA, Hebert PD, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. BMC genomics 9 (1):214
35. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SC (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS one 3 (7):e2802
36. CBOL (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences of the United States of America 106 (31):12794-12797
37. Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington R, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species‐level sampling in three divergent groups of land plants. Molecular Ecology Resources 9 (2):439-457
38. Little DP (2014) A DNA mini‐barcode for land plants. Molecular Ecology Resources 14 (3):437-446
39. Fraiture M-A, Herman P, Taverniers I, De Loose M, Deforce D, Roosens NH (2015) Current and new approaches in GMO detection: challenges and solutions. BioMed research international 2015
40. Holst-Jensen A, Spilsberg B, Arulandhu AJ, Kok E, Shi J, Zel J (2016) Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products. Analytical and bioanalytical chemistry 408 (17):4595-4614
41. Su Z, Ning B, Fang H, Hong H, Perkins R, Tong W, Shi L (2011) Next-generation sequencing and its applications in molecular diagnostics. Expert review of molecular diagnostics 11 (3):333-343
42. Shokralla S, Spall JL, Gibson JF, Hajibabaei M (2012) Next‐generation sequencing technologies for environmental DNA research. Molecular ecology 21 (8):1794-1805
43. Schuster SC (2007) Next-generation sequencing transforms today's biology. Nature methods 5 (1):16
44. Chen R, Dong J, Cui X, Wang W, Yasmeen A, Deng Y, Zeng X, Tang Z (2012) DNA based identification of medicinal materials in Chinese patent medicines. Scientific reports 2:958
45. Li R, Quan S, Yan X, Biswas S, Zhang D, Shi J (2017) Molecular characterization of genetically-modified crops: Challenges and strategies. Biotechnology Advances 35 (2):302-309
46. Yang L, Wang C, Holst-Jensen A, Morisset D, Lin Y, Zhang D (2013) Characterization of GM events by insert knowledge adapted re-sequencing approaches. Scientific reports 3

47. Iyengar A (2014) Forensic DNA analysis for animal protection and biodiversity conservation: A review. Journal for Nature Conservation 22 (3):195-205
48. Galimberti A, De Mattia F, Losa A, Bruni I, Federici S, Casiraghi M, Martellos S, Labra M (2013) DNA barcoding as a new tool for food traceability. Food Research International 50 (1):55-63
49. Fajardo V, González I, Rojas M, García T, Martín R (2010) A review of current PCR-based methodologies for the authentication of meats from game animal species. Trends in Food Science & Technology 21 (8):408-421
50. Wong EH-K, Hanner RH (2008) DNA barcoding detects market substitution in North American seafood. Food Research International 41 (8):828-837
51. Hanner R, Becker S, Ivanova NV, Steinke D (2011) FISH-BOL and seafood identification: Geographically dispersed case studies reveal systemic market substitution across Canada. Mitochondrial DNA 22 (sup1):106-122
52. Yancy HF, Zemlak TS, Mason JA, Washington JD, Tenge BJ, Nguyen N-LT, Barnett JD, Savary WE, Hill WE, Moore MM, Fry FS, Randolph SC, Rogers PL, hebert PD (2008) Potential use of DNA barcodes in regulatory science: applications of the Regulatory Fish Encyclopedia. Journal of Food Protection 71 (1):210-217
53. CITES (2015) CITES. Accessed 06.04 2016
54. Speciesplus (2015) Speciesplus. Accessed 15.10 2015
55. Chen F, Chan H, Wong K-L, Wang J, Yu M-T, But P, Shaw P-C (2008) Authentication of Saussurea lappa, an endangered medicinal material, by ITS DNA and 5S rRNA sequencing. Planta medica 74 (8):889-892
56. Lahaye R, Van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the floras of biodiversity hotspots. Proceedings of the National Academy of Sciences 105 (8):2923-2928
57. Glenn TC (2011) Field guide to next-generation DNA sequencers. Molecular Ecology Resources 11 (5):759-769
58. Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA, Larsen TH, Hsu WW, Benedick S, Hamer KC, Wilcove DS, Bruce C, Wang X, Levi T, Lott M, Emerson BC, Yu DW (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology letters 16 (10):1245-1257
59. Lammers Y, Peelen T, Vos RA, Gravendeel B (2014) The HTS barcode checker pipeline, a tool for automated detection of illegally traded species from high-throughput sequencing data. BMC bioinformatics 15:44
60. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21 (8):2045-2050
61. Nielsen UN, Wall DH (2013) The future of soil invertebrate communities in polar regions: different climate change responses in the Arctic and Antarctic? Ecology letters 16 (3):409-419
62. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". Proceedings of the National Academy of Sciences 103 (32):12115-12120
63. Gibson J, Shokralla S, Porter TM, King I, van Konynenburg S, Janzen DH, Hallwachs W, Hajibabaei M (2014) Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proceedings of the National Academy of Sciences 111 (22):8007-8012
64. Cheng X, Su X, Chen X, Zhao H, Bo C, Xu J, Bai H, Ning K (2014) Biological ingredient analysis of traditional Chinese medicine preparation based on high-throughput sequencing: the story for Liuwei Dihuang Wan. Scientific Reports 4: 5147
65. Tillmar AO, Dell'Amico B, Welander J, Holmlund G (2013) A universal method for species identification of mammals utilizing next generation sequencing for the analysis of DNA mixtures. PloS one 8 (12):e83761
66. Särkinen T, Staats M, Richardson JE, Cowan RS, Bakker FT (2012) How to open the treasure chest? Optimising DNA extraction from herbarium specimens. PloS one 7 (8):e43808
67. Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PloS one 6 (5):e19254
68. Ficetola GF, Coissac E, Zundel S, Riaz T, Shehzad W, Bessiere J, Taberlet P, Pompanon F (2010) An in silico approach for the evaluation of DNA barcodes. BMC Genomics 11:434

69. Piñol J, Mir G, Gomez‐Polo P, Agustí N (2015) Universal and blocking primer mismatches limit the use of high‐throughput DNA sequencing for the quantitative metabarcoding of arthropods. Molecular ecology resources 15 (4):819-830
70. Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. PloS one 10 (7):e0130324
71. Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21 (8):1834-1847
72. Valentini A, Pompanon F, Taberlet P (2009) DNA barcoding for ecologists. Trends Ecol Evol 24 (2):110-117
73. Ortea I, Pascoal A, Canas B, Gallardo JM, Barros-Velazquez J, Calo-Mata P (2012) Food authentication of commercially-relevant shrimp and prawn species: from classical methods to Foodomics. Electrophoresis 33 (15):2201-2211
74. Nicolè S, Negrisolo E, Eccher G, Mantovani R, Patarnello T, Erickson DL, Kress WJ, Barcaccia G (2012) DNA barcoding as a reliable method for the authentication of commercial seafood products. Food Technol Biotech 50:387-398
75. Alacs EA, Georges A, FitzSimmons NN, Robertson J (2009) DNA detective: a review of molecular approaches to wildlife forensics. Forensic Science, Medicine, and Pathology 6 (3):180-194
76. Veldman S, Otieno J, Gravendeel B, Andel Tv, Boer Hd (2014) Conservation of Endangered Wild Harvested Medicinal Plants: Use of DNA Barcoding. Novel Plant Bioresources: Applications in Food, Medicine and Cosmetics:81-88
77. Hebert PD, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Systematic biology 54 (5):852-859
78. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. Journal of molecular biology 215 (3):403-410
79. Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004) Identification of birds through DNA barcodes. PLoS biology 2:1657-1663
80. Ratnasingham S, Hebert PD (2007) BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Molecular ecology notes 7 (3):355-364
81. Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopedia of life: an introduction to DNA barcoding. Philosophical transactions of the Royal Society of London Series B, Biological sciences 360 (1462):1805-1811
82. Kress WJ, Erickson DL (2008) DNA barcodes: genes, genomics, and bioinformatics. Proceedings of the National Academy of Sciences 105 (8):2761-2762
83. Bucklin A, Steinke D, Blanco-Bercial L (2011) DNA barcoding of marine metazoa. Annual review of marine science 3:471-508
84. Bhargava M, Sharma A (2013) DNA barcoding in plants: evolution and applications of in silico approaches and resources. Molecular phylogenetics and evolution 67 (3):631-641
85. Kvist S (2013) Barcoding in the dark: a critical view of the sufficiency of zoological DNA barcoding databases and a plea for broader integration of taxonomic knowledge. Molecular phylogenetics and evolution 69 (1):39-45
86. Sandionigi A, Galimberti A, Labra M, Ferri E, Panunzi E, De Mattia F, Casiraghi M (2012) Analytical approaches for DNA barcoding data – how to find a way for plants? Plant Biosystems - An International Journal Dealing with all Aspects of Plant Biology 146 (4):805-813
87. Bernardo GD, Gaudio SD, Galderisi U, Cascino A, Cipollaro M (2007) Comparative evaluation of different DNA extraction procedures from food samples. Biotechnology progress 23 (2):297-301
88. Fernandes TJ, Oliveira MBP, Mafra I (2013) Tracing transgenic maize as affected by breadmaking process and raw material for the production of a traditional maize bread, broa. Food chemistry 138 (1):687-692
89. Gryson N (2010) Effect of food processing on plant DNA degradation and PCR-based GMO analysis: a review. Analytical and bioanalytical chemistry 396 (6):2003-2022
90. Olexová L, Dovičovičová Ľ, Kuchta T (2004) Comparison of three types of methods for the isolation of DNA from flours, biscuits and instant paps. European Food Research and Technology 218 (4):390-393
91. Ivanova NV, Dewaard JR, Hebert PD (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Molecular Ecology Notes 6 (4):998-1002

92. Bergerová E, Hrnčírová Z, Stankovská M, Lopašovská M, Siekel P (2010) Effect of thermal treatment on the amplification and quantification of transgenic and non-transgenic soybean and maize DNA. Food analytical methods 3 (3):211-218
93. Rasmussen RS, Morrissey MT, Hebert PD (2009) DNA barcoding of commercially important salmon and trout species (*Oncorhynchus* and *Salmo*) from North America. Journal of agricultural and food chemistry 57 (18):8379-8385
94. Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B: Biological Sciences 270 (1512):313-321
95. Trontelj P, Machino Y, Sket B (2005) Phylogenetic and phylogeographic relationships in the crayfish genus *Austropotamobius* inferred from mitochondrial COI gene sequences. Molecular phylogenetics and evolution 34 (1):212-226
96. Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD (2005) DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B: Biological Sciences 360 (1462):1847-1857
97. Barrett RD, Hebert PD (2005) Identifying spiders through DNA barcodes. Canadian Journal of Zoology 83 (3):481-491
98. Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PD (2006) DNA barcodes distinguish species of tropical Lepidoptera. Proceedings of the National Academy of Sciences of the United States of America 103 (4):968-971
99. Nagy ZT, Sonet G, Glaw F, Vences M (2012) First large-scale DNA barcoding assessment of reptiles in the biodiversity hotspot of Madagascar, based on newly designed COI primers. PloS one 7 (3):e34506
100. Shearer T, Van Oppen M, Romano S, Wörheide G (2002) Slow mitochondrial DNA sequence evolution in the Anthozoa (Cnidaria). Molecular Ecology 11 (12):2475-2487
101. Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. PLoS biology 3 (12):e422
102. Wörheide G (2006) Low variation in partial cytochrome oxidase subunit I (COI) mitochondrial sequences in the coralline demosponge *Astrosclera willeyana* across the Indo-Pacific. Marine Biology 148 (5):907-912
103. Huang D, Meier R, Todd PA, Chou LM (2008) Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. Journal of Molecular Evolution 66 (2):167-174
104. D'Amato ME, Alechine E, Cloete KW, Davison S, Corach D (2013) Where is the game? Wild meat products authentication in South Africa: a case study. Investigative genetics 4 (1):6
105. Cai Y, Zhang L, Shen F, Zhang W, Hou R, Yue B, Li J, Zhang Z (2011) DNA barcoding of 18 species of Bovidae. Chinese Science Bulletin 56 (2):164-168
106. Bitanyi S, Bjornstad G, Ernest EM, Nesje M, Kusiluka LJ, Keyyu JD, Mdegela RH, Roed KH (2011) Species identification of Tanzanian antelopes using DNA barcoding. Mol Ecol Resour 11 (3):442-449
107. Luo A, Zhang A, Ho SY, Xu W, Zhang Y, Shi W, Cameron SL, Zhu C (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. BMC Genomics 12:84
108. Sevilla RG, Diez A, Norén M, Mouchel O, Jérôme M, Verrez-bagnis V, Van Pelt H, Favre-krey L, Krey G, consortium Tf, Bautista JM (2007) Primers and polymerase chain reaction conditions for DNA barcoding teleost fish based on the mitochondrial cytochrome *b* and nuclear rhodopsin genes. Molecular Ecology Notes 7 (5):730-734
109. Shokralla S, Zhou X, Janzen DH, Hallwachs W, Landry JF, Jacobus LM, Hajibabaei M (2011) Pyrosequencing for mini-barcoding of fresh and old museum specimens. PloS one 6 (7):e21252
110. Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Front Zool 10 (1):34
111. Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Molecular marine biology and biotechnology 3 (5):294-299
112. Karlsson AO, Holmlund G (2007) Identification of mammal species using species-specific DNA pyrosequencing. Forensic science international 173 (1):16-20
113. Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC (1989) Dynamics of mitochondrial DNA evolution in animals: amplification and sequencing with conserved primers. Proceedings of the National Academy of Sciences 86 (16):6196-6200

114. Parson W, Pegoraro K, Niederstätter H, Föger M, Steinlechner M (2000) Species identification by means of the cytochrome *b* gene. International journal of legal medicine 114 (1-2):23-28
115. Elbrecht V, Taberlet P, Dejean T, Valentini A, Usseglio-polatera P, Beisel J-n, Coissac E, Boyer F, Leese F (2016) Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects. PeerJ Preprints 4:e1855v1851
116. Horreo JL, Ardura A, Pola IG, Martinez JL, Garcia-Vazquez E (2013) Universal primers for species authentication of animal foodstuff in a single polymerase chain reaction. Journal of the science of food and agriculture 93 (2):354-361
117. Kitano T, Umetsu K, Tian W, Osawa M (2007) Two universal primer sets for species identification among vertebrates. International journal of legal medicine 121 (5):423-427
118. Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolved rapidly in structure, but slowly in sequence. Journal of Molecular evolution 28 (1-2):87-97
119. Newmaster SG, Grguric M, Shanmughanandhan D, Ramalingam S, Ragupathy S (2013) DNA barcoding detects contamination and substitution in North American herbal products. BMC medicine 11 (1):222
120. Hilu K, Liang H (1997) The *matK* gene: sequence variation and application in plant systematics. American journal of botany 84 (6):830-839
121. Ogden R, McGough HN, Cowan RS, Chua L, Groves M, McEwing R (2009) SNP-based method for the genetic identification of ramin Gonystylus spp. timber and products: applied research meeting CITES enforcement needs. Endangered Species Research 9 (3):255-261
122. Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding *rbcL* gene complements the non-coding *trnH-psbA* spacer region. PloS one 2 (6):e508
123. Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PloS one 5 (1):e8613
124. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proceedings of the National Academy of Sciences 102 (23):8369-8374
125. CBOL P, Hollingsworth P, Forrest L, Spouge J, Hajibabaei M, Ratnasingham S, van der Bank M, Chase M, Cowan R, Erickson D, Fazekas A, Graham S, James K, Kim K-J, Kress W, Schneider H, van AlphenStahl J, Barrett S, van den Berg C, Bogarin D, Burgess K, Cameron K, Carine M, Chacón J, Clark A, Clarkson J, Conrad F, Devey D, Ford C, Hedderson T, Hollingsworth M, Husband B, Kelly L, Kesanakurti P, Kim J, Kim Y, Lahaye R, Lee H, Long D, Madriñán S, Maurin O, Meusnier I, Newmaster S, Park C, Percy D, Petersen G, Richardson J, Salazar G, Savolainen V, Seberg O, Wilkinson M, Yi D, Little D (2009) A DNA barcode for land plants. Proceedings of the National Academy of Sciences 106 (31):12794-12797
126. China Plant BOL G, Li DZ, Gao LM, Li HT, Wang H, Ge XJ, Liu JQ, Chen ZD, Zhou SL, Chen SL, Yang JB, Fu C, Zeng C, Yan H, Zhu Y, Sun Y, Chen S, Zhao L, Wang K, Yang T, Duan G (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proceedings of the National Academy of Sciences 108 (49):19641-19646
127. Shaw J, Lickey EB, Schilling EE, Small RL (2007) Comparison of whole chloroplast genome sequences to choose noncoding regions for phylogenetic studies in angiosperms: the tortoise and the hare III. American journal of botany 94 (3):275-288
128. Gonzalez MA, Baraloto C, Engel J, Mori SA, Pétronelli P, Riéra B, Roger A, Thébaud C, Chave J (2009) Identification of Amazonian trees with DNA barcodes. PloS one 4 (10):e7483
129. Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E (2009) Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama. Proceedings of the National Academy of Sciences 106 (44):18621-18626
130. Devey DS, Chase MW, Clarkson JJ (2009) A stuttering start to plant DNA barcoding: microsatellites present a previously overlooked problem in non-coding plastid regions. Taxon 58 (1):7-15
131. Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Pääbo S (1998) Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. Science 281 (5375):402-406
132. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic acids research 35 (3):e14

133. Faria M, Magalhães A, Nunes M, Oliveira M (2013) High resolution melting of trnL amplicons in fruit juices authentication. Food control 33 (1):136-141
134. De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P (2014) DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Molecular ecology resources 14 (2):306-323
135. Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proceedings of the National Academy of Sciences 101 (41):14812-14817
136. Hoareau TB, Boissin E (2010) Design of phylum-specific hybrid primers for DNA barcoding: addressing the need for efficient COI amplification in the Echinodermata. Molecular Ecology Resources 10 (6):960-967
137. Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for mitochondrial cytochrome *c* oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular ecology resources 13 (5):851-861
138. Irwin DM, Kocher TD, Wilson AC (1991) Evolution of the cytochrome *b* gene of mammals. Journal of molecular evolution 32 (2):128-144
139. Palumbi S (1991) Simple fool's guide to PCR.
140. Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. Biology letters 10 (9)
141. Riaz T, Shehzad W, Viari A, Pompanon F, Taberlet P, Coissac E (2011) ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. Nucleic acids research 39 (21):e145
142. Levin RA, Wagner WL, Hoch PC, Nepokroeff M, Pires JC, Zimmer EA, Sytsma KJ (2003) Family-level relationships of Onagraceae based on chloroplast *rbcL* and *ndhF* data. American Journal of Botany 90 (1):107-115
143. Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM (2012) DNA barcoding methods for land plants. Methods in molecular biology 858:223-252
144. Cuénoud P, Savolainen V, Chatrou LW, Powell M, Grayer RJ, Chase MW (2002) Molecular phylogenetics of Caryophyllales based on nuclear 18S rDNA and plastid *rbcL*, *atpB*, and *matK* DNA sequences. American Journal of Botany 89 (1):132-144
145. White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. PCR protocols: a guide to methods and applications 18:315-322
146. Sun Y, Skinner D, Liang G, Hulbert S (1994) Phylogenetic analysis of Sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. Theoretical and Applied Genetics 89 (1):26-32
147. Sang T, Crawford D, Stuessy T (1997) Chloroplast DNA phylogeny, reticulate evolution, and biogeography of *Paeonia* (Paeoniaceae). American Journal of Botany 84 (9):1120-1136
148. Tate JA, Simpson BB (2003) Paraphyly of *Tarasa* (Malvaceae) and diverse origins of the polyploid species. Systematic Botany 28 (4):723-737
149. Taberlet P, Gielly L, Pautou G, Bouvet J (1991) Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant molecular biology 17 (5):1105-1109
150. Quail M, Smith ME, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers. BMC Genomics 13 (1):341
151. Metzker ML (2010) Sequencing technologies-the next generation. Nature reviews genetics 11 (1):31-46
152. Hajibabaei M, Shokralla S, Zhou X, Singer GA, Baird DJ (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PloS one 6 (4):e17497
153. Carew ME, Pettigrove VJ, Metzeling L, Hoffmann AA (2013) Environmental monitoring using next generation sequencing: rapid identification of macroinvertebrate bioindicator species. Frontiers in Zoology 10 (1):45
154. Bertolini F, Ghionda MC, D'Alessandro E, Geraci C, Chiofalo V, Fontanesi L (2015) A next generation semiconductor based sequencing approach for the identification of meat species in DNA mixtures. PloS one 10 (4):e0121701

155. Loman NJ, Misra RV, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. Nature biotechnology 30 (5):434-439
156. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, Harkins TT, Cookson BT, Hoffman NG (2014) Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Applied and environmental microbiology 80 (24):7583-7591
157. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert J, Smith G, Knight R (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME journal 6 (8):1621-1624
158. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD (2013) Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Applied and environmental microbiology 79 (17):5112-5120
159. Hamady M, Walker JJ, Harris JK, Gold NJ, Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nature methods 5 (3):235-237
160. Smith AM, Heisler LE, Onge RPS, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N, Nislow C (2010) Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. Nucleic acids research 38 (13):e142
161. Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nature methods 10 (1):57-59
162. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28 (12):1647-1649
163. Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27 (6):863-864
164. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics:btu170
165. Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28 (23):3150-3152
166. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research 25 (17):3389-3402
167. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10 (10):996-998
168. Stoeckle MY, Gamble CC, Kirpekar R, Young G, Ahmed S, Little DP (2011) Commercial teas highlight plant DNA barcode identification successes and obstacles. Scientific reports 1 (42)
169. Song J, Shi L, Li D, Sun Y, Niu Y, Chen Z, Luo H, Pang X, Sun Z, Liu C, Lv A, Deng Y, Larson-Rabin Z, Wilkinson M, Chen S (2012) Extensive pyrosequencing reveals frequent intra-genomic variations of internal transcribed spacer regions of nuclear ribosomal DNA. PLoS 7 (8):e43971
170. Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4 (4):406-425
171. Jones M, Ghoorah A, Blaxter M (2011) jMOTU and taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. PloS one 6 (4):e19259
172. Kumar S, Carlsen T, Mevik B-H, Enger P, Blaalid R, Shalchian-Tabrizi K, Kauserud H (2011) CLOTU: an online pipeline for processing and clustering of 454 amplicon reads into OTUs followed by taxonomic annotation. BMC bioinformatics 12 (1):182
173. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nature methods 7 (5):335-336
174. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF (2009) Introducing

mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and environmental microbiology 75 (23):7537-7541
175. Yu DW, Ji Y, Emerson BC, Wang X, Ye C, Yang C, Ding Z (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods in Ecology and Evolution 3 (4):613-623
176. Ratnasingham S, Hebert PD (2013) A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. PloS one 8 (7):e66213
177. De Mattia F, Gentili R, Bruni I, Galimberti A, Sgorbati S, Casiraghi M, Labra M (2012) A multi-marker DNA barcoding approach to save time and resources in vegetation surveys. Botanical Journal of the Linnean Society 169 (3):518-529
178. Luo A, Zhang A, Ho SY, Xu W, Zhang Y, Shi W, Cameron SL, Zhu C (2011) Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. BMC genomics 12 (1):84
179. Lee SY, Ng WL, Mahat MN, Nazre M, Mohamed R (2016) DNA Barcoding of the Endangered Aquilaria (Thymelaeaceae) and Its Application in Species Authentication of Agarwood Products Traded in the Market. PloS one 11 (4):e0154631
180. Milner-Gulland E, Bukreeva O, Coulson T, Lushchekina A, Kholodova M, Bekenov A, Grachev IA (2003) Conservation: Reproductive collapse in saiga antelope harems. Nature 422 (6928):135-135
181. Ivanova NV, Kuzmina ML, Braukmann TWA, Borisenko AV, Zakharov EV (2016) Authentication of Herbal Supplements Using Next-Generation Sequencing. PloS one 11 (5):e0156426
182. Fahner NA, Shokralla S, Baird DJ, Hajibabaei M (2016) Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution, and annotation of four DNA markers. PloS one 11 (6):e0157505
183. Arulandhu AJ, Staats M, Peelen T, Kok EJ DNA metabarcoding of endangered plant and animal species in seized forensic samples. In: Genome, 2015. pp 188-189
184. Taylor H, Harris W (2012) An emergent science on the brink of irrelevance: a review of the past 8 years of DNA barcoding. Molecular Ecology Resources 12 (3):377-388
185. Parveen I, Gafner S, Techen N, Murch SJ, Khan IA (2016) DNA Barcoding for the Identification of Botanicals in Herbal Medicine and Dietary Supplements: Strengths and Limitations. Planta Medica 82 (14):1225-1235
186. Scholtens I, Laurensse E, Molenaar B, Zaaijer S, Gaballo H, Boleij P, Bak A, Kok E (2013) Practical experiences with an extended screening strategy for genetically modified organisms (GMOs) in real-life samples. Journal of agricultural and food chemistry 61 (38):9097-9109
187. Murray M, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic acids research 8 (19):4321-4326
188. Fazekas AJ, Kuzmina ML, Newmaster SG, Hollingsworth PM (2012) DNA barcoding methods for land plants. Methods in molecular biology 858:223-252
189. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19):2460-2461
190. Manning J, Boatwright JS, Daru BH, Maurin O, Bank Mvd (2014) A molecular phylogeny and generic classification of Asphodelaceae subfamily Alooideae: a final resolution of the prickly issue of polyphyly in the alooids? Systematic Botany 39 (1):55-74
191. Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y, Yu DW, Zhou X (2013) SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. Methods in Ecology and Evolution 4 (12):1142-1150
192. Tang M, Hardman CJ, Ji Y, Meng G, Liu S, Tan M, Yang S, Moss ED, Wang J, Yang C, Bruce C, Nevard T, Potts SG, Zhou X, Yu DW (2015) High‐throughput monitoring of wild bee diversity and abundance via mitogenomics. Methods in Ecology and Evolution 6 (9):1034-1043
193. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17 (1):10-12
194. Picard J (2013) Can We Estimate the Global Scale and Impact of Illicit Trade? Convergence: Illicit Networks and National Security in the Age of Globalization:37-60
195. Byard RW, Musgrave I (2010) Herbal medicines and forensic investigations. Springer,
196. Subedi A, Kunwar B, Choi Y, Dai Y, van Andel T, Chaudhary RP, de Boer HJ, Gravendeel B (2013) Collection and trade of wild-harvested orchids in Nepal. Journal of ethnobiology and ethnomedicine 9 (1):64

197. De Boer HJ, Ghorbani A, Manzanilla V, Raclariu A-C, Kreziou A, Ounjai S, Osathanunkul M, Gravendeel B (2017) DNA metabarcoding of orchid-derived products reveals widespread illegal orchid trade. Proc R Soc B 284 (1863):20171182
198. Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale C, Niekamp H, Perri E, Barbante A, Rosec J-P, Seyfarth R, Sovová T, Van Moorleghem C, van Ruth S, Peelen T, Kok E (2017) Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. GigaScience 6 (10):1-18
199. Raclariu AC, Paltinean R, Vlase L, Labarre A, Manzanilla V, Ichim M, Crisan G, Brysting A, de Boer H (2017a) Comparative authentication of Hypericum perforatum herbal products using DNA metabarcoding, TLC and HPLC-MS. Scientific reports 7 (1):1291
200. Raclariu A.C, Mocan A, Popa MO, Vlase L, Ichim MC, Crisan G, Brysting AK, de Boer H (2017b) Veronica officinalis product authentication using DNA metabarcoding and HPLC-MS reveals widespread adulteration with Veronica chamaedrys. Frontiers in Pharmacology 8
201. World Health Organization (2005) National policy on traditional medicine and regulation of herbal medicines: Report of a WHO global survey.
202. Sgamma T, Lockie-Williams C, Kreuzer M, Williams S, Scheyhing U, Koch E, Slater A, Howard C (2017) DNA barcoding for industrial quality assurance. Planta Medica
203. James C (2014) Global Status of Commercialized Biotech/GM Crops 2014. ISAAA Brief 46
204. Kleter GA, Kok EJ (2010) Safety assessment of biotechnology used in animal production, including genetically modified (GM) feed and GM animals – a review. Animal Science Papers and Reports 28 (2):105-114
205. Brod FC, van Dijk JP, Voorhuijzen MM, Dinon AZ, Guimaraes LH, Scholtens IM, Arisi AC, Kok EJ (2014) A high-throughput method for GMO multi-detection using a microfluidic dynamic array. Anal Bioanal Chem 406 (5):1397-1410
206. Randhawa GJ, Morisset D, Singh M, Žel J (2014) GMO matrix: A cost-effective approach for screening unauthorized genetically modified events in India. Food Control 38:124-129
207. Alexander J. Stein, Emilio Rodríguez-Cerezo (2010) Low-Level Presence of New GM Crops: An Issue on the Rise for Countries Where They Lack Approval. AgBioForum 13 (2):173-182
208. Commission E (2011) Regulation (EU) No 619/2011 of 24 June 2011 laying down the methods of sampling and analysis for the official control of feed as regards presence of genetically modified material for which an authorisation procedure is pending or the authorisation of which has expired. Off J Eur Union 166:9-15
209. Holst-Jensen A, Bertheau Y, de Loose M, Grohmann L, Hamels S, Hougs L, Morisset D, Pecoraro S, Pla M, Van den Bulcke M, Wulff D (2012) Detecting un-authorized genetically modified organisms (GMOs) and derived materials. Biotechnology advances 30 (6):1318-1335
210. Regulation EC (2003) No. 1829/2003 of the European Parliament and of the Council of 22nd September 2003 on genetically modified food and feed. Off J Eur Union Lt 268:1-23
211. European Network of GMO Laboratories (2009) Definition of Minimum Performance Requirements for Analytical Methods of GMO Testing.
212. Holst-Jensen A, Bertheau Y, Allnutt T, Broll H, De Loose M, Grohmann L, Henry C, Hougs L, Moens W, Morisset D, Ovesna J (2011) Overview on the detection, interpretation and reporting on the presence of unauthorised genetically modified materials. EUR 25008 EN
213. Heide BR, Heir E, Holck A (2007) Detection of eight GMO maize events by qualitative, multiplex PCR and fluorescence capillary gel electrophoresis. European Food Research and Technology 227 (2):527-535
214. Morisset D, Stebih D, Milavec M, Gruden K, Zel J (2013) Quantitative analysis of food and feed samples with droplet digital PCR. PloS one 8 (5):e62583
215. Leimanis S, Hernandez M, Fernandez S, Boyer F, Burns M, Bruderer S, Glouden T, Harris N, Kaeppeli O, Philipp P, Pla M, Puigdomenech P, Vaitilingom M, Bertheau Y, Remacle J (2006) A microarray-based detection system for genetically modified (GM) food ingredients. Plant Mol Biol 61 (1-2):123-139
216. Prins TW, van Dijk JP, Beenen HG, Van Hoef AA, Voorhuijzen MM, Schoen CD, Aarts HJ, Kok EJ (2008) Optimised padlock probe ligation and microarray detection of multiple (non-authorised) GMOs in a single reaction. BMC Genomics 9 (1):584

217. Block A, Debode F, Grohmann L, Hulin J, Taverniers I, Kluga L, Barbau-Piednoir E, Broeders S, Huber I, Van den Bulcke M, Heinze P, Berben G, Busch U, Roosens N, Janssen E, Zel J, Gruden K, Morisset D (2013) The GMOseek matrix: a decision support tool for optimizing the detection of genetically modified plants. BMC bioinformatics 14 (1):256
218. Ujhelyi G, Dijk JP, Prins TW, Voorhuijzen MM, Hoef AM, Beenen HG, Morisset D, Gruden K, Kok EJ (2012) Comparison and transfer testing of multiplex ligation detection methods for GM plants. BMC biotechnology 12 (1):4
219. Liang C, van Dijk JP, Scholtens IM, Staats M, Prins TW, Voorhuijzen MM, da Silva AM, Arisi ACM, den Dunnen JT, Kok EJ (2014) Detecting authorized and unauthorized genetically modified organisms containing vip3A by real-time PCR and next-generation sequencing. Analytical and bioanalytical chemistry 406 (11):2603-2611
220. Scholtens I, Laurensse E, Molenaar B, Zaaijer S, Gaballo H, Boleij P, Bak A, Kok E (2013) Practical experiences with an extended screening strategy for genetically modified organisms (GMOs) in real-life samples. Journal of agricultural and food chemistry 61 (38):9097-9109
221. Shao N, Jiang S-M, Zhang M, Wang J, Guo S-J, Li Y, Jiang H-W, Liu C-X, Zhang D-B, Yang L-T, Tao S-C (2014) MACRO: a combined microchip-PCR and microarray system for high-throughput monitoring of genetically modified organisms. Analytical chemistry 86 (2):1269-1276
222. Guo J, Yang L, Chen L, Morisset D, Li X, Pan L, Zhang D (2011) MPIC: A High-Throughput Analytical Method for Multiple DNA Targets. Analytical Chemistry 83 (5):1579-1586
223. Volpicella M, Leoni C, Costanza A, Fanizza I, Placido A, Ceci LR (2012) Genome Walking by Next Generation Sequencing Approaches. Biology 1 (3):495-507
224. Gawronski JD, Wong SM, Giannoukos G, Ward DV, Akerley BJ (2009) Tracking insertion mutants within libraries by deep sequencing and a genome-wide screen for Haemophilus genes required in the lung. Proceedings of the National Academy of Sciences 106 (38):16422-16427
225. Gabriel R, Eckenberg R, Paruzynski A, Bartholomae CC, Nowrouzi A, Arens A, Howe SJ, Recchia A, Cattoglio C, Wang W, Faber K, Schwarzwaelder K, Kirsten R, Deichmann A, Ball CR, Balaggan KS, Yanez-Munoz RJ, Ali RR, Gaspar HB, Biasco L, Aiuti A, Cesana D, Montini E, Naldini L, Cohen-Haguenauer O, Mavilio F, Thrasher AJ, Glimm H, von Kalle C, Saurin W, Schmidt M (2009) Comprehensive genomic access to vector integration in clinical gene therapy. Nature medicine 15 (12):1431-1436
226. Babekova R, Funk T, Pecoraro S, Engel K-H, Busch U (2008) Development of an event-specific Real-time PCR detection method for the transgenic Bt rice line KMD1. European Food Research and Technology 228 (5):707-716
227. Tan G, Gao Y, Shi M, Zhang X, He S, Chen Z, An C (2005) SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. Nucleic acids research 33 (13):e122
228. Spalinskas R, Bulcke M, Eede G, Milcamps A (2013) LT-RADE: An Efficient User-Friendly Genome Walking Method Applied to the Molecular Characterization of the Insertion Site of Genetically Modified Maize MON810 and Rice LLRICE62. Food Analytical Methods 6 (2):705-713
229. Schmidt M, Zickler P, Hoffmann G, Haas S, Wissler M, Muessig A, Tisdale JF, Kuramoto K, Andrews RG, Wu T, Kiem HP, Dunbar CE, von Kalle C (2002) Polyclonal long-term repopulating stem cell clones in a primate model. Blood 100 (8):2737-2743
230. Thirulogachandar V, Pandey P, Vaishnavi CS, Reddy MK (2011) An affinity-based genome walking method to find transgene integration loci in transgenic genome. Analytical biochemistry 416 (2):196-201
231. Xu W, Shang Y, Zhu P, Zhai Z, He J, Huang K, Luo Y (2013) Randomly broken fragment PCR with 5' end-directed adaptor for genome walking. Scientific reports 3
232. Trinh Q, Xu W, Shi H, Luo Y, Huang K (2012) An A-T linker adapter polymerase chain reaction method for chromosome walking without restriction site cloning bias. Analytical biochemistry 425 (1):62-67
233. Wang C, Li R, Quan S, Shen P, Zhang D, Shi J, Yang L (2015) GMO detection in food and feed through screening by visual loop-mediated isothermal amplification assays. Analytical and bioanalytical chemistry 407 (16):4829-4834
234. Ochman H, Gerber AS, Hartl DL (1988) Genetic applications of an inverse polymerase chain reaction. Genetics 120 (3):621-623
235. Zimmermann A, Lüthy J, Pauli U (2000) Event specific transgene detection in Bt11 corn by quantitative PCR at the integration site. LWT-Food Science and Technology 33 (3):210-216

236. Liu YG, Mitsukawa N, Oosumi T, Whittier RF (1995) Efficient isolation and mapping of Arabidopsis thaliana T-DNA insert junctions by thermal asymmetric interlaced PCR. The Plant Journal 8 (3):457-463
237. Yang L, Xu S, Pan A, Yin C, Zhang K, Wang Z, Zhou Z, Zhang D (2005) Event specific qualitative and quantitative polymerase chain reaction detection of genetically modified MON863 maize based on the 5'-transgene integration sequence. Journal of agricultural and food chemistry 53 (24):9312-9318
238. Pan A, Yang L, Xu S, Yin C, Zhang K, Wang Z, Zhang D (2006) Event-specific qualitative and quantitative PCR detection of MON863 maize based upon the 3′ -transgene integration sequence. Journal of cereal science 43 (2):250-257
239. Wang W-X, Zhu T-H, Lai F-X, Fu Q (2011) Event-specific qualitative and quantitative detection of transgenic rice Kefeng-6 by characterization of the transgene flanking sequence. European Food Research and Technology 232 (2):297-305
240. Xu J, Cao J, Cao D, Zhao T, Huang X, Zhang P, Luan F (2013) Flanking sequence determination and event-specific detection of genetically modified wheat B73-6-1. Acta biochimica et biophysica Sinica 45 (5):416-421
241. Siebert PD, Chenchik A, Kellogg DE, Lukyanov KA, Lukyanov SA (1995) An improved PCR method for walking in uncloned genomic DNA. Nucleic acids research 23 (6):1087-1088
242. Primers P (2007) High-efficiency thermal asymmetric interlaced PCR for amplification of unknown flanking sequences. Biotechniques 43:649-656
243. Riley J, Butler R, Ogilvie D, Finniear R, Jenner D, Powell S, Anand R, Smith J, Markham A (1990) A novel, rapid method for the isolation of terminal sequences from yeast artificial chromosome (YAC) clones. Nucleic acids research 18 (10):2887-2890
244. Hengen PN (1995) Vectorette, splinkerette and boomerang DNA amplification. Trends in biochemical sciences 20 (9):372-373
245. Orcheski BB, Davis TM (2010) An enhanced method for sequence walking and paralog mining: TOPO(R) Vector-Ligation PCR. BMC research notes 3 (1):61
246. Windels P, Taverniers I, Depicker A, Van Bockstaele E, De Loose M (2001) Characterisation of the Roundup Ready soybean insert. European Food Research and Technology 213 (2):107-112
247. Raymond P, Gendron L, Khalf M, Paul S, Dibley KL, Bhat S, Xie VR, Partis L, Moreau M-E, Dollard C, Coté M-J, Laberge S, Emslie KR (2010) Detection and identification of multiple genetically modified events using DNA insert fingerprinting. Analytical and bioanalytical chemistry 396 (6):2091-2102
248. Yuanxin Y, Chengcai A, Li L, Jiayu G, Guihong T, Zhangliang C (2003) T-linker-specific ligation PCR (T-linker PCR): an advanced PCR technique for chromosome walking or for isolation of tagged DNA ends. Nucleic acids research 31 (12):e68-e68
249. Leoni C, Gallerani R, Ceci L (2008) A genome walking strategy for the identification of eukaryotic nucleotide sequences adjacent to known regions. BioTechniques 44 (2):229-235
250. Rudi K, Fossheim T, Jakobsen aKS (1999) Restriction Cutting Independent Method for Cloning Genomic DNA Segments Outside the Boundaries of Known Sequences. BioTechniques 27:1170-1177
251. Spalinskas R, Van den Bulcke M, Milcamps A (2013) Efficient retrieval of recombinant sequences of GM plants by Cauliflower Mosaic Virus 35S promoter-based bidirectional LT-RADE. European Food Research and Technology 237 (6):1025-1031
252. Schmidt M, Schwarzwaelder K, Bartholomae C, Zaoui K, Ball C, Pilz I, Braun S, Glimm H, von Kalle C (2007) High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). Nat Methods 4 (12):1051-1057
253. Paruzynski A, Arens A, Gabriel R, Bartholomae CC, Scholz S, Wang W, Wolf S, Glimm H, Schmidt M, von Kalle C (2010) Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. Nature protocols 5 (8):1379-1395
254. Fraiture M-A, Herman P, Taverniers I, De Loose M, Deforce D, Roosens NH (2014) An innovative and integrated approach based on DNA walking to identify unauthorised GMOs. Food chemistry 147:60-69
255. Fraiture M-A, Herman P, Lefèvre L, Taverniers I, De Loose M, Deforce D, Roosens NH (2015) Integrated DNA walking system to characterize a broad spectrum of GMOs in food/feed matrices. BMC biotechnology 15 (1):76
256. Williams-Carrier R, Stiffler N, Belcher S, Kroeger T, Stern DB, Monde RA, Coalter R, Barkan A (2010) Use of Illumina sequencing to identify transposon insertions underlying mutant phenotypes in high-copy Mutator lines of maize. The Plant Journal 63 (1):167-177

257. Li J, Zhang JM, Li X, Suo F, Zhang MJ, Hou W, Han J, Du LL (2011) A piggyBac transposon-based mutagenesis system for the fission yeast Schizosaccharomyces pombe. Nucleic acids research 39 (6):e40
258. Langridge GC, Phan MD, Turner DJ, Perkins TT, Parts L, Haase J, Charles I, Maskell DJ, Peters SE, Dougan G, Wain J, Parkhill J, Turner AK (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants. Genome research 19 (12):2308-2316
259. Lepage E, Zampini E, Boyle B, Brisson N (2013) Time- and cost-efficient identification of T-DNA insertion sites through targeted genomic sequencing. PloS one 8 (8):e70912
260. Zastrow-Hayes GM, Lin H, Sigmund AL, Hoffman JL, Alarcon CM, Hayes KR, Richmond TA, Jeddeloh JA, May GD, Beatty MK (2015) Southern-by-Sequencing: A Robust Screening Approach for Molecular Characterization of Genetically Modified Crops. The Plant Genome 8 (1)
261. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD (2007) HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications. Genome research 17 (8):1186-1194
262. Liu S, Yeh CT, Ji T, Ying K, Wu H, Tang HM, Fu Y, Nettleton D, Schnable PS (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genet 5 (11):e1000733
263. van Opijnen T, Bodi KL, Camilli A (2009) Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat Methods 6 (10):767-772
264. Gallagher LA, Shendure J, Manoil C (2011) Genome-scale identification of resistance functions in Pseudomonas aeruginosa using Tn-seq. mBio 2 (1):e00315-00310
265. Taheri A, Robinson SJ, Parkin I, Gruber MY (2012) Revised selection criteria for candidate restriction enzymes in genome walking. PloS one 7 (4):e35117
266. Kanizay LB, Jacobs TB, Gillespie K, Newsome JA, Spaid BN, Parrott WA (2015) HtStuf: High-Throughput Sequencing to Locate Unknown DNA Junction Fragments. The Plant Genome 8 (1)
267. Willems S, Fraiture M-A, Deforce D, De Keersmaecker SC, De Loose M, Ruttink T, Herman P, Van Nieuwerburgh F, Roosens N (2016) Statistical framework for detection of genetically modified organisms based on Next Generation Sequencing. Food chemistry 192:788-798
268. Dobnik D, Spilsberg B, Bogozalec Kosir A, Holst-Jensen A, Zel J (2015) Multiplex Quantification of 12 European Union Authorized Genetically Modified Maize Lines with Droplet Digital Polymerase Chain Reaction. Analytical chemistry 87 (16):8218-8226
269. ISAAA (2016) Global Status of Commercialized Biotech/GM Crops: 2016. ISAAA Brief 52
270. Dobnik D, Spilsberg B, Bogozalec Kosir A, Holst-Jensen A, Zel J (2015) Multiplex Quantification of 12 European Union Authorized Genetically Modified Maize Lines with Droplet Digital Polymerase Chain Reaction. Anal Chem 87 (16):8218-8226. doi:10.1021/acs.analchem.5b01208
271. Fraiture M-A, Herman P, Papazova N, De Loose M, Deforce D, Ruttink T, Roosens NH (2017) An integrated strategy combining DNA walking and NGS to detect GMOs. Food Chemistry 232:351-358
272. Košir AB, Arulandhu AJ, Voorhuijzen MM, Xiao H, Hagelaar R, Staats M, Costessi A, Žel J, Kok EJ, Dijk JPv (2017) ALF: a strategy for identification of unauthorized GMOs in complex mixtures by a GW-NGS method and dedicated bioinformatics analysis. Scientific reports 7 (1)
273. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal 17 (1):pp. 10-12
274. Zhang J, Kobert K, Flouri T, Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30 (5):614-620
275. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow HP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC genomics 13 (1):341
276. Kircher M, Sawyer S, Meyer M (2011) Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic acids research 40 (1):e3-e3
277. Stein AJ, Rodríguez-Cerezo E (2010) Low-level presence of new GM crops: an issue on the rise for countries where they lack approval.
278. Querci M, Foti N, Bogni A, Kluga L, Broll H, Van den Eede G (2009) Real-time PCR-based ready-to-use multi-target analytical system for GMO detection. Food Analytical Methods 2 (4):325-336
279. Neufeld J (2014) The State of Food and Agriculture. Canadian Journal of Comparative Medicine 34
280. National Academies of Sciences E, Medicine (2016) Genetically engineered crops: experiences and prospects. National Academies Press,

281. Prusak A, Rowe G, Strojny J (2014) Is GMO" Sustainable"? A Review of the Environmental Risks of GM Plants in Comparison with Conventional and Organic Crops. Modern Management Review 19 (21 (4)):187-200
282. Dinon AZ, Prins TW, van Dijk JP, Arisi ACM, Scholtens IM, Kok EJ (2011) Development and validation of real-time PCR screening methods for detection of cry1A. 105 and cry2Ab2 genes in genetically modified organisms. Analytical and bioanalytical chemistry 400 (5):1433-1442
283. standardization ECf (2014) NPR-CEN/TS 16707:2014 en. 2016-06 Foodstuffs — Methods of analysis for the detection of genetically modified organisms and derived products — Nucleic acid extraction.
284. Rosa SF, Gatto F, Angers-Loustau A, Petrillo M, Kreysa J, Querci M (2016) Development and applicability of a ready-to-use PCR system for GMO screening. Food chemistry 201:110-119
285. Broeders S, Fraiture M-A, Vandermassen E, Delvoye M, Barbau-Piednoir E, Lievens A, Roosens N (2015) New qualitative trait-specific SYBR® Green qPCR methods to expand the panel of GMO screening methods used in the CoSYPS. European Food Research and Technology 241 (2):275-287
286. Park S-B, Kim H-Y, Kim J-H (2015) Multiplex PCR system to track authorized and unauthorized genetically modified soybean events in food and feed. Food Control 54:47-52
287. Scholtens IM, Molenaar B, van Hoof RA, Zaaijer S, Prins TW, Kok EJ (2017) Semiautomated TaqMan PCR screening of GMO labelled samples for (unauthorised) GMOs. Analytical and Bioanalytical Chemistry 409 (15):3877-3889
288. Fu W, Wei S, Wang C, Du Z, Zhu P, Wu X, Wu G, Zhu S (2017) A temperature-tolerant multiplex elements and genes screening system for genetically modified organisms based on dual priming oligonucleotide primers and capillary electrophoresis. Food chemistry 229:396-402
289. Prins TW, van Hoof RA, Scholtens IM, Kok EJ (2017) Novel TaqMan PCR screening methods for element cry3A and construct gat/T-pinII to support detection of both known and unknown GMOs. European Food Research and Technology 243 (3):481-488
290. Tan G, Gao Y, Shi M, Zhang X, He S, Chen Z, An C (2005) SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. Nucleic acids research 33 (13):e122-e122
291. Fraiture M-A, Herman P, Taverniers I, De Loose M, Van Nieuwerburgh F, Deforce D, Roosens NH (2015) Validation of a sensitive DNA walking strategy to characterise unauthorised GMOs using model food matrices mimicking common rice products. Food chemistry 173:1259-1265
292. Trinh Q, Shi H, Xu W, Hao J, Luo Y, Huang K (2012) Loop‐linker PCR: an advanced PCR technique for genome walking. IUBMB life 64 (10):841-845
293. Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics, proteomics & bioinformatics 13 (5):278-289
294. Hackl T, Hedrich R, Schultz J, Förster F (2014) proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30 (21):3004-3011
295. Ono Y, Asai K, Hamada M (2012) PBSIM: PacBio reads simulator—toward accurate genome assembly. Bioinformatics 29 (1):119-121
296. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB (2013) Characterizing and measuring bias in sequence data. Genome biology 14 (5):R51
297. Hestand MS, Van Houdt J, Cristofoli F, Vermeesch JR (2016) Polymerase specific error rates and profiles identified by single molecule sequencing. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 784:39-45
298. Frohman MA, Dush MK, Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Sciences 85 (23):8998-9002
299. Van De Wiel C, Groeneveld R, Dolstra O, Kok E, Scholtens I, Thissen J, Smulders M, Lotz L (2009) Pollen-mediated gene flow in maize tested for coexistence of GM and non-GM crops in the Netherlands: effect of isolation distances between fields. NJAS-Wageningen Journal of Life Sciences 56 (4):405-423
300. Savini C, Mazzara M, Munaro B, Van den Eede G (2008) Event-specific method for the quantification of cotton line MON15985 using real-time PCR-Validation report and protocol. EURL-GMFF Online Publication
301. Savini C, Bogni A, Grazioli E, Munaro B, Mazzara M, Van den Eede G (2008) Event-specific method for the quantification of maize line MON 89034 using real-time PCR. JRC Scientific and Technical Reports

302. Mazzara M, Grazioli E, Savini C, Van Den Eede G (2009) Report on the Verification of the Performance of a MON810 Event-specific Method on Maize Line MON810 Using Real-time PCR. Validation Report and Protocol. Publications Office of the European Union 10:1-92
303. Delobel C, Foti N, Grazioli E, Mazzara M, Van den Eede G (2013) Event-specific method for the quantification of maize line MON88017 using real-time PCR v. 1.01. Publications Office of the European Union, Luxembourg Google Scholar
304. Cock PJ, Chilton JM, Grüning B, Johnson JE, Soranzo N (2015) NCBI BLAST+ integrated into Galaxy. Gigascience 4 (1):39
305. Afgan E, Baker D, Van den Beek M, Blankenberg D, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning B, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic acids research 44 (W1):W3-W10
306. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26 (6):841-842
307. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature methods 9 (4):357
308. Arumuganathan K, Earle E (1991) Nuclear DNA content of some important plant species. Plant molecular biology reporter 9 (3):208-218
309. Centre ECsJr (2011). Technical guidance document from the European Union Reference Laboratory for Genetically Modified Food and Feed on the implementation of Commission Regulation (EU) 619/2011
310. Goecks J, Nekrutenko A, Taylor J (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 11 (8):R86
311. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome research 15 (10):1451-1455
312. Blankenberg Dea Current Protocols in Molecular Biology, vol Chapter 19, Unit 19.10.1-21. (John Wiley & Sons, Inc., 2010),
313. EUginius - the European GMO reference database. Available at: http://wwweuginiuseu/euginius/pages/homejsf Accessed: 30th December 2016
314. Booker A, Agapouda A, Frommenwiler DA, Scotti F, Reich E, Heinrich M (2017) St John's wort (Hypericum perforatum) products–an assessment of their authenticity and quality. Phytomedicine
315. Ghorbani A, Saeedi Y, de Boer HJ (2017) Unidentifiable by morphology: DNA barcoding of plant material in local markets in Iran. PloS one 12 (4):e0175722
316. Shi Y, Zhao M, Yao H, Yang P, Xin T, Li B, Sun W, Chen S (2017) Rapidly discriminate commercial medicinal Pulsatilla chinensis (Bge.) Regel from its adulterants using ITS2 barcoding and specific PCR-RFLP assay. Scientific reports 7:40000
317. Arulandhu AJ, van Dijk J, Staats M, Hagelaar R, Voorhuijzen M, Molenaar B, van Hoof R, Li R, Yang L, Shi J, Scholtens I, Kok E (2018) NGS-based amplicon sequencing approach; towards a new era in GMO screening and detection. Food Control 93:201-210
318. Paruzynski A, Arens A, Gabriel R, Bartholomae CC, Scholz S, Wang W, Wolf S, Glimm H, Schmidt M, von Kalle C (2010) Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. Nature protocols 5 (8):1379-1395
319. Debode F, Huber I, Macarthur R, Rischitor P, Mazzara M, Herau V, Sebah D, Dobnik D, Broeders S, Roosens N, Busch U, Berben G, Morisset D, Zel J (2017) Inter-laboratory studies for the validation of two singleplex (tE9 and pea lectin) and one duplex (pat/bar) real-time PCR methods for GMO detection. Food Control 73:452-461
320. Žel J, Demšar T, Štebih D, Milavec M, Gruden K (2015) Extraction of DNA from different sample types–a practical approach for GMO testing. Acta BiologicaI Slovenica Ljubljana 58:2
321. Coello RP, Justo JP, Mendoza AF, Ordoñez ES (2017) Comparison of three DNA extraction methods for the detection and quantification of GMO in Ecuadorian manufactured food. BMC research notes 10 (1):758
322. Mishra P, Kumar A, Nagireddy A, Mani DN, Shukla AK, Tiwari R, Sundaresan V (2016) DNA barcoding: an efficient tool to overcome authentication challenges in the herbal market. Plant biotechnology journal 14 (1):8-21

323. Raclariu AC, Heinrich M, Ichim MC, de Boer H (2018) Benefits and limitations of DNA barcoding and metabarcoding in herbal product authentication. Phytochemical Analysis 29 (2):123-128
324. Rach J, Bergmann T, Paknia O, DeSalle R, Schierwater B, Hadrys H (2017) The marker choice: Unexpected resolving power of an unexplored CO1 region for layered DNA barcoding approaches. PloS one 12 (4):e0174842
325. Fraiture M-A, Vandamme J, Herman P, Roosens NH (2018) Development and validation of an integrated DNA walking strategy to detect GMO expressing cry genes. BMC biotechnology 18 (1):40
326. Fraiture M-A, Saltykova A, Hoffman S, Winand R, Deforce D, Vanneste K, De Keersmaecker SC, Roosens NH (2018) Nanopore sequencing technology: a new route for the fast detection of unauthorized GMO. Scientific reports 8 (1):7903

Summary

Next generation DNA sequencing based strategies; towards a new era for the traceability of endangered species and genetically modified organisms

Food products are often composed of multiple ingredients that are in addition generally heavily processed, this makes it very challenging to determine the ingredient composition. Traditional molecular biological techniques, such as, specific PCR followed by Sanger sequencing or TaqMan PCR are most frequently applied to identify species/varieties in food/feed products. In the last decade, next generation sequencing (NGS) technologies have been developed and have been widely applied in medical science and other areas, such as agricultural and environmental sciences. The aim of this thesis was to use detailed genetic differences to identify species/varieties in feed/food products based on advanced analytical NGS based strategies. The study focused on the identification of two target groups: (a) endangered species and (b) GMOs. Elucidating genetic composition was subdivided in three main topics: enrichment, NGS based strategy and identification. For both applications novel molecular assays were developed and coupled to an apt NGS technology, data analysis was performed with a dedicated bioinformatics pipelines that were developed for the specific needs per application.

With respect to endangered species identification, in *chapter 2* it was shown that no dedicated method was available to identify endangered plant and animal species in real-life samples. To address this issue, in *chapter 3*, a multi-locus DNA metabarcoding approach was developed comparing 12 plant and animal barcode and mini-barcode markers, and the method was validated across 16 laboratories. The results showed that the approach was sensitive enough to identify species present at 1% and consistent and reproducible results were observed across the laboratories for all the analysed experimental mixtures and real-life samples. The combination of multiple barcodes enabled the identification of all the species used in the experimental mixtures, and additionally increased the quality assurance for detection. Furthermore, in *chapter 4* the applicability of the multi-locus DNA metabarcoding approach was evaluated on 18 traditional medicines (TMs) belonging to different matrices. It was shown that an adequate DNA clean-up system is necessary to remove impurity from real-life samples, in the metabarcoding analysis of the TMs mainly mini-barcode accounted for the identification of the taxa. Regarding to the identified species in the TMs, only a few declared species on the label could be identified across the TMs, however, many undeclared species were identified in the TMs including the endangered species (*Ursus arctos*). The conclusion for the first part of the thesis was that a combination of universal plant and animal barcode and mini-barcode markers can provide high resolution for species detection, without being limited by matrix, DNA integrity or species composition of a sample.

With respect to the identification of GMOs, the AM-SEQ NGS-based GMOs screening approach was developed and evaluated (*chapter 6*). The obtained results from the NGS based screening were compared to the currently applied two-step TaqMan PCR based GMO screening. This comparison showed that high abundant targets could be detected similarly, however, low abundant targets could not always detected in one of the two methods. With the use of a broader NGS-based screening strategy more GMOs and related targets could be identified compared to the more limited two-step TaqMan PCR based GMO screening. Additionally, some identified low abundant targets could not be explained, which might indicate the presence of Unknown GMOs (UMGOs) or, alternatively, the donor organism. To identify the unknown sequence of a UGMO a genome walking (GW) approach is necessary, and in *chapter 5* the available GW approaches were summarised and from this literature review it was concluded that at that moment no GW method was available to full fill the requirements of UGMOs identification, such as, 0.1% detection limit and enrichment of UGMOs target in a background of GMOs. To address these issues, in *chapter 7*, Amplification of Linearly-enriched Fragments (ALF) approach was developed and combined with PacBio SMRT NGS technology. The ALF approach was subsequently evaluated on real-life mimicking samples, where sequences related to GMOs present at 1% could be

identified. The longest enriched fragment was around 2.5 kbp and a data analysis model was used to distinguish the sequences belonging to known GMOs from the unknown sequences by a sequence of data mapping. With the data analysis model, previous unknown sequence information of a GMO was obtained, showing that the ALF approach can be used to identify the unknown sequence of a UGMO in real-life samples. For the second part of the thesis it was concluded that NGS based GMO screening is an accurate and reliable screening method for GMOs, additionally, the combination of a genome walking approach and NGS is sensitive enough to identify previously unknown sequences for GMO present at low abundance.

In general, it can be concluded that the use of NGS-based screening methods can provide accurate and reliable information on the detailed genetic differences of species/varieties present in complex food/feed products. Using enrichment of known targets both well-known species as well as known and unknown GM sequences could be identified, not limited by the complexity of a sample. The results of this thesis show that NGS-based approaches have the potential to be effectively used for food composition screening, and the developed methods can aid Customs, regulatory agencies, and food industries in monitoring food and feed samples.

About the author

Alfred Joseph Arulandhu was born in Madurai, Tamil Nadu, India, on the 1st of January 1986. In 2009, he obtained his Bachelor's degree in Biotechnology from Jeppiaar Engineering College affiliated to Anna University in Chennai, India. In 2010 he moved to Netherlands where he studied at Wageningen University and obtained his Master's degree in the field of Cellular and Molecular Biology in 2013. In 2014 he continued his study at Wageningen University to obtain a PhD degree. Alfred's PhD was funded by the European project DECATHLON and his research topic was to develop new analytical methods for Food authenticity in the field of GMOs and endangered species identification. The results of this research are described in this thesis.

List of manuscripts and publications

Arulandhu AJ, Staats M, Hagelaar R, Peelen T, Kok EJ**.** "The application of multi-locus DNA metabarcoding in traditional medicines" (submitted)

Arulandhu AJ, van Dijk JP, Staats M, Hagelaar R, Voorhuijzen M, Molenaar B, van Hoof R, Li R, Yang L, Shi J, Scholtens I, Kok EJ. "NGS-based amplicon sequencing approach; towards a new era in GMO screening and detection" *Food Control* 2018; 93: 201-210.

Arulandhu AJ, Staats M, Hagelaar R, Voorhuijzen MM, Prins TW, Scholtens I, Costessi A, Duijsings D, Rechenmann F, Gaspar FB, Barreto Crespo MT, Arne Holst-Jensen A, Birck M, Burns M, Haynes E, Hochegger R, Klingl A, Lundberg L, Natale C, Niekamp H, Perri E, Barbante A, Rosec JP, Seyfarth R, Sovová T, van Moorleghem C, van Ruth S, Peelen T, Kok EJ. "Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples". *GigaScience* 2017; 6(10): 1-8.

Košir AB, **Arulandhu AJ**, Voorhuijzen MM, Xiao H, Hagelaar R, Staats M, Costessi A, Žel J, Kok EJ, van Dijk JP. "ALF: a strategy for identification of authorised GMOs in complex mixtures by a GW-NGS method and dedicated bioinformatics analysis". *Scientific Report* 2017; 7(1): 14155.

Arulandhu AJ, van Dijk JP, Dobnik D, Holst-Jensen A, Shi J, Zel J, Kok EJ. "DNA enrichment approaches to identify unauthorised genetically modified organisms (GMOs)". *Analytical and Bioanalytical Chemistry* 2016; 408(17): 4575-4593.

Staats M, **Arulandhu AJ**, Gravendeel B, Holst-Jensen A, Scholtens I, Peelen T, Prins TW, Kok E. "Advances in DNA metabarcoding for food and wildlife forensic species identification". *Analytical and Bioanalytical Chemistry* 2016; 408(17): 4615-4630.

Holst-Jensen A, Spilsberg B, **Arulandhu AJ**, Kok E, Shi J, Zel J. "Application of whole genome shotgun sequencing for detection and characterization of genetically modified organisms and derived products". *Analytical and Bioanalytical Chemistry* 2016; 408(17): 4595-45614.

Overview of completed training activities

Discipline specific activities

- Meeting EU Partners - DECATHLON (2014), Naturalis, The Netherlands
- Meetings endangered species and UGMO detection - DECATHLON (2014-2016), RIKILT Wageningen University, The Netherlands
- Oral presentation - International research consortium (2015), iBET, Portugal
- Oral presentation - 6th International barcode of life conference (2015), University of Guelph, Canada.
- Poster presentation – 6th RAFA symposium (2015), UCT Prague, Czech Republic
- Poster presentation – Food fraud postgrad symposium (2016), Wageningen University and Research, The Netherlands
- Technical Secondment on Data analysis (2016), SJTU, China
- Oral presentation – 11th Rapid Methods Europe conference (2016), The Netherlands.
- Oral presentation - DECATHLON final meeting (2016), RIKILT Wageningen University, The Netherlands.
- Oral presentation - Wageningen PhD symposium (2017), Wageningen University and Research, The Netherlands.
- Oral presentation - 7th International barcode of life conference (2017), University of Johannesburg, South Africa.

General courses

- VLAG-PhD Week (2014), VLAG graduate school, The Netherlands.
- Information literacy and Endnote (2014), Wageningen University library, The Netherlands
- Coaching writing skills (2014-2016), Wageningen University languages service, The Netherlands
- Techniques for writing and presentation a scientific paper (2015), Wageningen Graduate School, The Netherlands

- Presenting with impact (2016), Wageningen Graduate School, The Netherlands
- Scientific Writing (2016), Wageningen Graduate School, The Netherlands
- Career assessment (2016), Wageningen Graduate School, The Netherlands
- Career perspective (2017), Wageningen Graduate School, The Netherlands

Optional courses and activities

- VLAG research proposal (2014), Wageningen Graduate School, The Netherlands
- Project meetings with stakeholders (2014-2016), RIKILT Wageningen University, The Netherlands
- Food fraud postgraduate symposium organising committee (2016), Wageningen University and Research, The Netherlands
- Expertise group/business unit meeting (2014-2017), RIKILT Wageningen University, The Netherlands

Teaching obligations

- Mentoring undergraduate student (2016), RIKILT Wageningen University, The Netherlands
- Assisting Food packing course (2017), FQD Wageningen University, The Netherlands

Acknowledgements

Unity is strength, this thesis is the outcome of a project with collaborations all over the globe, a project where even someone's small contribution is appreciated. I express my gratitude to Esther Kok, Jeroen van Dijk and Martijn Staats for guiding me in the voyage of finishing my PhD in a rough sea. You all inspire me in some way that impacted me and led to my scientific and personal development. I thank Prof. Dr. Saskia van Ruth for her effort in helping me in the process of finalising this thesis and supporting me when it was needed. My gratitude continues by acknowledging the help from the RIKILT and FQD staff, specifically Marleen Voorhuijzen, Rico Hagelaar, Bonnie Molenaar, Stephanie Zaaijer, Ingrid Scholtens, Theo Prins, Richard van Hoof, Jorg Nijland, Irene Konig-Lamers, Arno Bak, Prof. Dr. Vincenzo Fogliano, Lysanne Hoksbergen and Kimberley Boss.

"No friendship is an accident" a special gratitude to Valentina Acierno, Isabelle Silvis, Viola Ghio, Ningjing Liu, Yuzheng Yang and Jing Yan, you all are more like a family to me. "Behind every successful man stands a woman", my wife Thea van den Berg is that person, she supported and motivated me in each and every step I took to come to this point in my life. I would also like to thank all my family and friends, especially Fop and Jeltje van Driel and the members of student chaplaincy, for being supportive and encouraging me to finish my PhD. A special thanks for Dikkie and Lani for teaching me the art of relaxing. I am very thankful to my father Arulandu Sebastian and my mother the late Anthony Rajathi for all their care, prayers and encouragement in pursuing my goals. Last, but not least, I thank my heavenly Father for guiding me in each and every moment of my life.

Acknowledgements of financial support

This work was supported by the DECATHLON project (613908), which was funded by the European Commission under Seventh Framework Programme (FP7)

Propositions

1. The evolution of food authenticity methods will drive the evolution of food fraud.

   (this thesis)
2. Identification of species and GM variants is unlimited when affordable NGS methods are available.

   (this thesis)
3. Viruses are harmful, but the use of viruses for gene therapy will reverse this perception.
4. As evolution drives diversity, conservation efforts to maintain diversity is hindering evolution.
5. A species is more than its ATGC content.
6. In the scientific community knowledge should be valued over money.

Propositions belonging to the thesis entitled:

"Next generation DNA sequencing based strategies; towards a new era for the traceability of endangered species and genetically modified organisms"

Alfred Joseph Arulandhu

Wageningen, 19th December 2018.