DNA metabarcoding for the identification of species within vegetarian food samples

Megan Dawn De Jager

A minithesis submitted in partial fulfilment of the requirements for the degree of Magister Scientiae in the Department of Biotechnology, University of the Western Cape.

Supervisor: Prof. Eugenia D’Amato and Co-Supervisor: Carlotta Pietroni (PHD

November 2021

Keywords

Food Authentication

DNA Metabarcoding

DNA Extraction

Polymerase Chain Reaction

Multiplex reaction

Next Generation Sequencing

High-throughput Sequencing

Library building

Illumina sequencing by synthesis

Bioinformatics

Species identification

Abstract

DNA metabarcoding for the identification of species within vegetarian food samples

Megan D. De Jager

MSc, Thesis, Department of Biotechnology, University of the Western Cape

Aims

DNA metabarcoding has recently emerged as a valuable supplementary tool to ensure food authenticity within the global food market. However, it is widely known that highly processed food samples are one of DNA metabarcoding's greatest shortfalls due to high DNA degradation, presence of PCR inhibitors and the incomplete removal of several undesirable compounds (such as polysaccharides) that makes the amplification of desired DNA challenging.

This project has two main aims, the first of which was to determine and develop a cost and time effective DNA metabarcoding system that could successfully describe to species level the ingredient composition of highly processed vegetarian food products. The DNA metabarcoding system was thoroughly evaluated and tested by combining well-researched primers with varying concentrations into a multiplex reaction. The combination of plant and animal primers selected that yielded the best results were used to determine the species composition in the samples.

The second aim is to determine the possible presence of meat contaminants within the highly processed vegetarian food samples. Numerous studies have shown that food adulteration is a wide-spread phenomenon throughout the world due to the economic gains it can provide. Animal

primers were introduced into the multiplex reaction to aid in the identification of any meat products that could have been inserted into the vegetarian products to lower the overall cost to company.

Methodology

Thirty-two highly processed vegetarian food samples were collected in the Cape Town area from local and franchised supermarkets. DNA was extracted using the Chloroform/Isoamyl alcohol method best suited for plant-based samples followed by amplification of the following mini-barcoding regions: the mitochondrial 16S ribosomal rRNA, cytochrome B, tRNALeu – trnL – UAA intron and the ribosomal internal transcribed spacer region – ITS2 for plant and fungi identification. The PCR products were purified using the Qiaquick kit and library preparation and building was conducted using the TruSeq DNA PCR-free Library kit. Final purification was completed using AMPure XP kit and the pooled libraries were sequenced on an Illumina Miseq using 300bp paired-end run. Statistical and bioinformatic analysis on the NGS raw sequence reads was performed in R version 3.6.3.

Results

The results of the data analysis showed that the cytochrome B primer couldn't detect any animal DNA in the vegetarian samples, however animal-derived sequences were detected in the positives present, validating the efficacy of the multiplex reaction. Mitochondrial 16S ribosomal rRNA was only able to detect plant-based DNA due to the structural homology between chloroplast and mitochondrial DNA. The fungal ribosomal internal transcribed spacer region – ITS2 detected sequences deriving from "Viridiplantae". This result could have been due to the fungal and plant ribosomal internal transcribed spacer region – ITS2 sharing a reverse primer during amplification. The trnL region was able to detect the presence of undeclared coriander, mustard and wheat in 8 (29%), 6 (21%) and 5 (18%) samples respectively. Additionally, trnL was able to detect the presence of tobacco in 11 (35%) samples. This could have been due to cross-contamination

between samples being co-extracted and amplified at the same time for separate studies. The PITS2 region was able to detect the presence of undeclared barley, mustard and wheat in 8 (25%), 4 (14%) and 4 (14%) samples respectively.

Our results show the possibility of DNA metabarcoding for the authentication of a wide range of species present in highly processed vegetarian samples using a single assay. However, further optimization of the technique for the identification of both plant and animal species within vegetarian samples needs to be performed before the wide-spread implementation of this technology would be both feasible and viable. Eliminating primer biases, decreasing the risk of homology between different primers in the same assay as well as preventing the amplification of sequencing of undesirable DNA need to be further explored and ultimately mitigated before DNA metabarcoding can be widely seen as an effective and cost-effective method for authentication and food control.

Date: 05/11/2021

Declaration

I, Megan Dawn De Jager, hereby declare that *DNA metabarcoding for the identification of species within vegetarian food samples* is my own work, that it has not been submitted before for any degree or examination in any other university, and that all the sources I have used or quoted have been indicated and acknowledged as complete references.

Megan Dawn De Jager

November 2021, Cape Town

Acknowledgements

First and foremost, I would like to extend my sincerest thanks to Professor Eugenia D'Amato for your time, effort and guidance throughout this study opportunity. Your words of encouragement to take on this Master's journey is what had gotten me here and what has ultimately got me over the finish line. Thank you for motivating for the bursaries I have received and the belief that you had in me to achieve all that I have. To my amazing Co-supervisor, Carlotta Pietroni – words can't express how grateful I am for your guiding hand, patience and the willingness to share all the knowledge that you had garnered from your own studies. Without your constant encouragement and perseverance, I don't think I would have gotten through my initial years of my thesis. You were a shoulder to cry on and an ally to rely on and for that I will be eternally grateful.

A special thank you to Prof. Julian May from the Centre of Excellence in Food Security for motivating and successfully acquiring my funding in order to make this research possible. Without it, this thesis title would never have made it off the ground. On behalf of every student who has ever received funding from you, thank you for making our dreams a reality.

To my colleagues at the Forensic DNA lab, thank you for keeping me sane during my laboratory experiments. Your light-hearted warmth and welcoming aura always made me feel better after a rough day in the lab. Thank you for sharing your wealth of knowledge and making my Master's journey as special as it was. The process would have not been the same if you all weren't there to walk with me through it.

To my loving and supportive parents – Julie and Hendrik – and my brothers and sisters – Chadae, Cindy, Hennie and Luan – thank you for your continuous support and faith in my ability to tackle this monumental task of a thesis. Thank you for believing in me when my faith in myself wavered. It has been a tough time, but this final thesis is a testament of human will to never give up and to never give in to the fear of failure and disappointment.

To my Bubs – Michael – Thank you for your unwavering support and constant belief that I would concur and overcome all the hardships and hiccups this thesis had to throw at me. Thank you for the times when I needed to hear how far I had come and how proud you were of me for getting to where I am. The cuddles and cry sessions helped a lot too. May we continue to grow strength by strength in everything we face and overcome.

“Strength doesn’t come from what you can do. It comes from overcoming the things you once thought you couldn’t *” – Rikki Rogers*

List of Abbreviations

EMA	-	Economically Motivated Adulteration
GMA	-	Grocery Manufacturers Association
FSA	-	Food Standards Agency
DNA	-	Deoxyribonucleic Acid
PCR	-	Polymerase Chain Reaction
GMO	-	Genetically Modified Organisms
qPCR	-	Quantitative Polymerase Chain Reaction
ELISA	-	Enzyme-linked Immunosorbent Assay
IgG	-	Immunoglobulin G
NGS	-	Next Generation Sequencing
rRNA	-	Ribosomal Ribonucleic Acid
nDNA	-	Nuclear Deoxyribonucleic Acid
mtDNA	-	Mitochondrial Deoxyribonucleic Acid
cpDNA	-	Chloroplast Deoxyribonucleic Acid
LSC	-	Large Single Copy
SSC	-	Small Single Copy
IR	-	Inverted Repeat
ITS2	-	Internal Transcribed Spacer 2 Region
rDNA	-	Recombinant Deoxyribonucleic Acid
ASV	-	Amplicon Sequencing Variant
OTU	-	Operational Taxonomic Unit
DADA2	-	Divisive Amplicon Denoising Algorithm
mg	-	Milligrams
ml	-	Millilitre
CTAB	-	Cetyltrimethylammonium Bromide
˚C	-	Degrees Celcius
µl	-	Microlitre
rpm	-	Revolutions per minute

TE	-	Tris-Ethylene-diamine-tetra-acetic acid
ng/µL	-	Nanograms per microlitre
10X	-	10 times
dNTP	-	Deoxyribonucleotide triphosphate
$MgCl_2$	-	Magnesium chloride
Taq	-	Thermus aquaticus
BSA	-	Bovine serum albumin
bp	-	Base pairs
mM	-	Millimolar
U/µl	-	Units per microlitre
min	-	Minute
EDTA	-	Ethylenediaminetetraacetic acid
TBE	-	Tris Borate EDTA
w/v	-	Weight to volume
Kb	-	Kilobase
EB	-	Elution Buffer
PB	-	Phosphate Buffer
Tris-HCl	-	Tris hydroxymethyl aminomethane hydrochloride
NaCl	-	Sodium Chloride
ERP2	-	End Repair Mix 2
ATL	-	A-Trailing Mix
LIG 2	-	Ligation Mix 2
STL	-	Stop Ligation Buffer
RNase	-	Ribonuclease
SNP	-	Single Nucleotide Polymorphism
FDL	-	Forensic DNA Lab
UWC	-	University of the Western Cape
EMA	-	Economically Motivated Adulteration
GMA	-	Grocery Manufacturers Association

FSA	-	Food Standards Agency
5’	-	5 Prime
3’	-	3 Prime
cpDNA	-	Chloroplast DNA
nDNA	-	Nuclear DNA
mtDNA	-	Mitochondrial DNA
ITS2	-	Internal Transcribed Spacer 2
RNA	-	Ribonucleic acid
CytB	-	Cytochrome B
SBS	-	Sequencing by Synthesis
GC	-	Guanine – Cytosine
AT	-	Adenine – Thymine
qPCR	-	Quantitative PCR
CNV	-	Copy Number Variants
SV	-	Structural Variants
ASV	-	Amplicon Sequence Variant
OTU	-	Operational Taxonomic Unit

List of Figures

List of Tables

Supplementary Table List

Chapter 1: Literature Review

1.1. Vegetarianism: health benefits and ethical considerations

Vegetarianism is the common practice of abstaining from consuming animal or animal-related products. These include diets that mainly consist of fruits, vegetables, nuts and legumes. There are many different variations of vegetarianism (**Figure 1.1**) and most individuals choose the diet that aligns well with their lifestyle, ethical standards and beliefs. Many people across the world have adopted this healthier way of living for countless different reasons, but scientific research has shown that one of the main advantages of a purely vegetarian diet is the numerous health benefits that are associated with it (Ion, 2007). Studies have reported that diets consisting largely of fruits and vegetables dramatically decreases the levels of cholesterol, saturated fatty acids and animal protein that is associated with health issues such as obesity, diabetes, high blood pressure and cardiovascular diseases (McEvoy, Temple and Woodside, 2012). These particular diseases have been positively linked to the over-consumption of fresh and processed red meat, which have led many to exclude meat from their diet to avoid such negative implications. The reason behind the inhibitory effect of a plant-based diet to most lifestyle-related diseases is largely due to the myriad of dietary fibre, antioxidants and phytochemicals that are present within the fruits, vegetables, nuts and legumes (McEvoy, Temple and Woodside, 2012). The fibre found in most plants aids in actively reducing the levels of cholesterol found in the blood, while their naturally low saturated fatty acid content helps to decrease the blood viscosity, which in turn lowers blood pressure (McEvoy, Temple and Woodside, 2012). The lower sugar content, as well as the presence of complex carbohydrates, contributes to the effectiveness of insulin within the body, which naturally reduces the prevalence of type 2 diabetes mellitus (McEvoy, Temple and Woodside, 2012). A study conducted in 2012 illustrated that there was an overall 37% risk reduction in cases of coronary heart disease-related deaths in adults that consumed nuts 4 times weekly, with an 8.3% reduction for each weekly serving of nuts. Further studies have shown that the consumption of lignans and soy proteins that contain phyto-oestrogens may have a protective role against breast cancer development in women (McEvoy, Temple and Woodside, 2012). These research articles have found a definitive link between diet and disease and have concluded that converting to a greener and cleaner lifestyle may be the key to lengthening life expectancy rates.

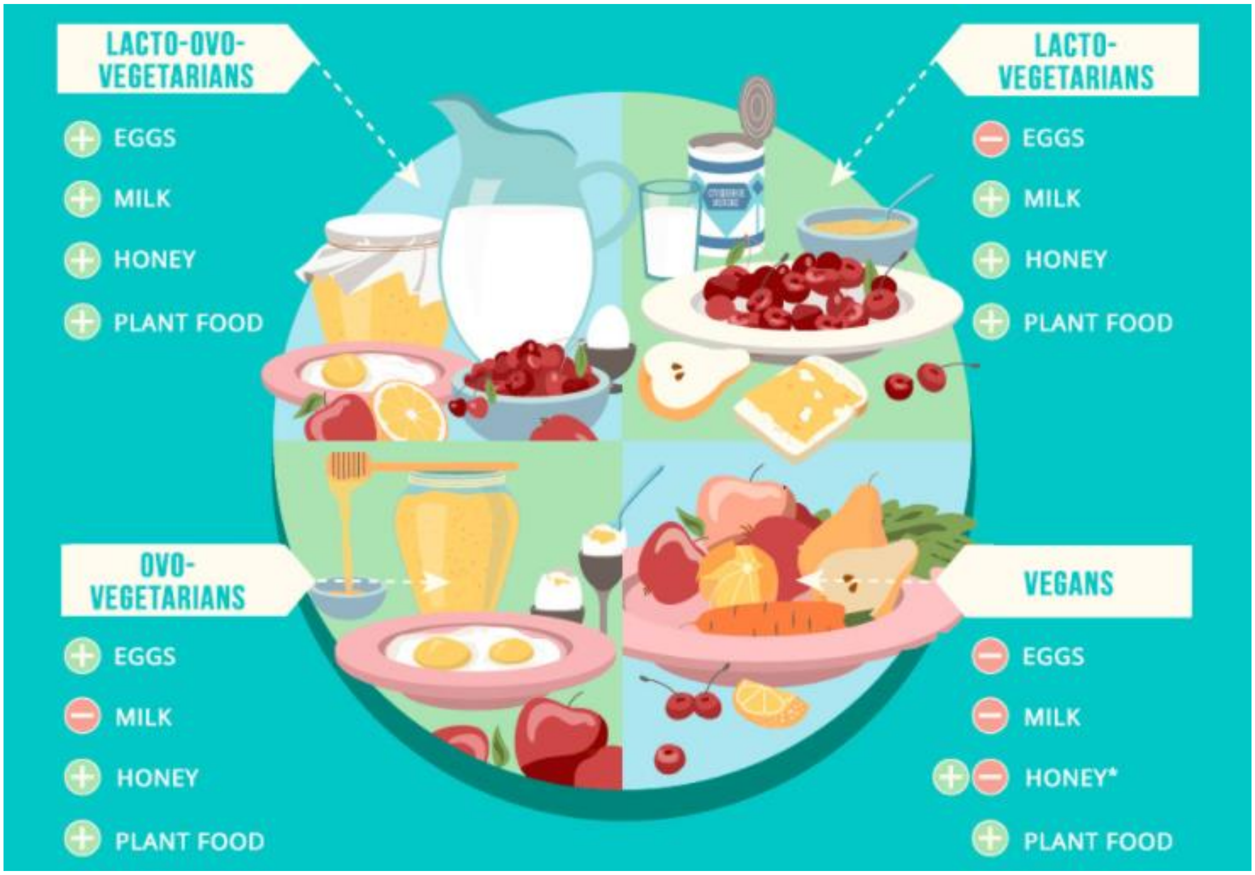

Figure 1. 1: Four different vegetarian diets and their requirements for conformity (Johnston, 2018). The choice of which vegetarian diet to follow purely depends on the ethical, socio-economic or religious reasons of the individual.

The positive health impact is not the only reason communities have decided to change to healthier, plant-based diets. Many do not condone the treatment of animals that are marked for human consumption, mainly due to the maltreatment and conditions given to these animals before death (Ion, 2007). Often animals are forced into small, overpopulated enclosures, handled cruelly and not given a suitable standard of living before slaughter. This has convinced many to remove meat from their diet, as they do not want to be a part of the cruelty and suffering of these animals based on their moral values (Ion, 2007). Other individuals prefer to abstain from certain meats for religious reasons, as their culture does not allow for the consumption of animals that may be found sacred or form an integral part of their religion.

Regardless of the reasons behind vegetarianism, many individuals have chosen to live this particular lifestyle, and there shouldn't be any infringement on their right to do so. However, this has not always been the case. There have been many reports and articles that have shown incidences of food fraud across the globe.

1.2. Food fraud

Food fraud or Economically Motivated Adulteration (EMA) is a collective term used to describe the intentional and deliberate misrepresentation of food or food ingredients; or misleading statements made about the product for economic gain (Di Pinto *et al.*, 2015). **Figure 1.2** illustrates the various types of food fraud and relevant description of each. The authentication of food contents has been one of the main quality-related issues that have vexed the food industry throughout history, as it has become increasingly difficult to identify when the crime has been committed, especially with heterogeneous products (Di Pinto *et al.*, 2015).

1.2.1. Economical, Environmental and health impact

The extent of food adulteration across the world is not fully understood, as many manufacturers do not intentionally create products that could pose as safety risk to the consumer, in an attempt to avoid detection. However, the Grocery Manufacturers Association (GMA) has estimated that fraud costs the global food industry between $10 billion to $15 billion annually (Johnson, 2014). While food fraud that results in food safety or a public health event may cost the businesses involved between 2% and 15% of their annual revenues and may even lead to possible bankruptcies and liquidation (Johnson, 2014).

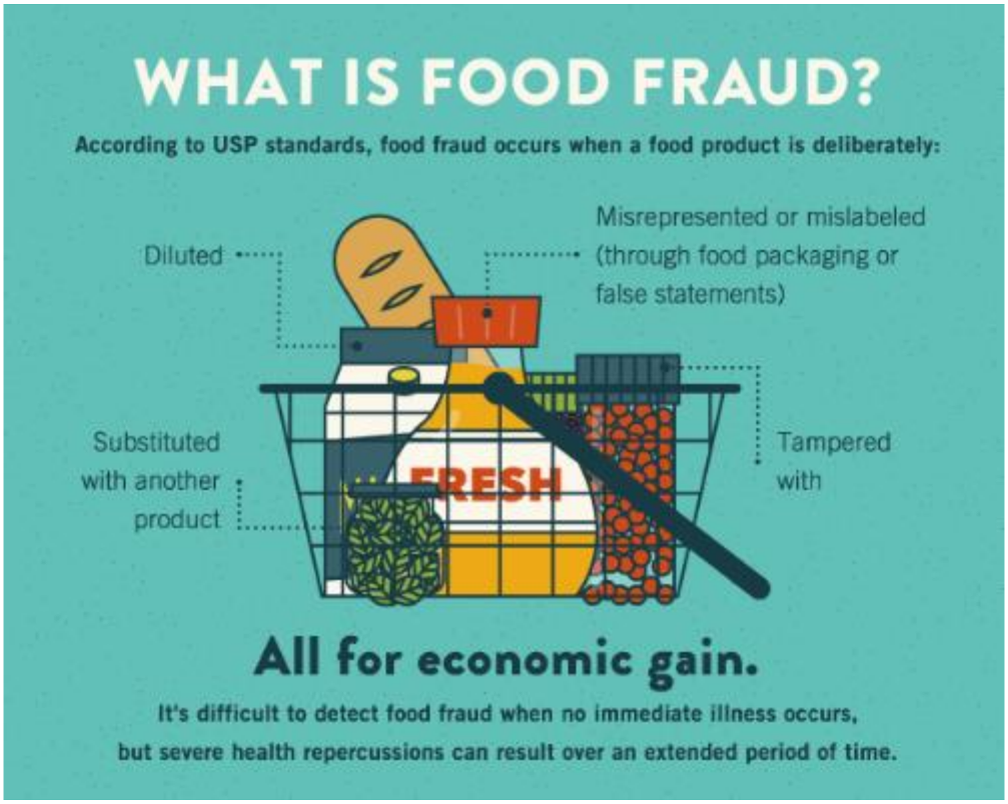
WHAT IS FOOD FRAUD?
According to USP standards, food fraud occurs when a food product is deliberately:
Diluted
Misrepresented or mislabeled (through food packaging or false statements)
Substituted with another product
Tampered with
FRESH
All for economic gain.
It's difficult to detect food fraud when no immediate illness occurs, but severe health repercussions can result over an extended period of time.

Figure 1. 2: The various types of food fraud (Lau, 2021)**.** Food can be manipulated in several ways to lower the costs involved in their production, which may include diluting, substituting or mislabelling.

One of the major concerns involving food fraud is the insertion of endangered species into food products as they provide a similar or a comparatively cheaper alternative. It is not only a threat towards the prevention of commercial fraud but also contributes in the further decline of red-listed or endangered species, which may ultimately lead to species extinction (Di Pinto *et al.*, 2015). Food fraud is also considered a significant food safety hazard as any undeclared ingredients may be harmful to human health, such as ingredients that may incite an allergic reaction that could lead to anaphylaxis or even death severity (Di Pinto *et al.*, 2015). Possible reasons behind this criminal act could connect to potentially increasing the value, reducing the cost or diluting the product to increase the profits generated from the product (Spink and Moyer, 2011).

1.2.2. Case of global vegetarian food fraud

A recent 2018 investigation conducted by the Food Standards Agency (FSA) in the United Kingdom had shown that two out of 10 of the Tesco’s and Sainsbury’s ready-made vegetarian meals contained traces of meat (Simpson, 2018). The Telegraph reported that a German government-accredited food safety organization had uncovered DNA traces of pork in the Sainsbury branded Meat-Free Meatballs and turkey in Tesco's Wicked Kitchen BBQ Butternut Macaroni and Cheese (Simpson, 2018). These allegations had caused an enormous public outcry from members of the community who had placed their trust in the company, as many had chosen these meals based on their social, ethical or religious reasons. A thorough investigation of the company has yet to be conducted, but this type of adulteration could be attributed to poor regulation of manufacturing protocols (Simpson, 2018).

1.3. Food authentication

Government legislation enforces the proper labelling of food and beverages which aims to reassure the consumer by providing them with all the information required to make an informed decision about the product (Georgiou, 2017). To enforce such legislation, the state-appointed legislation bodies recruit various scientific methods to certify that the food products in the market comply with their labelling (Georgiou, 2017). Food authentication is the analytical process of verifying food products compliance with the description presented on the label (Danezis *et al*., 2016). This information may include, but is not limited to, species origin (genetic, species or geography), production method (free-range, organic or traditional methods) and method of processing (freezing, microwave heating or irradiation) (Danezis *et al*., 2016). Consumers and government leaders are increasingly interested in discovering the geographical origin and quality of the products due to the mass globalization of food markets that have been said to affect the availability and variability of certain items. Therefore, authenticity testing has become a key criterion for food product legislation, particularly for the protection of regional foods. A few examples of food authentication techniques are discussed below.

1.3.1. Molecular analysis – DNA methods and Polymerase Chain Reaction

Molecular analysis in food authentication is considered the leader in all major authentication studies and is rapidly replacing previously utilized techniques. Nucleotide and protein-based methods are primarily used for species identification and detection (Dopheide *et al*., 2019). DNA methods are often preferred over protein-based techniques purely for the stability of double-stranded DNA molecules in experimental analysis. Most of the DNA-based methods often depend heavily on the amplification – creation of copies – of specific DNA regions by PCR (Dopheide *et al*., 2019). PCR refers to the *in vitro* synthesis and amplification of short DNA fragments that can be up to several kb in length. It was first invented by Karry Mullis, who utilized a thermo-sensitive polymerase that was later replaced by the thermo-stable enzyme known as *Taq* polymerase (Mullis *et al*., 1986). The polymerase was isolated from a bacterium that live in 75˚C hot springs found in the Yellowstone National Park (Dopheide *et al*., 2019). PCR achieves the purpose of creating copies in three major steps, specifically named denaturation, annealing and extension or elongation (**Figure 1.3** for visualization). The first step involves the separation of two DNA strands using heat-based techniques in order to create template strands for replication (Kadri, 2020). The annealing stage joins the primers to the DNA template strands when the mixture is cooled, allowing the polymerase to bind and extend by copying the template strand. Lastly, during the elongation step, the temperature is raised to the optimal functioning temperature of the specific polymerase, which allows the enzyme to extend the complementary strand in the 3' – 5' direction (Kadri, 2020). This process allows for the exponential increase in DNA copy number before reaching an eventual plateau, which often depends on the concentration of DNA template and overall PCR efficiency. However, limitations surrounding the use of this technology often lie with the demand for prior knowledge of primer design as well as the limited length of the amplicons in case of degraded DNA samples (which should not exceed 100 - 150 bp in length).

PCR for DNA amplification only really became accessible to scientists in 1989 due to the publication written by Kocher *et al* that described the use of universal primers for animal mitochondrial DNA as well as the emergence of the first PCR machines (Kadri, 2020). This technology had revolutionized the scientific field as it offered easier analysis of genetic polymorphisms and limited the constraints that accompanied preserving tissues in liquid nitrogen for fresh tissue samples (Kadri, 2020). Before the advent of Next Generation Sequencing or high-throughput sequencing, the majority of papers evaluating DNA polymorphisms in conservation genetics or molecular ecology involved PCR-based methods (Kadri, 2020). Techniques involving PCR amplification are generally revered for its simplicity, sensibility, specificity and quick execution (Mafra *et al*., 2008). PCR techniques tailored for food analysis, especially ones involving GMO's and allergens, presented a more quantitative approach over the traditional qualitative analysis that is usually associated with traditional PCR techniques (Mafra *et al*., 2008).

One quantitative PCR-based method known as real-time PCR (qPCR) utilizes specific probes or labelled primers that allow for the simultaneous detection and confirmation of fragments, which increases the reliability of this application for food analysis (Mafra *et al*., 2008). This approach can detect multiple species from a mixture and has the ability to quantify the PCR products formed during the amplification process (Kang, 2019). Despite the urgent need for applications such as real-time PCR for food analysis, the basic principles regarding the method development and validation of the technique have yet to be fully evaluated (Kang, 2019).

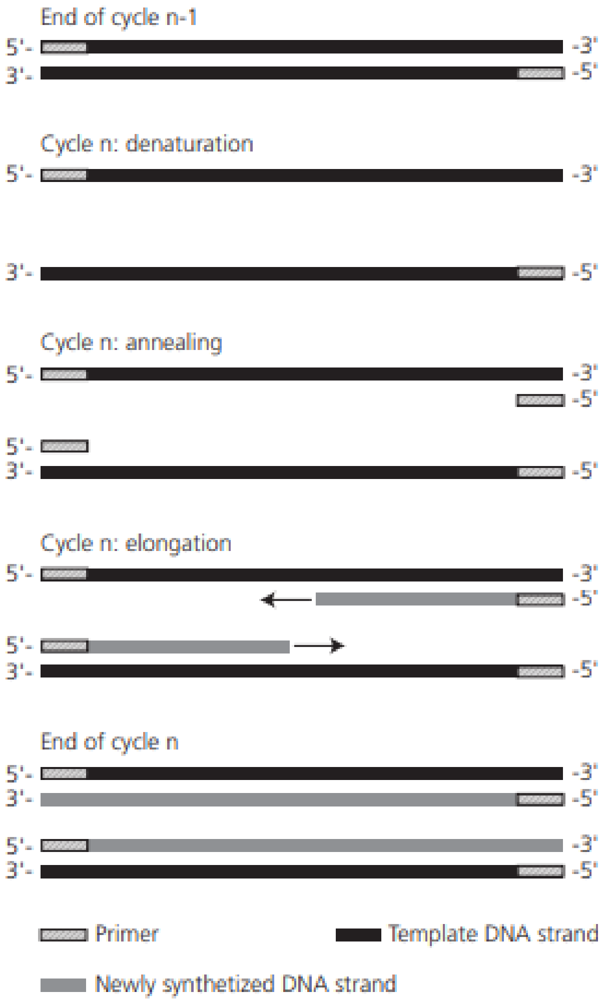

Figure 1. 3: An image representing the three main stages of PCR. These are broken down into denaturation, annealing and elongation in the 3' – 5' direction (Mafra *et al*., 2008).

1.3.2. Protein-based techniques

Allergies related to food have become an increasing health concern worldwide. Manifestations of food allergies can vary from minor digestive issues and slight skin irritations to severe symptoms that can even be life-threatening (Prado *et al*., 2016). Therefore, governments have revised various food legislation laws that obligate manufacturers to indicate the presence of certain allergenic ingredients on their food labels. For many years, the presence of potentially allergenic ingredients has been detected using DNA or protein-based approaches (Prado *et al*., 2016). The choice of the appropriate method depends on the type and stability of the allergy-inducing ingredients detected.

Protein-based approaches rely on the stability of the target protein present in the food, or another protein that indicates the presence of the offending food, for appropriate detection (Prado *et al*., 2016). Although protein-based techniques can be influenced to target more than one protein for allergen detection, the method presents unique challenges. The protein content can frequently be affected by food processing protocols as well as biological variations that can be influenced by seasonal and geographical impacts (Prado *et al*., 2016). Additionally, some thermal processes have the potential to reduce the solubility of the target proteins which can compromise the results of analysis (Prado *et al*., 2016). In contrast, some food allergenic ingredient present at high protein content and low DNA content such as eggs, while in other cases, the opposite is true. The choice of DNA or protein-based approaches should depend on the specific requirements of the ingredients to be analyzed (Prado *et al*., 2016).

Additionally, food allergen detection using protein-based techniques have traditionally relied on immunoassays, which employ an antibody-based detection approach that targets proteins associated with the allergenic food (Prado *et al*., 2016). Enzyme-linked Immunosorbent Assay (ELISA) uses IgG antibodies obtained from immunized animals for the detection of allergens in food products in the food industry (Prado *et al*., 2016). ELISA’s offer great sensitivity and is relatively easy to execute. Many commercial ELISA kits are currently available for the detection of numerous allergens such as eggs, peanuts, soybeans and many others. Lateral flow assays and dipstick tests are an easy-to-use, cost-effective and fast-acting variant of traditional ELISA kits that are performed on a membrane strip (Prado *et al*., 2016). However, there are several major drawbacks that need to be considered before moving forward with this approach: i) complex food matrix interference can affect the colorimetric absorption measurement used to detect the allergenic proteins, ii) possible cross-reactivity of certain proteins and antibodies can lead to false positives which increase when polyclonal antibodies are used, iii) food processes used to prepare food for consumption can change the structural integrity of the protein conformation which can influence epitope recognition (Prado *et al*., 2016). Despite all the drawbacks, ELISA remains the method of choice for the detection and quantification of food allergens due to its cost-effectiveness and ease of execution (Prado *et al*., 2016).

1.3.3. Chromatography techniques

The chromatographic analysis allows for the reliable and rapid separation of chemically similar compounds in heterogenous food products, and can often be used in food authentication and food allergen detection studies (Coskun, 2016). For food authentication, this technique must overcome numerous challenges specific to heterogenous food products. The chemically diverse nature of food that possesses a wide range of polarities allows chromatographic methods to generate a unique chemical fingerprint that differentiates and authenticates the molecules (Coskun, 2016). This method relies on the identification of minimal analytic differences between the patterns associated with the compounds or the unique markers presented. Due to the enormous chemical complexity of most food products, high-resolution chromatographic techniques such as liquid and gas chromatography have been proposed as possible alternatives for this challenging medium (Coskun, 2016). Liquid chromatography primarily focuses and targets three key characteristics of chemical compounds, namely molecular size, electric charge and polarity. This technique is typically used to detect vitamins, amino acids, proteins and carbohydrates while Gas chromatography analyzes volatile and semi-volatile molecules (Coskun, 2016). In summary, chromatography methods analyze the patterns of a specific food product profile and link it to a characteristic target value associated with the food identity origin. This method is the most valuable in the identification and authentication of high-quality products with cost-effective or sub-standard ingredients such as honey, wines and olive oils.

A technique known as DNA metabarcoding has shown some potential in providing a suitable method in the identification, differentiation and assignment of multiple species within heterogeneous food samples (Staats *et al.*, 2016). This method has been used in previous studies as a molecular tool for analyzing alleged wildlife crime cases and the detection of adulteration in the food industry (Staats *et al.*, 2016). This strategy has previously been tested for the detection of species within highly processed food materials containing degraded DNA, which could aid in the identification of endangered and hazardous species in food items (Staats *et al.*, 2016)

1.4. DNA metabarcoding

The capability of this approach is dependent on the combination of two powerful techniques known as DNA barcoding and high-throughput sequencing. DNA barcoding is an established diagnostic tool that involves the PCR amplification and sequencing of distinct, standardized regions of DNA followed by a comparison of these sequences to a reference database (Hebert *et al*., 2003). Next Generation Sequencing is a high throughput sequencing method used to reveal species composition in heterogeneous or environmental samples (Dormontt *et al.*, 2018). With regards to DNA metabarcoding, the primary aim of the PCR is to generate a large number of copies of a DNA template that will be sequenced using high-throughput sequencing technologies (Hebert *et al*., 2003).

1.4.1. Criteria for a good barcoding system

According to Taberlet *et al*., the best DNA barcoding systems conform to a set of key criteria. Firstly, the barcoding system must be standardized so that the same DNA region can be used to classify species based on their taxonomy and should be variable enough to differentiate between all species (Taberlet *et al.*, 2012). The DNA regions of interest must contain significant phylogenetic data to easily assign and differentiate between taxonomic classes such as family or genus and must be especially robust with reliable DNA amplification, sequencing and conserved priming sites (Taberlet *et al.*, 2012). This is especially important when heterogeneous sample sets are involved due to the mixture of DNA and the difficulty in identifying the species by morphological means. Additionally, the standardized region needs to be short enough for the amplification of highly degraded samples (Taberlet *et al.*, 2012).

1.4.2. DNA Barcode Regions

1.4.2.1. Mitochondrial DNA

The mitochondria are described as short, circular organelles that are present in nearly all eukaryotic cells, and are the only cytoplasmic organelles that are known to carry genetic elements *(*Schon *et al*., 2012). An overview of the mitochondrial genome is featured in **Figure 1.4** below. They are primarily maternal transmitted, non-recombining and considered to have an elevated mutation rate than that of nuclear DNA (nDNA), which aids in studies that focus on the identification of different species (Schon *et al*., 2012).

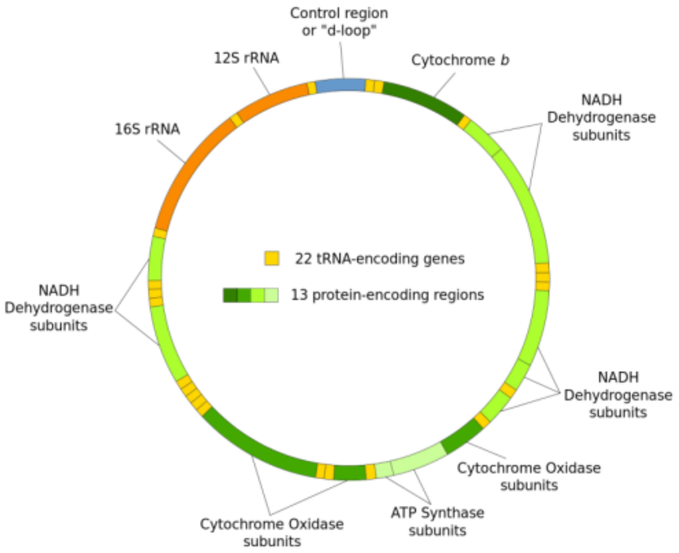

Figure 1. 4: Structure of the mitochondrial genome. It is approximately 16 595 bp long. The sections highlighted in various shades of green depict the plethora of protein-encoding regions. The mitochondrial control region (CR) is the longest non-coding region in mtDNA and is considered the most variable region in the mitochondrial region. It is considered the ideal sequence for genetic analysis due to its exceptionally fast evolutionary rate. Image taken from: (http://www.contexo.info/dna_basics/chromosomes/mitochondria/)

Initial studies conducted on animal mitochondrial DNA (mtDNA) have shown that it possesses key characteristics that make it an ideal genetic marker (Schon *et al*., 2012). Firstly, multiple copies of mtDNA are present within the cell, which makes the amplification of mitochondrial DNA easier than most parts of the nuclear DNA (nDNA). Additionally, mitochondrial DNA contains regions that alternate between variable and conserved sections on the same molecule which enables the design of universal primers (Schon *et al*., 2012). These universal primers have the ability to amplify pieces of mtDNA from any species, without the need for prior knowledge of the unknown species. These primers target highly conserved regions that are variable among species. This variability ensures the effective resolution and discrimination of the species through the use of a single barcode. These particular traits of mtDNA, along with the relatively inexpensive applications revolving around it, act as an important tool for population genetics (Schon *et al*., 2012).

1.4.2.1.1. 16S and 12S Mitochondrial rRNA

The mitochondrial genome is a large organelle that encodes for two ribosomal ribonucleic acid (rRNA) subunits that are involved in the translation of messenger RNAs that synthesize mitochondrial proteins (Staats *et al*., 2016). These subunits are known as 12S and 16S rRNA. The location of each gene on the mitochondrial genome are portrayed in **Figure 1.4** as well as cytochrome B (CytB), which will be discussed later (Staats *et al*., 2016). These genes are the most widely used genetic markers for species identification of highly degraded or problematic samples such as bones, feathers and meat products (Staats *et al*., 2016).

Scientists have shown that the 16S and 12S rRNA regions found in the mitochondrial genome contain highly conserved internal regions across all taxa that are valuable for designing universal primers (Staats *et al*., 2016). These are alternated with short hypervariable regions that are highly species-specific and are different across all taxa that will allow for species identification (Staats *et al*., 2016). A 250 bp barcode marker developed by Sarri *et al* ensured the successful amplification of the 16S rRNA region across all sample types such as cheese, fish fillets, birds and highly processed meats (Staats *et al*., 2016). Additionally, Karlsson had been able to identify a total of 28 different mammals including game and domestic species using the 16S and 12S regions, further validating the use of this primer set for species identification (Staats *et al*., 2016).

1.4.2.1.2. Cytochrome B

This particular marker has made appearances in studies that relate to wildlife protection and conservation within an environment, as it has an extensive track record for the identification of animal species in meat products (D’Amato *et al*., 2013). The early discovery and availability of the CytB DNA marker had encouraged numerous studies regarding molecular evolution, which is responsible for a large number of sequences available today (D’Amato *et al*., 2013). The most widely studied fragment of CytB is around 358 bp in length and its high inter and intraspecies variation has made it an attractive target for most phylogenetic studies (D’Amato *et al*., 2013). Additionally, this gene is located on the mitochondrial genome, which makes it easier to isolate and amplify using PCR-based methods (D’Amato *et al*., 2013).

1.4.2.2. Chloroplast DNA

cpDNA are circular organelles that can range from 115 to 165 kilobase pairs in length that contain a LSC and a SSC region that are separated by two copies of a large IR region (Liu *et al*., 2018). In general, chloroplast genomes are seen to be more conserved than nuclear or mitochondrial genomes with regards to their organization, structure and gene content. Their nucleotide substitution rate of their genes are higher than mitochondrial and lower than nuclear (Liu *et al*., 2018). Due to its highly conserved gene content, small size and simple structure, it has been widely analyzed in genome evolution studies for a broader understanding in intron gene losses at higher taxonomic levels as well as genome size variations. Additionally, it is a useful tool to track demographic history, analyze species divergence and hybridization as well as species identification due to its non-recombinant nature and their generally uniparental inheritance (Liu *et al*., 2018).

cpDNA regions have mostly been screened using traditional methods due to its efficacy in describing related taxa for analysis, however recent studies conducted on whole-genome research has uncovered a more systematic approach that has the capacity to take into account the mutational rates of chloroplast genomes (Liu *et al*., 2018). Using this method, informative regions found in chloroplast specific hotspots can be identified for a specific tribe, family or genus (Liu *et al*., 2018). Due to recent developments in NGS technology, an increasingly better understanding of cpDNA sequences have been isolated and assembled, which has provided a more efficient and cost-effective way to obtain information regarding differential gene expression and phylogenomics studies (Liu *et al*., 2018).

1.4.2.2.1. Plant trnL

Scientists across the world have struggled to find a suitable DNA marker to distinguish the majority of plant species, as the chloroplast and mitochondrial genomes often evolve too slowly to indicate enough variation to identify species (Taberlet *et al*., 2007). Discovered around 15 years ago, the chloroplast *trn*L (UAA) intron region has been extensively analyzed and examined, providing researchers with a wealth of knowledge and database resources to draw on when designing their own experiments (Taberlet *et al*., 2007). **Figure 1.5** portrays the whole *trn*L (UAA) intron region, which is roughly between 254 to 767 base pairs in length, while the shorter P6 loop fragment of the intron is roughly between 10 to 143 base pairs long.

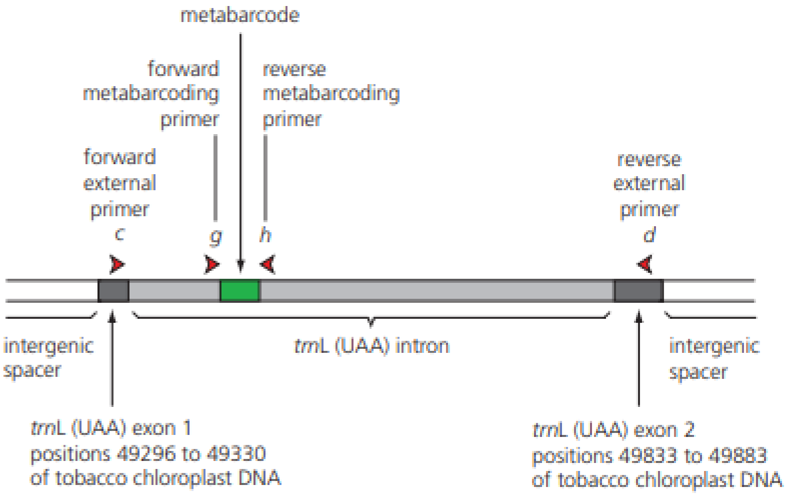
metabarcode
forward metabarcoding primer
reverse metabarcoding primer
forward external primer
c
g
h
reverse external primer
d
intergenic spacer
*trn*L (UAA) intron
intergenic spacer
*trn*L (UAA) exon 1 positions 49296 to 49330 of tobacco chloroplast DNA
*trn*L (UAA) exon 2 positions 49833 to 49883 of tobacco chloroplast DNA

Figure 1. 5: A visual representation of the complete *trnL* (UAA) intron region indicated in light grey. The highlighted green section depicts the P6 loop, which is primarily used to amplify highly degraded plant DNA (Taberlet *et al*., 2007).

The primer set targeting plant species has been widely used in applications regarding highly degraded DNA as the P6 loop has been known to be successfully amplified in those circumstances, making it an essential tool for the food forensic industry and in ancient DNA studies (Taberlet *et al*., 2007). It has a highly robust and conserved amplification system, making it an ideal candidate for amplifying multiple species simultaneously. However, the major drawback of the *trn*L intron region is its substantially low-resolution power (with only 67.3% of the species from Genebank being correctly identified) while the resolution is even lower for the P6 loop fragment with only 19.5% of species being unambiguously identified (Taberlet *et al*., 2007). Despite this disadvantage, it is still seen to be one of the better primer sets to use for plant species identification due to its vast reference database support and its superiority in the successful reconstruction of phylogenies between species when compared to other primer sets.

1.4.2.3. Nuclear DNA

Nuclear DNA is the DNA enclosed in each cell nucleus within a eukaryotic organism. Some scientists suggest that despite the mitochondrial DNA's durability within the cell, nuclear loci are far superior for DNA quantification because of its diploid copy number, which enhances the predictive success of the DNA profiling at species level (Ng *et al*., 2014). Targeting nuclear DNA not only promotes successful identification and quantification of analyzed samples with traces of template DNA, but this approach could also provide multiple target sites that can simultaneously and species-specifically quantify DNA templates from a variety of species (Ng *et al*., 2014).

1.4.2.3.1. Plant /Fungi ITS2

The internal transcribed spacer 2 region (ITS2) found in nuclear ribosomal RNA has been considered one of the most significant plant genetic markers in molecular evolution, as it shows a high degree of sequence variability at the species category (Han *et al*., 2013). Additionally, the key criteria of an ideal DNA barcode has mostly been fulfilled, which includes the ease at which the DNA can be amplified, the availability of conserved regions that allow for the development of universal primers due to its small size as well as the regions of variability that ensure that closely-related species can be identified (Han *et al*., 2013). These features are especially important when dealing with samples that may contain partially or highly degraded DNA, such as the cases involving food or ancient sample sets (Han *et al*., 2013). Compared to the whole ITS region, ITS2 sub-region is a more suitable candidate for species identification due to its short length and higher efficiency in PCR amplification. Numerous studies have been conducted, and this exposure has grown the availability of structural information that will permit research at a higher taxonomic level, which will ultimately lead to the improved robustness and accuracy of the construction of phylogenetic trees (Han *et al*., 2013).

1.4.3. High-throughput sequencing technology

NGS encompasses all technologies that involve deep, high-throughput, massively parallel DNA sequencing for large-scale applications (Kulski, 2016). It had first emerged a few decades after the discovery of the Sanger sequencing method that was developed in 1977. The NGS technologies differ from that of the Sanger sequencing method in that they provide massively parallel analysis that allows for incredibly high-throughput sequencing from multiple complex samples at a reduced cost (Kulski, 2016). Second generation sequencing methods involve the preparation and amplification of sequencing libraries before sequencing the amplified DNA clones, while some third generation sequencing undertakes single molecular sequencing without the need for the creation of time-consuming and costly libraries (Kulski, 2016). NGS has reduced the time needed to generate gigabase-sized sequences from many years to days or hours and is accompanied by a substantial price reduction. For example, J.C. Venter, a renowned scientist associated with the consortium dedicated to the Human Genome Project, took almost 15 years to sequence the entire human genome for $1 million using the Sanger sequencing method, while J.D Watson and his team sequenced a genome using NGS within 2 months and for 100th of the price of the Sanger sequencing run (Kulski, 2016).

1.4.3.1. Roche 454 pyrosequencing by synthesis (SBS)

The first commercially successful second-generation sequencing system was the Roche 454 pyro SBS that was developed by 454 Life Sciences in 2005 (Kulski, 2016). This application utilizes sequencing chemistry that detects and measures visible light that is produced by the repeated nucleotide incorporation into the newly synthesized DNA chain (Kulski, 2016). This system was miniaturized and was able to produce more than 200 000 reads at around 100 to 150 bp per read in 2005 and was improved upon to produce an average length of 700 bp in 2008. The major limitations regarding this technology are the high cost of reagents, high error rates in homopolymer repeats as well as the announcement that Roche will no longer supply or service these 454 pyrosequencing reagents or chemicals (Kulski, 2016).

1.4.3.2. Illumina sequencing by synthesis

Another sequencing technology known as Illumina was purchased and commercialized by Solexa Genome Analyzer in 2007 and is currently regarded as the most successful sequencing system with more than 70% dominance in the market (Kulski, 2016). The Illumina is different to that of the Roche 454 sequencer in that its sequence by synthesis technology utilizes removable fluorescently labelled chain-terminating nucleotides that can produce a large output of data with a reduced reagent cost (Kulski, 2016). The single-stranded template DNA is washed over a flow cell and is bound to the surface due to the nature of the complementarity between the short oligonucleotides present on the flow cell and the adapter sequences that are attached to the DNA fragments. Solid-phase bridge amplification follows a blend of unlabelled nucleotides and an appropriate enzyme is washed over the flow cell, which allows the incorporation of nucleotides that form double-stranded bridges (Kulski, 2016). This newly synthesized double-stranded DNA is denatured which leaves the single strands attached to the flow cell so that these fragments may be amplified in clusters. After clustering, the sequence cycle initiates the addition of four different fluorescently labelled reverse terminator nucleotides along with the reaction reagents (Kulski, 2016). A laser is passed over the flow cell that excites the fluorescence in the labelled nucleotides of each cluster and the signal is captured and recognized. This cycle repeats until the sequence run has completed.

1.4.3.3. NGS limitations

Regardless of which NGS platform that is chosen, all systems produce unique sequencing errors and biases that need to be detected and corrected (Kulski, 2016). The major limitations with regards to sequencing errors across all platforms are related to the high-frequency indel polymorphisms, GC and AT rich regions, substitution errors, homopolymeric regions and replication bias (Kulski, 2016). An important element to consider for producing unbiased, high-quality and interpretable data from NGS is the achievement of sufficient depth and coverage of the sample data to infer statistical accuracy (Kulski, 2016). Preparing good quality sequence libraries is essential for producing good sequence depth and coverage as lower values may contribute to higher instances of errors stemming from incorrect base calling and mapping, which can have an effect on the statistical significance of nucleotide variants, single nucleotide polymorphisms and identifying true genotypes (Kulski, 2016).

1.4.4. Bioinformatic analysis of metabarcoding data

Over recent years, NGS has proven its versatility and robustness as an application for multiple fields of research such as clinical oncology, food forensics, genomics among many others (Wadapurkar & Vyas, 2018). However, the storage of the numerous data files and the difficulty in inferring significant conclusions from the large raw data sets remains one of the leading computational challenges that researchers face (Wadapurkar & Vyas, 2018). Therefore, NGS raw data can be seen as incredibly complex to interpret and analyze correctly, and thus requires the assistance of bioinformatics tools to help lighten the workload. In essence, bioinformatics can be defined as a science that utilizes computational tools that collect, classify, store and visualize any biochemical or biological data (Wadapurkar & Vyas, 2018).

1.4.4.1. General workflow

The bioinformatic analysis of the metabarcoding data can be described in a few simple steps: i) pre-processing and quality filtering of the sequencing data is performed, ii) detection and identification of unique amplicon sequence variants (ASVs) or operational taxonomic units (OTU) by clustering the sequences set to a predetermined threshold, iii) taxonomic assignment of the OTU's or ASV's and iv) biodiversity analysis (Alberdi *et al*., 2018). During the pre-processing step, it is vitally important to remove or denoise any possible sequencing errors from the sequencing data set. There are two possible approaches that can be applied to ensure quality filtering: i) the appropriate use of a designed tool or algorithm that is programmed to actively identify and resolve sequencing errors present in the dataset, ii) the use of quality filtering approaches that can actively remove any poor-quality sequences that don't meet the criteria set by predetermined thresholds (Alberdi *et al*., 2018). Proper planning and research needs to be conducted as both of these strategies could influence or affect the end result. Identification of the ASV's and OTU's are achieved by analyzing reads that differ by less than a predetermined dissimilarity threshold, followed by clustering them in distinct operational molecular OTU's (Piper *et al*., 2019). Caution should be taken when clustering, as this may result in either overestimating or underestimating the species read count present in the samples (Piper *et al*., 2019). It is also common practice to use bioinformatics tools to remove any chimeras or remaining artefacts that are left behind by the sequencing process. Tools, such as DADA2, can effectively distinguish between correct biological sequences and artefacts, even those that differ by as little as one nucleotide, that are created in the PCR amplification and subsequent sequencing step (Piper *et al*., 2019). With regards to taxonomic assignment and ultimately, diversity analysis, the choice of metabarcode is particularly vital. The chosen sequenced region must be variable and descriptive enough to distinguish the required biological entity at the level of resolution that is currently under investigation (Arulandhu *et al*., 2017). Especially in food microbiome analysis, often the resolution that is required is at species level. In addition, the capability of the metabarcode to identify taxa and infer species-level resolution is highly dependent on the reference database curated for the purpose of taxonomic assignment (Arulandhu *et al*., 2017). Different databases come with their own set of challenges as many of them contain sequencing errors, mismatched sequences or may over or under represent certain taxa (Arulandhu *et al*., 2017).

1.4.4.2. Tools for Metabarcoding analysis

Once the raw data is collected from the Next Generation sequencer, the first step is to trim away any unnecessary information away from the reads, such as primer and adapter sequences and to assess the quality of the NGS reads (Wadapurkar & Vyas, 2018). This involves an evaluation that removes, corrects or trims any raw data reads that do not conform to the predetermined standards set by the study, which includes errors such as poor read quality and errors in base calling.

1.4.4.2.1. Decisive Amplicon Denoising Algorithm 2

DADA2 is an open-source software package integrated into R that infers exact amplicon sequence variant (ASVs) from high-throughput amplicon sequencing data generated from NGS (Callahan *et al*., 2016). In summary, the DADA2 pipeline utilizes the demultiplexed fastQ files as input and outputs the appropriate sequence variants with their sample-specific abundances after the removal of chimera, bimera and substitution errors (Callahan *et al*., 2016). DADA2 provides an array of tools that converts raw amplicon sequencing data into a comprehensive feature table defining sample composition. Over the years, the DADA2 package has been reviewed as a more robust, sensitive and specific algorithm when compared to the most commonly used Operational Taxonomic Unit (OTU) protocols, and can resolve ASVs that differ by as little as one nucleotide (Callahan *et al*., 2016).

1.4.4.2.2. Decontam

High throughput sequencing methods have transformed the way in which microbial community and microbiome analysis is performed. However, the accuracy of the method is limited, as it introduces contaminating DNA sequences that are not truly present in the microbiome community during sequencing (Davis *et al*., 2017). These contaminating DNA sequences can be introduced from numerous sources, such as the reagents that are used during the sequencing reaction, which can gravely interfere with downstream processes. The decontam package can provide simple statistical tools that can identify and visualize varying contaminating DNA sequences, which allows them to be effectively removed from the true community dataset (Davis *et al*., 2017). A detailed description and guide of the decontam package can be found here: https://benjjneb.github.io/decontam/vignettes/decontam_intro.html. To put it simply, decontam provides a simple interface that takes in your format of sequence features and classifies each possible sequence contaminant based on previous signature contaminants that were present in previous studies (Davis *et al*., 2017). The first contaminant identification method, known as frequency, involves identifying the distribution of the frequency of each of the sequence features as a function of the input DNA concentration (Davis *et al*., 2017). The second contaminant identification method, known as prevalence, identifies the prevalence (indicates the presence/absence across all samples) of each sequence in true positive samples and compares that to the prevalence of the negative controls. No matter which method you choose, the decontam package reduces the prevalence of false positives in exploratory analysis, minimizes batch effects between different studies and sequencing runs and improves the accuracy of your sequencing dataset (Davis *et al*., 2017).

1.4.5. Advantages, technical limitations and challenges of taxon identification with DNA metabarcoding

The current popular advantage of DNA metabarcoding for taxon identification of species from ancient and modern heterogeneous samples has skyrocketed the development and availability of high throughput sequencing platforms and has paved a promising future for the identification of complex sample types (Zhang *et al.*, 2018). However, this technique does not come without several technical limitations that often results in generating both false negatives and false positives (Zhang *et al.*, 2018). This strategy relies heavily on well-designed primers that amplify a DNA region of interest in taxonomically complex samples. Therefore, difficulties often involve uncovering an appropriate target DNA region that would be able to amplify across all taxa, dealing with PCR errors and sequencing artefacts, compiling an extensive reference sequence database and deciding on suitable bioinformatics steps to analyze variable sequence divergence threshold among species (Zhang *et al.*, 2018). Choosing one or more suitable genetic markers is essential in the success and accuracy of the application as it affects both PCR amplification success and species-level resolution (Zhang *et al.*, 2018). Further research in compiling comprehensive reference databases and DNA sequencing may provide solutions to the current challenges facing DNA metabarcoding and may expand to other fields of research in the prospective future.

There are important difficulties to consider when selecting this application for a research study. During the PCR amplification in the library building step, artificially generated sequences known as chimeras are synthesized. Chimeras are artefacts that arise from an incomplete extension, which act as primers that amplify sequences incorrectly (Taberlet *et al.*, 2012). This gives the illusion of novel sequences and inflates the diversity found within a sample, which can dramatically change the outcome of a study. It is difficult to differentiate between chimeras and novel sequences, however, there are certain bioinformatic tools available that can identify chimeras and remove them. Another possible challenge facing DNA metabarcoding primers is tag jumping, which is the incorrect incorporation of tags to certain samples that could confuse the process of demultiplexing and assignment in the bioinformatics step. Along with chimeras, tag jumping may be jointly responsible for the inflation of diversity within samples as it can synthesize sequences that have incorrect but used tag combinations, which results in the erroneous assignment of sequences to samples (Taberlet *et al.*, 2012). It is not known where the phenomenon of tag jumping originated from; however, scientists have suggested that it could occur during the library building and index phase and that it may be a particular hurdle of Illumina technology. An additional concern involving DNA metabarcoding is the variation in binding efficiencies of the primers across all taxa, which may cause an unequal representation of all species that can lead to some being masked or lost in the process (Taberlet *et al.*, 2012). Unfortunately, there is not much that can be done to solve this issue. Lastly, taxa representation in databases can be minimal, which would make it increasingly difficult to match a sequence to their respective species identity. Therefore, appropriate research should be conducted on the desired barcodes before committing as it may alleviate taxonomic assignment issues towards the end of the study (Taberlet *et al.*, 2012).

1.5. Aims

The purpose of this study was to develop a robust and effective workflow that could be used to identify species present in processed vegetarian products using a metabarcoding approach. With that being said, the aim was to investigate and test suitable primers that could be used to achieve that purpose, while simultaneously establishing a cost-effective and time-efficient workflow that most investigative laboratories may adopt. Additionally, we would like to determine if any food fraud has been committed on the sample set obtained for this study and if so, determine what is present in the samples. It is also advantageous to learn more about the fungal communities present within these samples to determine possible pathogenic organisms that may cause food poisoning or other forms of bodily harm.

1.6. References

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*(1), 134–147. https://doi.org/10.1111/2041-210X.12849

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F. B., Barreto Crespo, M. T., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hochegger, R., Klingl, A., Lundberg, L., Natale, C., … Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, *6*(10), 1. https://doi.org/10.1093/gigascience/gix080

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

CFR - Code of Federal Regulations Title 21. (n.d.). Retrieved May 16, 2021, from https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm

Chauhan, A. (2020). *Food fraud – an evolving crime with profit at its heart - New Food Magazine*. https://www.newfoodmagazine.com/article/109059/food-fraud-an-evolving-crime-with-profit-at-its-heart/

Coghlan, M. L., Haile, J., Houston, J., Murray, D. C., White, N. E., Moolhuijzen, P., Bellgard, M. I., & 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. https://doi.org/10.1371/journal.pgen.1002657

Coissac, E. (2012). OligoTag: A program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods in Molecular Biology*, *888*, 13–31. https://doi.org/10.1007/978-1-61779-870-2_2

Coskun, O. (2016). Separation Tecniques: CHROMATOGRAPHY. *Northern Clinics of Istanbul*, *3*(2), 156. https://doi.org/10.14744/nci.2016.32757

D'Amato, M. E., Alechine, E., Cloete, K. W., Davison, S., & Corach, D. (2013). Where is the game? Wild meat products authentication in South Africa: a case study. *Investigative Genetics*, *4*(1), 6. https://doi.org/10.1186/2041-2223-4-6

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC - Trends in Analytical Chemistry*, *85*(November), 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2017). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/221499

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., & Newcomb, R. D. (2019). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution*, *10*(1), 120–133. https://doi.org/10.1111/2041-210X.13086

Eisenstein, M. (2018). Microbiology: Making the best of PCR bias. *Nature Methods*, *15*(5), 317–320. https://doi.org/10.1038/nmeth.4683

Elnifro, E. M., Ashshi, A. M., Cooper, R. J., & Klapper, P. E. (2000). Multiplex PCR: Optimization and Application in Diagnostic Virology. *Clinical Microbiology Reviews*, *13*(4), 559. https://doi.org/10.1128/CMR.13.4.559-570.2000

Fatica, A., Di Lucia, F., Marino, S., Alvino, A., Zuin, M., De Feijter, H., Brandt, B., Tommasini, S., Fantuz, F., & Salimei, E. (2019). Study on analytical characteristics of Nicotiana tabacum L., cv. Solaris biomass for potential uses in nutrition and biomethane production. *Scientific Reports*, *9*(1), 1–8. https://doi.org/10.1038/s41598-019-53237-8

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017a). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/s41467-017-01312-x

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017b). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/s41467-017-01312-x

Griffith, C. (2016). Surface Sampling and the Detection of Contamination. In *Handbook of Hygiene Control in the Food Industry: Second Edition* (pp. 673–696). Elsevier Inc. https://doi.org/10.1016/B978-0-08-100155-4.00044-3

Han, J., Zhu, Y., Chen, X., Liao, B., Yao, H., Song, J., Chen, S., & Meng, F. (2013). The Short ITS2 Sequence Serves as an Efficient Taxonomic Sequence Tag in Comparison with the Full-Length ITS. *BioMed Research International, 2013,* 1–7. https://doi.org/10.1155/2013/741476

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/RSPB.2002.2218

Https://www.westerncape.gov.za/text/2016/August/regulations_-_relating_to_the_labelling_and_advertising_of_foodstuffs_-_r_1055_of_2002. (n.d.). *(No Title)*. Retrieved May 16, 2021, from https://www.westerncape.gov.za/text/2016/August/regulations_-_relating_to_the_labelling_and_advertising_of_foodstuffs_-_r_1055_of_2002.pdf

Johnston, R. (2018). *Vegetarian food | Smart Food Expo*. https://smartfoodexpo.ru/en/vegetarianskoe-pitanie

Kadri, K. (2020). Polymerase Chain Reaction (PCR): Principle and Applications. In *Synthetic Biology - New Interdisciplinary Science*. IntechOpen. https://doi.org/10.5772/intechopen.86491

Kang, T. S. (2019). Basic principles for developing real-time PCR methods used in food analysis: A review. In *Trends in Food Science and Technology* (Vol. 91, pp. 574–585). Elsevier Ltd. https://doi.org/10.1016/j.tifs.2019.07.037

Koen, N., Blaauw, R., & Wentzel-Viljoen, E. (2016). Food and nutrition labelling: the past, present and the way forward. *South African Journal of Clinical Nutrition*, *29*(1), 13–21.

https://doi.org/10.1080/16070658.2016.1215876

Koetsier, G., Cantor, E., & Biolabs, E. (2019). *A Practical Guide to Analyzing Nucleic Acid Concentration and Purity with Microvolume Spectrophotometers*.

Kulski, J. K. (2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. In *Next Generation Sequencing - Advances, Applications and Challenges*. InTech. https://doi.org/10.5772/61964

Lau, J.-E. (2021). *Food Fraud - Part I - What Is Food Fraud? | Food Technology |Science Meets Food*. https://sciencemeetsfood.org/food-fraud-what-is-food-fraud/

Liu, C., Qi, R.-J., Jiang, J.-Z., Zhang, M.-Q., & Wang, J.-Y. (2019). Development of a Blocking Primer to Inhibit the PCR Amplification of the 18S rDNA Sequences of Litopenaeus vannamei and Its Efficacy in Crassostrea hongkongensis. *Frontiers in Microbiology*, *0*(APR), 830. https://doi.org/10.3389/FMICB.2019.00830

Liu, L., Wang, Y., He, P., Li, P., Lee, J., Soltis, D. E., & Fu, C. (2018). Chloroplast genome analyses and genomic resource development for epilithic sister genera Oresitrophe and Mukdenia (Saxifragaceae), using genome skimming data. *BMC Genomics*, *19*(1), 235. https://doi.org/10.1186/s12864-018-4633-x

Mafra, I., Ferreira, I. M. P. L. V. O., & Oliveira, M. B. P. P. (2008). Food authentication by PCR-based methods. *European Food Research and Technology*, *227*(3), 649–665. https://doi.org/10.1007/s00217-007-0782-x

Martin, M. (2011). *Cutadapt removes adapter sequences from high-throughput sequencing reads | Martin | EMBnet.journal*. EMBnet.Journal 17, 10. https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

Matlock, B. (2019). *Assessment of Nucleic Acid Purity*. www.thermoscientific.com

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*(1), 263–273.

https://doi.org/10.1101/sqb.1986.051.01.032

Ng, J., Satkoski, J., Premasuthan, A., & Kanthaswamy, S. (2014). A nuclear DNA-based species determination and DNA quantification assay for common poultry species. *Journal of Food Science and Technology*, *51*(12), 4060–4065. https://doi.org/10.1007/s13197-012-0893-7

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *0 3*(SUPPL.42). https://doi.org/10.1002/0471250953.bi0301s42

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). *Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance*. *8*, 1–22. https://doi.org/10.1093/gigascience/giz092

Prado, M., Ortea, I., Vial, S., Rivas, J., Calo-Mata, P., & Barros-Velázquez, & J. (2016). Advanced DNA-and Protein-based Methods for the Detection and Investigation of Food Allergens. *Critical Reviews in Food Science and Nutrition*, *56*, 2511–2542. https://doi.org/10.1080/10408398.2013.873767

Schon, E. A., DiMauro, S., & Hirano, M. (2012). Human mitochondrial DNA: roles of inherited and somatic mutations. *Nature Reviews. Genetics*, *13*(12), 878–890. https://doi.org/10.1038/nrg3275

Simpson, F. (2018, June 9). *Sainsbury's, Tesco Investigated After Meat Traces Reported in Vegetarian Food*. https://www.businessinsider.com/sainsburys-tesco-investigated-after-meat-traces-reported-in-vegetarian-food-2018-6?IR=T

Sint, D., Raso, L., & Traugott, M. (2012). Advances in multiplex PCR: Balancing primer efficiencies and improving detection success. *Methods in Ecology and Evolution*, *3*(5), 898–905. https://doi.org/10.1111/j.2041-210X.2012.00215.x

Skypala, I. J. (2019). Food-induced anaphylaxis: Role of hidden allergens and cofactors. In *Frontiers in Immunology* (Vol. 10, Issue APR, p. 673). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2019.00673

Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T. W., & Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry*, *408*(17), 4615–4630. https://doi.org/10.1007/s00216-016-9595-8

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. https://doi.org/10.1111/j.1365-294X.2012.05470.x

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. https://doi.org/10.1093/nar/gkl938

UNITE - Resources. (n.d.). Retrieved October 10, 2020, from https://unite.ut.ee/repository.php

Wadapurkar, R. M., & Vyas, R. (2018). Computational analysis of next generation sequencing data and its applications in clinical oncology. *Informatics in Medicine Unlocked*, *11*, 75–82. https://doi.org/10.1016/J.IMU.2018.05.003

Wheeler, R. (2016). *Management of allergens in the food industry – still a hot topic*. https://www.foodfocus.co.za/home/Industry-Topics/food-safety/Management-of-allergens-in-the-food-industry

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., & Chen, S. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE*, *5*(10). https://doi.org/10.1371/journal.pone.0013102

Zhang, G. K., Chain, F. J. J., Abbott, C. L., & Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolutionary Applications*, *11*(10), 1901–1914. https://doi.org/10.1111/eva.12694

Chapter 2: Describing the ingredient composition of processed vegetarian samples using multi-locus DNA metabarcoding.

2.1. Introduction

2.2.1. Why is food labelling important?

Food labelling and packaging inform the consumer with regards to the composition and nature of the products to avoid any confusion and protect the customer from any adverse risks. These risks are commonly associated with known food allergens and ingredients that could potentially cause cardiovascular disease when consumed in larger quantities (Koen *et al*, 2016). A typical food label includes the marketing information, brand name, safe storage, food preparation/composition information and the declaration of principal ingredients, including all the potential allergens, so that the consumer can make an informed decision regarding their nutrition (Koen *et al*., 2016). Nutrition labelling is considered a population-based approach that can positively influence the dietary habits of consumers that could ultimately contribute to the accomplishment of government-regulated public health objectives (Koen *et al*., 2016). In addition, nutrition labelling can be considered a valuable educational toolkit for health care professional's mission to educate clients on better nutrition and healthier lifestyle choices (Koen *et al*., 2016).

2.1.2. Vegetarian food and food fraud

Due to the ever-changing opinions and information readily available on the internet, many individuals have conducted their own research, and have decided to switch to a more sustainable, plant-based lifestyle. Although there is no legal definition explaining the word "vegan", the Food Standards Agency (FSA) has filled in the gaps by providing voluntary guidance on the use of the term vegetarian and veganism with regards to food labelling (Chauhan, 2020). These guidelines include advice around controlling cross-contamination and providing manufacturers meaningful and actionable steps on how to demonstrate that foods presented as 'vegetarian' or 'vegan' have not been contaminated during storage, preparation, cooking or display (Chauhan, 2020).

Over the years, there has been a rise in interest around plant-based foods that look, taste and have a similar texture to that of meat products (Chauhan, 2020). However, many consumers are wary of the possibility that these foods can be susceptible to food fraud because of the consumer's potential to be misled by similar but non-plant ingredients. A recent report issued by the Guardian brought to light a multitude of issues that are prompting distrust among loyal customers in England (Chauhan, 2020). In response to this growing trend seen in the United Kingdom, the Centre for Food Safety in the US has appealed to the Food and Drink Administration to block all sales related to untested meat-like, plant-based burgers because the products contain untested lab-produced food dyes that have the potential to create a 'bleeding effect' that may have unknown consequences to consumer health (Chauhan, 2020).

The majority of instances of consumer doubt and distrust should be considered a major red flag for manufacturers and producers of plant-based products. Items labelled as natural or organic are regarded as premium or high-value items which often fall victim to either fraud or the fallout of consumer doubt. That is why it is imperative to create systems and develop strategies that increase consumer confidence by addressing concerns regarding misrepresentation of ingredients and presenting honest facts backed by scientific research and reliability testing.

2.1.3. Aim

In this study, our aim was to develop an effective metabarcoding system that could describe ingredients at the species level of plant-based products and possible meat contaminant traces by using reliable DNA metabarcoding assays targeting animals. Due to the incredible advancement of metabarcoding technologies and the limited knowledge of the vegetarian food mislabelling prevalence of products sold in South Africa, we aimed to identify the most comprehensive species composition of 32 processed food samples collected in the Cape Town region.

2.2. Materials and Methods

The protocols and procedures exhibited in this thesis have been developed and modified by the Forensic DNA Laboratory.

2.2.1. Collection and preparation of food samples

During May to June 2018, thirty-two vegetarian food products were purchased and collected in the Cape Town area from local and franchised supermarkets. These samples consisted of highly processed vegetarian food items such as sausages, schnitzels and burgers. A simplified overview of the ingredients listed on the provided packaging and their presence in the sample data are stated in **Table 2.1**. The processed vegetarian frozen food samples mostly consisted of a mixture of different plant and fungi species, while some were held together by certain animal-derived components, such as egg whites, that were fully disclosed on the packaging (Please refer to **Table 2.2.** for those samples). Some of the samples may have been processed and packaged in factories that cater, and provide their services to, a wide variety of brands which could subject the vegetarian products to animal component and/or allergens contamination.

Table 2. 1: Simplified overview of listed ingredients and their incidence in the sample data (Number of samples counted with the ingredient, #S). Ingredients that could contain two or more ingredients were left blank, as there could be a potential match to multiple species.

Labeled as	Species match	#S
Vegetable Oil		26
Garlic	*Allium sativum*	20
Onion	*Allium cepa*	19
Sea Salt		16
Soya	*Glycine Max*	14
Wheat	*Triticum spp.*	13
Mustard	*Sinapis alba or Brassica juncea*	11
Potato	*Solanum tuberosum*	11
Mycoprotein	*Fusarium venenatum*	10
Yeast	*Saccharomyces cerevisiae*	8
Egg	*Gallus gallus*	7
Black Pepper	*Piper nigrum*	5
Pea	*Pisum sativum*	5
Sugar	*Saccharum officinarum*	5
Bean	*Phaseolus vulgaris*	5
Cumin	*Cuminum cyminum*	4
Maize	*Zea mays*	4
Milk	*Bos Taurus*	4

Sage	*Salvia officinalis*	4
Chickpea	*Cicer arietinum*	3
Chilli	*Capsicum frutescens*	3
Coriander	*Coriandrum sativum*	3
Ginger	*Zingiber officinale*	3
Tomato	*Lycopersicon esculentum*	3
All Spice	*Cinnamomum verum* *Syzygium aromaticum* *Myristica fragrans* *Zingiber officinale*	2
Bay Leaf	*Laurus nobilis*	2
Butternut	*Cucurbita moschata*	2
Cayenne Pepper	*Capsicum annuum*	2
Celery	*Apium graveolens*	2
Clove	*Syzygium aromaticum*	2
Olive Oil	*Olea eurpaea*	2
Oyster Mushroom	*Pleurotus ostreatus*	2
Oregano	*Origanum vulgare*	2
Parsley	*Petroselinum crispum*	2
Quinoa	*Chenopodium quinoa*	2
Barley	*Hordeum vulgare*	2
Seaweed	*Chondrus crispus*	2
White Pepper	*Piper nigrum*	2
Rice	*Oryza sativa*	2
Butter Bean	*Phaseolus lunatus*	1
Carrot	*Daucus carota subsp. sativus*	1
Chia	*Salvia hispanica*	1
Curry Leaf	*Murraya koenigii*	1
Lemon	*Citrus limon*	1
Lentil	*Lens culinaris*	1
Marjoram	*Origanum majorana*	1
Masala		1
Mint	*Mentha spp.*	1
Mushroom	*Agaricus bisporus*	1
Oat	*Avena sativa*	1
Rosemary	*Salvia rosmarinus*	1
Paprika	*Capsicum annuum*	1
Spinach	*Spinacia oleracea*	1
Sweet Potato	*Ipomoea batatas*	1
Tapioca	*Manihot esculenta*	1
Tumeric	*Curcuma longa*	1

Table 2. 2: Vegetarian samples containing animal products (food additives and flavourants). These ingredients were flagged and acknowledged as a possible reason for the presence of DNA of animal origin within the sample. Sample IDs are the unique tag given to each product for the protection of the manufacturers/producers identity.

Sample IDs	Ingredient	Animal Origin
F238	Egg/Reconstituted Free-Range Egg White	Chicken
	Milk	Cow
F239	Egg/Reconstituted Free-Range Egg White	Chicken
	Milk	Cow
F243	Egg/Reconstituted Free-Range Egg White	Chicken
	Milk	Goat
	Milk	Cow
F245	Egg/Reconstituted Free-Range Egg White	Chicken
	Milk	Goat
	Milk	Cow
F254	Egg	Chicken
F255	Egg	Chicken
F260	Mature vegetarian cheddar cheese	Cow
	Soft cheese (buttermilk and cream)	Cow
	Egg/Reconstituted dried free range egg white	Chicken
	Milk	Cow
F261	Reconstituted dried free range egg white	Chicken
	Milk	Goat
	Milk	Cow
F263	Egg/ Reconstituted dried free range egg white	Chicken
	Milk	Goat
	Milk	Cow

The samples were brought and catalogued to the Forensic DNA Laboratory (FDL) based in the University of the Western Cape (UWC), where each sample was assigned a specific reference number that could be traced to reveal the details of the sample. These details included a description, the brand, ingredients listed, all notable information as well as the pictures of the packaging before aliquoting. Once all of the samples had been referenced and catalogued, the food items were removed from their packaging and approximately 200 mg to 450 mg of the sample were weighed out using a Balance Precision PS 4500 R2 scale (Radwag) and placed into Safe-Lock 2 ml Micro Test Tubes (later referred to as test tubes) to be used as input for DNA extractions. A plastic pestle was used to manually break and mix the more cellulosic samples. What remained of the samples, as well as the tubes containing the aliquots, were kept in a -20˚C freezer storage for preservation.

2.2.2. Cetyltrimethylammonium Bromide (CTAB) DNA extraction

Multiple preliminary tests were conducted on 10 vegetarian samples to determine the optimal protocol for the extraction of plant-based material. Once a suitable extraction protocol had been modified and established, it was performed on the remainder of the samples as well as a set of positive controls. All sample preparations were performed in a NuAIRE™ Biosafety Cabinet (to minimize the exposure of the samples to contaminants.

2.2.3. Sample lysis

For plant-based material, a Cetyltrimethylammonium Bromide (CTAB) solution was prepared (Please refer to Appendix A: Solutions). To accelerate the dissolution of the CTAB, the powder was dissolved in a small volume of distilled water, incubated at 55˚C for 15 minutes before the addition of the reagents referred to in Appendix A were made to a final volume of distilled water. In the aliquoted food sample test tubes, 1ml of lysis buffer and 20 µl of 20 mg/ml Proteinase K was transferred. Proteinase K is used to digest any contaminating proteins found in the sample. Following the previous step, each tube was wrapped in parafilm before shaking incubation, using a Labnet™ Vortemp 56 Shaking Incubator, was performed at 65˚C overnight. Once the incubation period had elapsed, the samples were centrifuged using a Labnet™ Prism Microcentrifuge at 5200 rpm for 10 minutes, the clean supernatant transferred to a sterile tube and the cellular debris discarded. This step was repeated to ensure that no cellular debris carryover was present in the following steps.

2.2.4. Chloroform/Isoamyl alcohol DNA extraction

Following the lysis step, the purified supernatant of each sample was transferred to Safe-Lock 1.5ml Micro test tubes with an equal volume of Chloroform:Isoamyl alcohol 24:1, where they were then thoroughly mixed by shaking. Chloroform:Isoamyl alcohol prevents the emulsification of a solution. Once the samples had been centrifuged at 10 000 rpm for 10 minutes, the upper aqueous supernatant was carefully transferred to a clean test tube while the lower, organic solution was largely left undisturbed. During the preliminary trials, the recommended step that followed involved DNA precipitation by adding one volume of 100% ice-cold isopropanol to each test tube before vortexing and incubating on ice for 30 minutes. The ice-cold isopropanol precipitates the DNA. However, during the experimental protocol, 70% ice-cold ethanol equal to the volume of the aliquot was transferred to all sample test tubes in error before following the remainder of the preliminary trial procedure. To rectify this issue, an additional one-time volume of 100% ethanol was transferred to the supernatant, vortexed and was allowed to incubate on ice for 30 minutes for DNA precipitation (https://bitesizebio.com/2839/dna-precipitation-ethanol-vs-isopropanol/). Ethanol was used as it precipitates small volumes of DNA effectively, while reducing the risk of precipitating excess salt. From this point, both protocols converged, whereby the samples were centrifuged for 15 minutes at 13 000 rpm, the supernatant discarded and the DNA pellet washed with ~400 µl ice-cold 70% ethanol before gently tapping the tube to avoid disturbing the pellet. Once suitably mixed, the test tubes were centrifuged for a further 5 minutes spinning at 13 000 rpm, the supernatant discarded and the previous steps repeated twice. Lastly, the samples were allowed to air dry for 30 minutes before the addition of 150 µl of Tris-Ethylene-diamine-tetra-acetic Acid (TE buffer) and were left to rehydrate for 15 minutes at 55˚C. TE buffer solubilizes the DNA while protecting it from degradation. The pure DNA extracts were placed in DNA LoBind tubes for long-term storage and a small aliquot was normalized to 10 ng/µL for future PCR experiments.

The DNA extractions were performed following the recommendations set out by good laboratory practice with regards to exposure to chloroform: the standard centrifuge was placed into a Vivid Air™ fume hood to ensure that no chloroform vapours would escape into the lab and double gloves were worn to prevent skin irritation.

2.2.5. DNA Quantitation using Nanodrop and Qubit dsDNA HS Assay kit

Firstly, DNA quantitation of the food samples was achieved by applying 1 µl of sample to a NanoDrop™ 2000 UV/VIS Spectrophotometer. To verify these results, a Qubit™ dsDNA HS (High Sensitivity) Assay kit was performed on diluted DNA extraction aliquots and the results measured on a Qubit™ 2.0 Fluorometer. The assay was conducted by following the standard protocol provided by the manufacturer.

2.2.6. Primer selection

The aim of this study was to identify a fast and cost-effective protocol that provides a consistently high quality and quantity of DNA from a mixture of plant tissue cells, based on the ingredient description. Additionally, we investigated the possible presence of other sources of DNA (mammal and birds). We selected these primers based on literature; targeting plants, fungi and animal sources (refer to **Table 2.3**). The primer sequences, and their modifications, are reported in **Table 2.4.** All the PCR primer sequences (forward and reverse) contained a short, unique oligonucleotide identifier (tag) at the 5'end of the DNA strand. A unique set of tags was designed for this study using oligoTag (Coissac, 2012). The tags created were designed to be 7 base pairs in length with a minimum Hamming distance of 3 between tags to reduce the probability of assignation errors between the samples. Additionally, they were designed to contain no more than 3 guanine's or/and Cytosine's and the overall length of the homopolymers were limited to two. The tagging designed for both ends primer allowed for the sorting of sequences corresponding to the linked sample during the bioinformatic processing step in order to reduce the probability of sample misalignment (Coissac, 2012). A restrictive dual-indexing approach was utilized as described by Frøslev *et al* (Frøslev *et al*., 2017a).

Table 2. 3: Primers used to amplify the DNA target group of interest. Target groups are colour-coded according to the type of organism they isolate. Key: **Plant = Green**, **Fungi = Dark green**, **Mammals, birds and vertebrates = Red**.

Main target group	Target DNA and gene	F Primer	R Primer	Reference
Plant	Chloroplast trnL	trnLc	trnLh	Taberlet *et al.*, 2007
Plant	Internal transcribed spacer (ITS) 2	S2F	ITS4	Chen *et al.*, 2015 and White *et al.*, 2018
Plant/Fungi	Internal transcribed spacer (ITS) 2	gITS7	ITS4	Ihrmark *et al.*, 2012 and White *et al.*, 2018
Mammal	Mitochondrial 16S ribosomal RNA (16S)	16Smam1	16Smam2	Taylor, 1994
Vertebrate	Cytochrome b (Cytb)	L14816	H15173	Parsonet *et al.*, 2000

Table 2. 4: The primer sequences obtained from literature and their respective modifications highlighted in red. Primers are reported in the 5' → 3' direction. Please refer to Table 2.1 for the target group for each primer.

Primer	F or R	Original Primer Sequence	Modification
trnLc	Forward	CGAAATCGGTAGACGCTACG	None
trnLh	Reverse	CCATTGAGTCTCTGCACCTATC	None
S2F	Forward	ATGCGATACTTGGTGTGAAT	None
ITS4	Reverse	**TCCTC**CGCTTATTGATATGC	GCTTATTGATATGC**TTAARYTCAGC**
gITS7	Forward	GTGARTCATCGARTCTTTG	None
ITS4	Reverse	**TCCTC**CGCTTATTGATATGC	GCTTATTGATATGC**TTAARYTCAGC**
16Smam1	Forward	CGGTTGGGG**T**GACCT**C**GGA	CGGTTGGGG**Y**GACCT**Y**GGA
16Smam2	Reverse	GCTGTTATCCCT**A**GGGTA**A**CT	GCTGTTATCCCT**R**GGGTA**R**CT**TG**
L14816	Forward	CCATCCAACATCTCAGCATGATGAAA	None
H15173	Reverse	CCCCTCAGAATGATATTTGTCCTCA	None

Due to their performance in previous authentication studies, we decided to choose the following mini barcode regions for the tentative identification of both animal, plant and fungal species (Please refer to **Table 2.5** below).

Table 2. 5: An overview of the mini-barcodes chosen for the authentication of plant, fungi and animal species. The table includes the shortened version of each primer name as well as the reasoning behind each choice with an appropriate reference.

Mini Barcode	Reason for Choice
Animal Sources:	
16S ribosomal RNA (16S rRNA) **(Coghlan *et al*., 2012)**	• Used in taxa identification in food products. • Identify wildlife species in traditional medicines.
Cytochrome B (CytB) **(D'Amato *et al*., 2013)**	• Previously used in food authentication studies. • Large reference database available for sequence comparison.
Plant/Fungi Sources:	
tRNALeu – trnL – UAA intron **(Taberlet *et al*., 2007)**	• Previously tested in food and biodiversity authentication studies in highly degraded DNA.
Nuclear transcribed spacer 2 (ITS2) **(Han *et al*., 2013)**	• Standard universal barcode for the identification of both plant and fungal species.

Depending on the region of interest, the appropriate PCR primers were selected according to **Table 2.6** below. Due to the products being of a processed nature, we included animal and plant primers that can target short barcode regions (16S rRNA and trnL ~ 150 nucleotides (nt) in length). These two primers were chosen to account for the possible presence of highly degraded DNA and could supplement for poor amplification performance among the other selected primers. The remaining primer sets: CytB and ITS2 commonly produce longer amplicons (>300 nt) which can provide better accuracy and discrimination at the species level.

Table 2. 6: Overview of the multiplex primers that were selected for this study with reference to their modifications (_m and highlighted in red), if applicable. The primer sequences are reported in the 5'→ 3'. The final concentrations of the primers in the multiplex are reported in µM. Modifications were appropriate in certain circumstances in order to avoid GC rich regions.

F or R	Primer	Primer sequences	µM	Dataset
Animal Primers				
CYTB_F	L14816	CCATCCAACATCTCAGCATGATGAAA	0.15	Cyt_B
CYTB_R	H15173	CCCCTCAGAATGATATTTGTCCTCA	0.15	
16s_F	16Smam1**_m**	CGGTTGGGG**Y**GACCT**Y**GGA	0.55	
16s_R 1	16Smam2**_m**	GCTGTTATCCCT**R**GGGTA**R**CT**TG**	0.55	M_16S
Plant/Fungi Primers				
ITS2P_F	S2F	ATGCGATACTTGGTGTGAAT	0.45	P_ITS2
ITS2_R	ITS4**_m**	GCTTATTGATATGC**TTAARYTCAGC**	0.45	
ITS2F_F	gITS7	GTGARTCATCGARTCTTTG	0.25	F_ITS2
trnL_F	trnLg	CGAAATCGGTAGACGCTACG	0.45	
trnL_R	trnLh	CCATTGAGTCTCTGCACCTATC	0.45	P_trnL

2.2.7. PCR amplification

Initially, each of the primers chosen for this study were tested in singleplex to determine the amplification efficiency of the primers when present in heterogenous plant samples. Once the tests had proven the efficacy of the primers, numerous primer combination PCR tests were conducted to determine the most optimal combination of primers that allowed for the most balanced DNA template optimization while limiting potential amplification biases. Ultimately, the best performing primer combination and concentration was utilized for this study. Only one PCR multiplex including plant and animal primers were chosen and utilized for the purpose of this study. The PCR's were performed using 1 µl of 10 ng/µL DNA aliquot in a final volume of 25 µL. The reagent mixture contained 2.5 µl of GeneAmp ® 10X PCR Buffer II (Applied Biosystems), 2.5 µl of 25 mM MgCl2 (Applied Biosystems), 1 µl of 20 mg/ mL Bovine Serum Albumin (BSA), 0.5 µL of 83 dNTP (each 10 mM), 0.2 µL of 5U/µL AmpliTaq Gold DNA polymerase (Applied Biosystems), 1.25 µl of forward as well as reverse primers and 14.8 µL of molecular biology water (Lonza). The final concentration of each primer set is shown in **Table 2.6**. Negative controls, extraction blanks and positive controls were prepared for the amplification stage and library preparation. As they will be used in further analysis, the positive animal controls were prepared as follows in **Table 2.7** below.

Table 2. 7: Meat positive controls prepared to analyse the effectiveness of the animal primers. These positives were extracted using the CTAB lysis and Chloroform/Isoamyl extraction method described earlier in the chapter. PCR_P1 was created by adding tissue from cow, pig and chicken in equal DNA concentration. PCR_P2 was created by co-extracting equal parts cow, pig and chicken tissue.

Sample Name	Sample ID	Total 5ng/µL	Final Volume (60)	Qubit [ng/µL]
Cow	PCR_P1	1.7	0.63	160
Pig		1.7	0.42	236
Chicken		1.7	0.36	274
H_2O			58.59	
Mix (Cow_Pig_Chicken)	PCR_P2	5	0.57	526

The performance of the PCR multiplex amplification prior to pooling was visualized using gel electrophoresis on 2% agarose. The 2% (w/v) agarose gels were prepared using a 1X TBE buffer solution. Loading buffer was prepared by transferring 1µl of Biotium GelRed™ Nucleic Acid Gel Stain into 1ml of methyl orange buffer. Once the gels were set, 4µl of the loading buffer and 5µl of each sample were mixed and ran alongside 1 µl of HyperLadder™ 25 bp and HyperLadder™ 100 bp. Gel Electrophoresis was run using a 1X TBE running buffer set at 150 volts for approximately 30 minutes. Once the time had elapsed, the gels were removed from the tanks and placed into an ENDURO^{TM} GDS Gel Documentation System where images of the fluorescence were captured. For each sample, the products of 3 independent PCR amplifications were sequenced for 32 samples, for a total of 130 PCRs including 4 negative and 30 positive controls. The reverse and forward primers utilised for PCR amplification included in their sequence a unique oligonucleotide sequence (tag), which is later used during bioinformatic processing to assign reads to samples. A total of 80 primers uniquely tagged were designed, which allowed for the pooling of up to 80 PCR products in 1 sequencing library. Therefore, the data for this study was produced by sequencing 2 libraries which contain 64 and 32 pooled samples.

2.2.8. Library Preparation and Sequencing

2.2.8.1. Purification of pooled PCR products using the Qiaquick kit

An appropriate volume of EB buffer for samples was aliquoted into a 2 ml microcentrifuge tube and placed on a heat block set to 37˚C. Approximately 300 µl of the pooled PCR products of each library was placed into separate 2ml microcentrifuge tubes containing 1500 µl of PB buffer with pH indicator. These were mixed thoroughly by pipetting, ensuring that the mixture had changed its colour to yellow. If the mixture turned an orange or violet colour, the pH of the solution was corrected by adding 3M sodium acetate and mixed thoroughly. After the appropriate pH had been achieved for all samples, approximately 650 µl of each sample was added to separate spin columns and centrifuged for 1 minute at 13 000 rpm before discarding the flow-through. This step was repeated 3 times. Once this had been completed, a volume of 740 µl of PE buffer was transferred into each of the spin columns before being centrifuged for 1 minute at 13 000 rpm. After the time had elapsed, the flow-through was discarded and the empty spin columns centrifuged for an additional 2 minutes at 13 000 rpm. The columns were placed in clean 1.5 ml microcentrifuge tubes, 30 µl of warm EB buffer transferred to the centre of the membrane of each column and were incubated at 37˚C for 15 minutes. Once the incubation time had elapsed, the spin columns were centrifuged at 13 000 rpm for 1 minute, the filter turned to the opposite side and centrifuged once more at 13 000 rpm for 1 minute. Each sample was then transferred into clean 1.5 ml LoBind Eppendorf tubes. The samples were quantified using a Qubit dsDNA HS Assay Kit and Qubit 2.0. Fluorometer. An aliquot of each sample was taken and diluted within range of 0.5 ng/µL to be analyzed on an Agilent 2200 TapeStation. Purified DNA inputs with concentrations exceeding 250ng were used to calculate the appropriate volume required for 50 µl reaction.

2.2.8.2. Library preparation and purification

Following the purification of the PCR products of each library, library building was carried out using TruSeq DNA PCR-free Library kit (Illumina). Approximately 40 µl of ERP2, 10 µl of EB buffer and 50 µl of the purified PCR product were transferred and thoroughly mixed. The tubes were placed in a thermocycler set to 30˚C for 30 minutes. After the time had elapsed, the samples were purified using a MinElute PCR purification kit (Qiagen). After purification, the samples were transferred into clean PCR tubes. A volume of 17.5 µl of product and 12.5 µl of ATL was transferred into their respective clean PCR tube and mixed thoroughly by a vortex. Once the samples had been spun down, they were placed in a thermocycler set to the following parameters: 37˚C for 30 minutes, 70˚C for 5 minutes and 4˚C for 5 minutes. For the next step, an appropriate adapter index was assigned to each sample set (AD001 to Library 1 and AD002 to Library 2). After this assignment, 30 µl of each product was transferred to clean PCR tubes and the following reagents were added to these tubes: 2.5 µl of Resuspension buffer, 2.5 µl of LIG 2 and 2.5 µl DNA Adapter Index. Once these tubes had been thoroughly mixed and spun down, they were placed in a thermocycler set to 30˚C for 10 minutes. After the time had elapsed, the PCR tubes were placed on a cooler and 5 µl of Stop Ligation Buffer (STL) were added to each sample and mixed. These tubes were incubated for a further 5 minutes on the cooler tray. The samples were purified using a MinElute PCR purification kit (Qiagen) and quantified using a Qubit HS kit (Thermo Fisher Scientific) and Qubit 2.0 Fluorometer (Thermo Fisher Scientific). An aliquot of each sample was taken and diluted within range of 0.5 ng/µL to be analyzed on an Agilent 2200 TapeStation. The remainder of the samples was placed in a -20˚C freezer.

2.2.8.3. Library purification with AMPure XP kit

A final purification with AMPure XP beads was performed to selectively bind DNA fragments 100 bp or larger to paramagnetic beads. Around half an hour before the start of the protocol, the AMPure XP bead suspension was taken out of the fridge and allowed to equilibrate to room temperature. In preparation for the following steps, 80% ethanol was prepared from molecular grade ethanol and RNase free water. The exact volume of 15 µl of each of the samples was transferred into separate 1.5 ml tubes and the volume was adjusted to 50 µl with RNase free water. The AMPure beads were vortexed thoroughly before 75 µl of bead suspension was transferred into each of the PCR product libraries. The solutions were gently but thoroughly pipetted at least 10 times to ensure that the samples were completely homogenous before incubation at room temperature for 15 minutes. The tubes were then placed on a magnet for 2 minutes or until the supernatant had completely cleared before it was discarded, having paid particular attention to not disturb the beads. While the tubes remained on the magnet stand, 200 µl of 80% ethanol was transferred into the tubes without disturbing the beads. The tubes were allowed to incubate for 30 seconds before the supernatant was discarded and the beads left undisturbed. This step was repeated, ensuring all excess ethanol had been removed. After the excess ethanol had been removed, the tubes were air-dried on the magnetic stand for 5 minutes at room temperature. After the time had elapsed, 30 µl of nuclease EB buffer was added to each sample and the tubes were removed from the magnet. The beads were resuspended thoroughly and incubated at room temperature for 2 minutes. The tubes were placed back on the magnetic stand for 1 minute at room temperature. The supernatant with the eluted library was then transferred to new DNA Lo-Bind tubes. The eluted library was quantified using Qubit HS kit and an aliquot of each sample was taken and diluted within range of 0.5 ng/µL to be analyzed on an Agilent 2200 TapeStation. The pooled libraries were then sequenced on a Miseq (Illumina) using one 300 bp paired-end run (MiSeq Reagent Kits v2), at the Danish National High-throughput DNA Sequencing Centre.

2.2.9. Sequencing Data

2.2.9.1. Preprocessing Sequencing Data

Statistical analysis on the NGS raw sequence reads was performed in R version 3.6.3 (https://cran.r-project.org/bin/windows/base/old/3.6.3/). The raw sequencing reads were initially processed using a custom script provided by Frøslev *et al* that assigned demultiplex reads to specific samples based on their unique tag sequences before merging the two sense reads. Cutadapt (Martin, 2011) was used to trim the tags and primer sequences from the reads. The DADA2 package built-in R was used for data filtering and chimera removal using the *removeBimeraDenovofunction* before the sense and antisense reads were merged to create an ASV table. The package utilizes a parametric error mode for filtering, which is able to use quality information within its error model (Callahan *et al*., 2016). This package is able to efficiently control errors so that biological sequences that differ by 1 bp can be retained and avoids unintentionally collapsing of closely related species during the clustering process.

2.2.9.2. Taxanomic Assignment

For the FITS2 dataset (please refer to **Table 2.5**), the ASVs were taxonomically classified across multiple ranks with the native implementation in DADA2 (“assignTaxonomy” function, minBoot = 50) of the naïve Bayesian classifier method trained on the DADA2-curated general release of the UNITE ITS database (Fasta format, version no 8) (https://unite.ut.ee/repository.php). A

BLAST of the fungi ITS2 ASV sequences was performed against the NCBI Genbank nucleotide database (Callahan *et al*., 2016). All ASVs assigned to the kingdom “Viridiplantae” by BLAST were excluded from further analysis.

For the other datasets in **Table 2.5**, we performed a BLAST (v.2.8.1) of the ASVs (At least 90% for query coverage, 80% for sequence identity and the first 100 matches retained). We identified the best BLAST matches according to the thresholds set using the custom script provided: (https://github.com/tobiasgf/biowide_synthesis/blob/master/R/unitax_lineage.R). A species-level annotation was assigned if the best BLAST match and the ASV had a similarity between 97 and 100%. In contrast, a genus-level annotation was assigned between 95–98% identity while family annotation was between 90–95% identity (Callahan *et al*., 2016).

2.2.9.3. Authentication Analysis

The labelling of the vegetarian products was considered inaccurate if any ASV was present in relative abundance higher than 0.02 (2%) of the total reads present. ASVs that had a relatively low abundance (<0.02) but were present in all 3 independent PCRs, were retained for further analysis. Low abundance ASVs were excluded if they were only present in one or two of the PCR replicates.

2.3. Results and Discussion

2.3.1. Chloroform/Isoamyl alcohol DNA extraction performance

In this study, Chloroform/Isoamyl alcohol was used as the preferred DNA extraction method due to its ability to effectively break down the polysaccharide sugars present in the plant cell wall (Matlock, 2019). DNA quantity and quality were evaluated using spectrophotometric and fluorometric techniques; **Supplementary Table 2.3** provides an overview of concentrations obtained using Qubit fluorescence quantitation and Nanodrop absorbance measurements, as well as the purity obtained using Nanodrop Spectrometry. The Qubit quantitation analysis showed that the samples tested achieved a DNA concentration within the range of 32.4 – 394 ng/µL, with an average of 185.3 (99.2 standard deviation). In contrast, the Nanodrop spectrophotometric measurements showed that the samples tested achieved a DNA concentration within the range of 456.3 – 5814 ng/µL, with an average of 1627 ng/µL (1090.2 standard deviation). DNA extraction could provide enough DNA, significantly above the 10 ng/µL input, for PCR amplification for each sample. The noticeable difference in DNA concentrations between the Qubit and Nanodrop is largely due to the UV absorbances' inability to distinguish between DNA, RNA and protein (Koetsier *et al*., 2019). Moreover, Nanodrop Spectrometry quantitative values are easily influenced by various contaminants (free nucleotides, organic compounds and salts) and the sensitivity is often inadequate, especially at lower concentrations (Koetsier *et al*., 2019). These factors can account for the large discrepancy between the two methods, as Qubit dsDNA HS Assay kit is highly selective for double stranded DNA only, providing a more accurate and reliable concentration value for the calculation of DNA input for further downstream reactions (Koetsier *et al*., 2019).

Nanodrop measurements were used as an indication of the DNA extract's purity. A260/280 absorbance values higher than 1.78 but below 2 (indicator of pure DNA) were detected in 18 (57%) out of the 32 sample extracts, indicating the presence of DNA (Matlock, 2019). The rest of the samples, 14 out of the 32 samples, had A260/280 absorbance values higher than 2, indicating the presence of RNA and proteins that may have been co-extracted at some point during the extraction protocol (Matlock, 2019). The secondary measure of nucleic acid purity, 260/230 were in the range of 0.5 – 1.79. The expected values for pure nucleic acid are within the range of 2 – 2.2, with higher values (2.3 – 2.4) commonly accepted as pure dsDNA in buffered solutions, while lower values between 2.1 – 2.2 are usually associated with pure RNA. 260/230 ratios lower than 2.0 were detected for all sample extracts, suggesting a significant presence of contaminants within the DNA extracts. Contaminants that absorb at 230nm outnumbered those absorbing at 280 nm, which could suggest the presence of significant carbohydrate carryover (a unique problem with plant-based samples, residual phenols, EDTA or proteins and polysaccharides carried over from the food ingredients).

The purity of the DNA extracts were deemed acceptable for our study. The reliability and selectivity of the protocol provided reassurance that the DNA extracts obtained were suitable for further downstream reactions.

2.3.2. DNA metabarcoding performance overview

A total of 4 marker datasets were created from the sequencing of amplified PCR products obtained with the primers in **Table 2.6**: (i) P_trnL (trnL-g and trnL–h primers targeting plants), (ii) F_ITS2 (gITS7 and ITS4 primers targeting fungi), (iii) P_ITS2 (S2F and ITS4 primers targeting fungi), (iv) M_16S (16Smam1 and 16Smam2 targeting mammals) and Cyt_B (L14816 and H15173).

We obtained approximately 484 563 reads from the sequencing process in total for the study samples. P_trnL accounted for the majority of the raw sequencing reads at 355 132 (73.29% of the sequencing reads). F_ITS2 and P_ITS2 followed at 74 324 (15.34%) and 52 552 (14.80%) respectively. An exceptionally low number of reads was obtained for the M_16S dataset at 2 555 (0.007%) in the food samples, however 305 294 reads were found in the positive controls (meat ingredients only) used in the study. Additionally, Cyt_B provided 0 reads in the samples, however inspection of the positive controls that were prepared according to **Table 2.7** above showed that a substantial number of sequencing reads (approximately 3639 reads) were collected when the marker was exposed to pure meat samples, providing confirmation of the primer sets' efficacy when exposed to its appropriate target DNA.

Figure 2.1 below shows an overview of the relative abundance for each primer set reads in the samples, which provides a general overview of how well each primer set performed in the multiplex PCR reaction. Approximately half of the samples: F_237, F_238, F_241, F_242, F_244, F_245, F_246, F_247. F_248, F_255, F_259, F_264, F_265, F_266¸ F_267, F_268 and F_269 (17 samples out of the 31 tested) contained primarily sequencing reads that were obtained from P_trnL (more than 50% of the relative abundance). On the other hand, 10 of the 31 samples (F_239, F_243, F_249, F_250, F_253, F_256, F_257, F_260, F_261 and F_262) contained sequencing reads primarily obtained from F_ITS2 after filtering out "Viridiplantae" or plant-based data from further analysis.

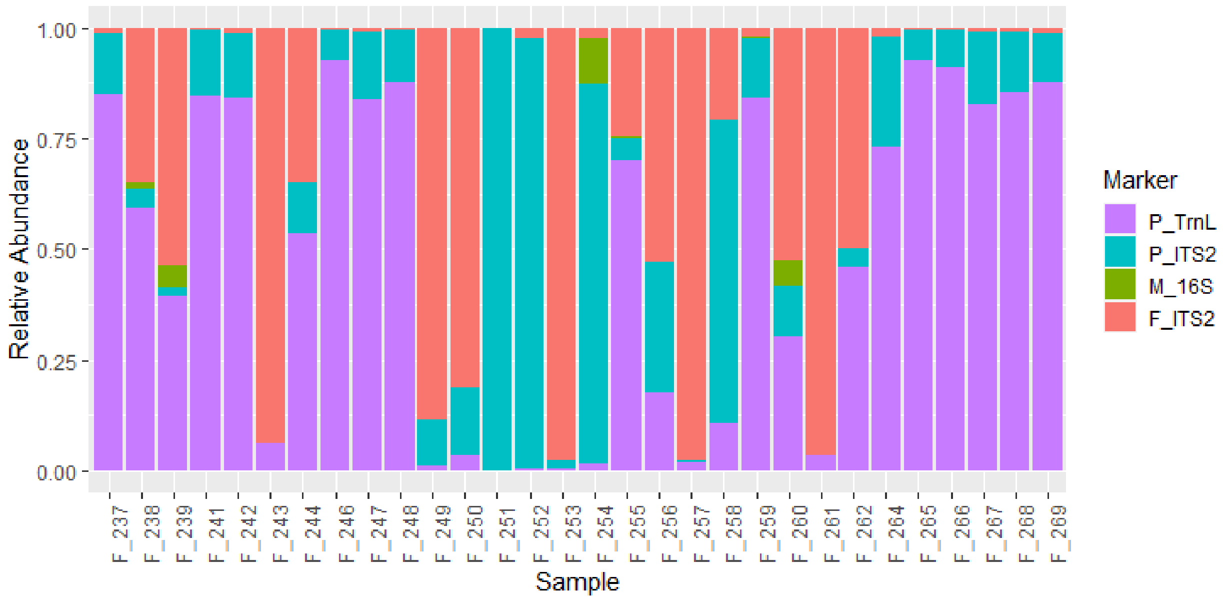

Figure 2. 1: The relative abundance of each primer set (y-axis) obtained for each vegetarian sample. The x-axis shows the sample code allocated to each sample. The plot indicates the relative abundance of sequencing reads obtained for each primer. The legend indicates the appropriate bar indicating the relative abundance for each primer set.

Figure 2.1 shows that the sequencing depth (relative abundance of reads assigned to the markers for each sample) between markers targeting plants and fungi were very unbalanced.

This high number of reads for P_trnL and F_ITS2 can be considered an expected outcome of the study as all the samples tested were primarily vegetarian and a large portion of the ingredients originated from plants and herbs. Additionally, a greater concentration of the P_trnL primer was inserted into the multiplex reaction during the PCR procedure, which may have caused some of the over-representation of P_trnL sequencing reads when directly compared to the other primer sets for each sample (Taberlet *et al*., 2007). Another possible reason behind the over-representation could be that the *taq* polymerase enzyme used in the PCR amplification step may have had a primer bias towards the P_trnL primers over the other primers present in the multiplex (Eisenstein, 2018). Some polymerases have a higher affinity for certain primers and will therefore prefer to bind with them when they are exposed to them in a complex mixture (Eisenstein, 2018).

Some of the samples containing F_ITS2 in higher relative abundance compared to other markers. Samples with ID codes: F_239, F_243, F_261 and F_262 all contained varying percentages of mycoprotein in their ingredient composition which could have contributed to the high representation of F_ITS2 for those samples. Mycoprotein is a form of single-celled protein or fungal protein that is specifically produced for human consumption and acts as a viable vegetarian alternative to a traditional meat protein source. Therefore, the higher relative abundance of F_ITS2 is expected for the samples stated above as the ingredients list indicated on the packaging that these products had a higher mycoprotein percentage. On the other hand, the remaining samples (F_249, F_250, F_253, F_256, F_257 and F_260) had heterogenous compositions that gave no indication of any fungal ingredients according to the labelling presented on the packaging. This can be due to F_ITS2 and P_ITS2 sharing a reverse primer. This close relation between the F_ITS2 and P_ITS2 primer sets can make the equal amplification of desired DNA difficult, with one target region being overamplified over the other. This challenge can be remedied by amplifying P_ITS2 and F_ITS2 separately in different balanced multiplexes or it can be achieved by amplifying the target DNA in singleplex.

Low number of sequencing reads for M_16S and CytB was expected and it is considered a positive result as the libraries mainly consisted of purely vegetarian samples, with the exception of few positive controls of animal origin to test the efficiency of the primer sets in a multiplex setting.

Ideally, this study aimed to create a multiplex that would be reproducible, robust and generate sequencing reads that provide the same relative abundance across all the plant-based markers (P_TrnL, P_ITS2 and F_ITS2) and animal-based DNA (M_16S and CytB) if applicable. Additionally, it was expected that there would be lower sequencing reads present for M_16S and

CytB as there were no animal-based DNA present in the study sample. As previously stated, **Figure 2.1** shows the relative abundance of sequencing reads generated by each primer set for the vegetarian sample set, and that the relative abundance was not equally balanced across the different samples tested. This is an unfortunate but not uncommon disadvantage of multiplex PCR building, as the optimization of the technique poses several difficulties, which include poor sensitivity and specificity and/or preferential amplification of certain specific target DNA (Elnifro *et al*., 2000). Preferential amplification of one target sequence over another (bias in template-to-product ratios) is a known anomaly in multiplex PCRs that are designed to amplify more than one target simultaneously (Elnifro *et al*., 2000). This is apparent as trnL generally provided more sequencing reads compared to its P_ITS2 counterpart. Due to several experimental studies and theoretical modelling techniques, there are two known classes of processes that induce this kind of bias, PCR drift and PCR selection (Elnifro *et al*., 2000). PCR drift is the stochastic fluctuation of the interaction of PCR reagents that occurs particularly in the early cycles of the reaction. On the other hand, PCR selection is a mechanism that specifically favours the generation and amplification of certain target templates due to the properties of the target DNA, their flanking sequences or it's entire genome (Elnifro *et al*., 2000). The addition of more than one primer pair increases the chance of obtaining spurious amplification products that may be amplified more efficiently than the desired target. This undesirable amplification not only consumes necessary reaction components, but it also produces impaired rates of annealing and extension of target DNA (Elnifro *et al*., 2000).

The trial-and-error approach was used to evaluate the suitability of the performance of the chosen primer pairs when present in a multiplex reaction. As seen above in **Figure 2.1**, the multiplex reaction did not perform as expected, as a successful multiplex reaction would generate sequencing reads with a balanced primer efficacy across all primer sets. Extensive preparation and optimization assays are crucial for the success of balanced primer efficacy in multiplex reactions (Zhang *et al*., 2018). Alternatively, the problems or biases experienced in this metabarcoding study can usually be obviated by not using multiplexes or building sequencing libraries that contain products from different markers in equimolar concentration.

2.3.3. Undisclosed species detected using universal primers

2.3.3.1. trnL and P_ITS2 marker

Figure 2.2 provides a visual overview of the species composition in relative abundance obtained using the trnL primer set. Each plot indicates the species representation for that individual sample and are differentiated by their unique sample code. Samples F_261 and F_263 were removed from further analysis for the trnL primer set as they did not have ASVs that were resolved to species level.

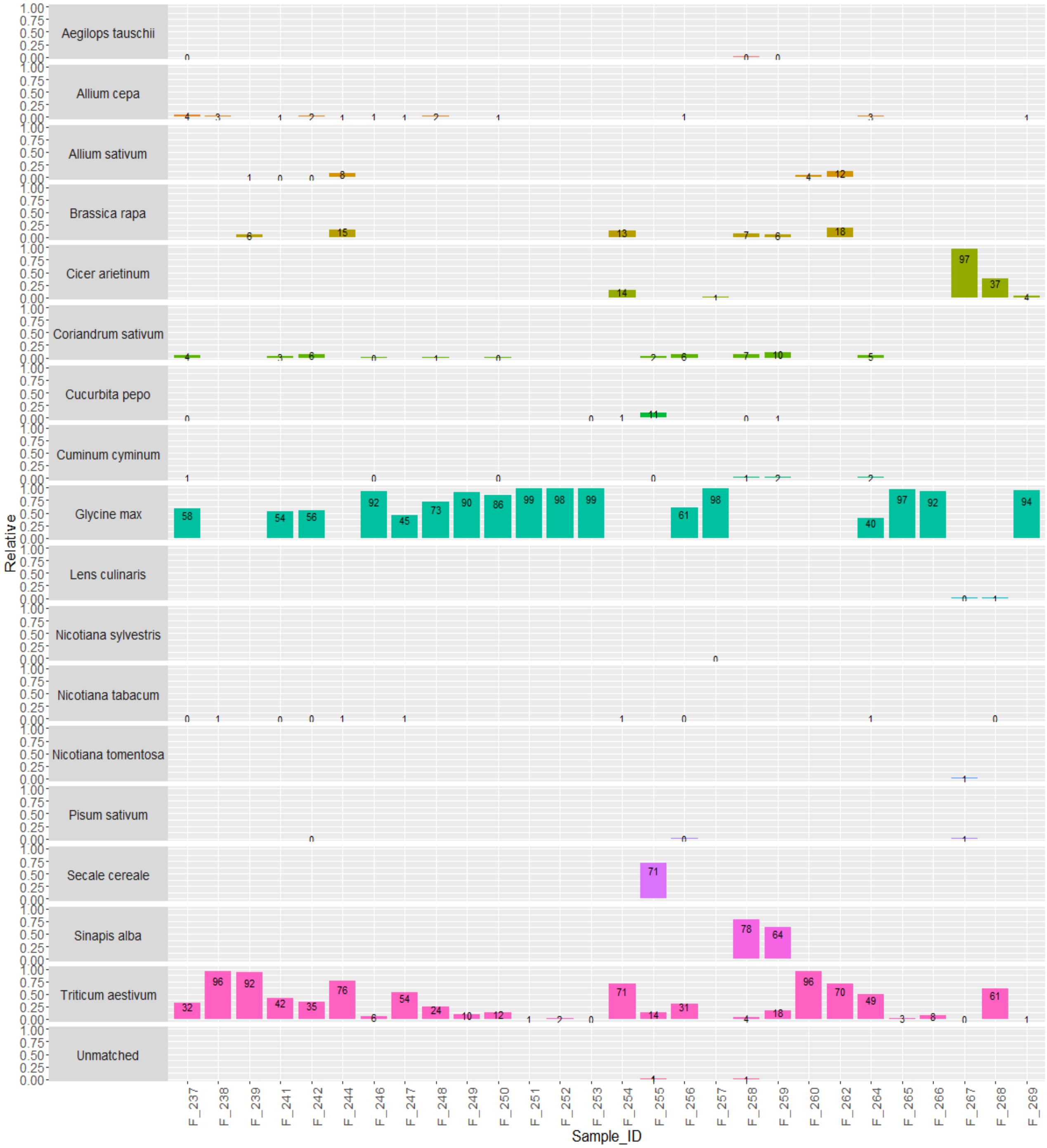

Figure 2. 2: The relative abundance of each species for each sample represented by the trnL primer set. F_2** represents the sample code allocated to each sample for discretion. The plot indicates the relative abundance of the sequencing reads depicting species information for each sample with the sequencing reads > 100 for the trnL primer set. The legend indicates the species represented, while the percentages show the relative abundance of the species within the sample when compared to other species present.

Figure 2.3 provides a visual overview of the species composition obtained using the P_ITS2 primer set. Each plot indicates the species representation for that individual sample and are differentiated by their unique sample code. Samples F_243, F_245, F_261 and F_263 were removed from further analysis for this primer set as they did not have ASVs that were assigned to species level.

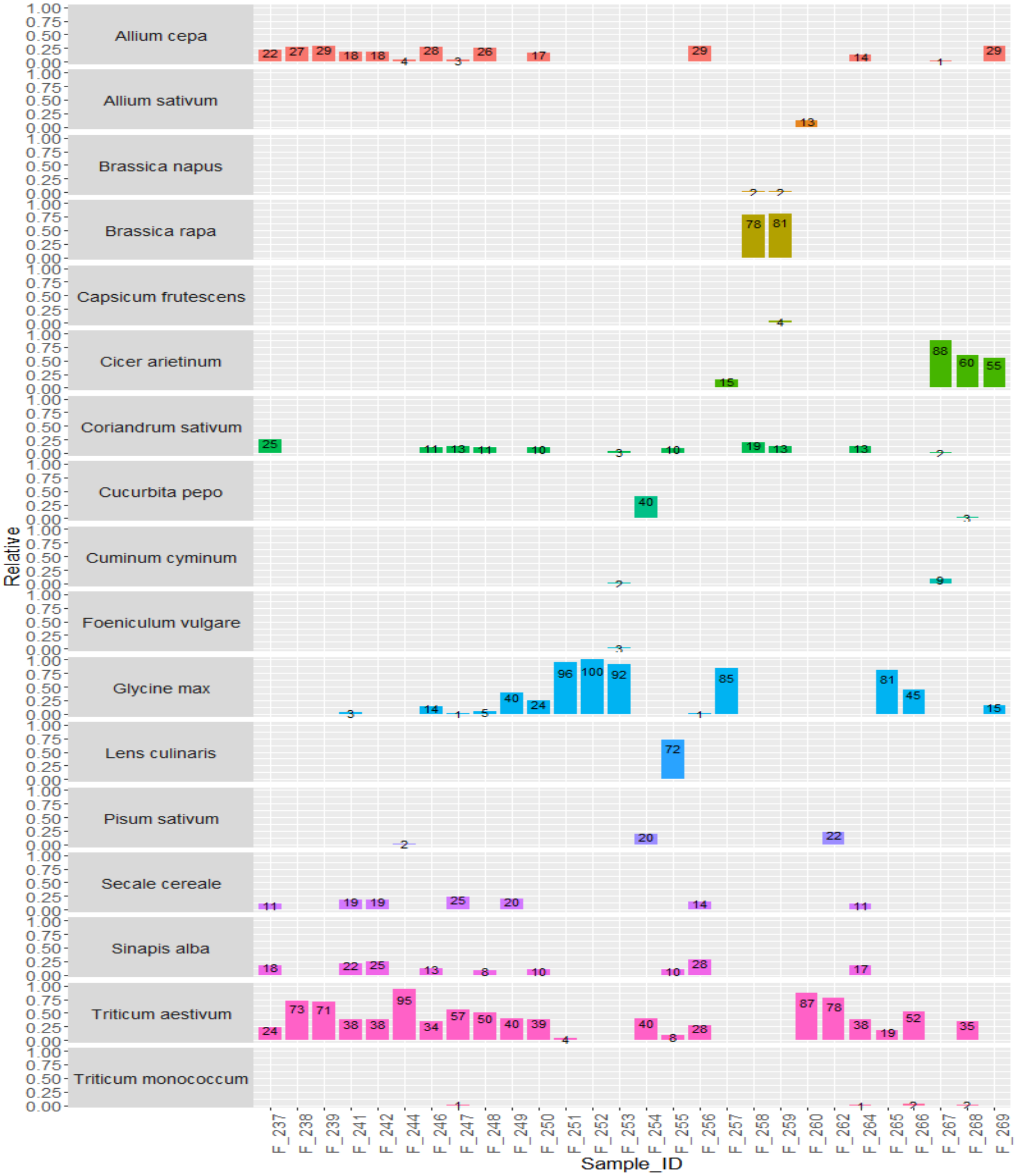

Figure 2. 3: The relative abundance of each species for each sample represented by the P_ITS2 primer set. F_2** represents the sample code allocated to each sample for discretion. The plot indicates the relative abundance of the sequencing reads depicting species information for each sample with the sequencing reads > 100 for the trnL primer set. The legend indicates the species represented, while the percentages show the relative abundance of the species within the sample when compared to other species present.

Supplementary Table 2.2 indicates the sequencing reads of species that were not explicitly labelled on the packaging for the P_ITS2 marker tested. Ingredients such as *Cuminum cyminum* (cumin) found in F_253, *Coriandrum sativum* (coriander) in samples F_237, F_246, F_247, F_248, F_250, F_253 and F_264, *Sinapis alba* (mustard) in samples F_255.

Cumin, coriander and mustard ingredients that were detected by the markers above can be grouped into the blanket category known as “herbs and spices” and a large majority of manufacturing companies do not explicitly list these ingredients in order to hide a proprietary product formula. Government bodies have different legislation and criteria regarding what is considered safe with regards to food, and food manufacturers are required to follow the rules and regulations laid out by their state. The Food and Drug Administration (FDA) has a set of rules and regulations regarding when and how herbs and spices should be extensively listed to maintain consumer safety. The FDA’s criteria for an ingredient to be considered a spice is as follows: “Any aromatic vegetable substance in the whole, broken or ground form and it primarily functions as seasoning rather than nutritional” (*CFR - Code of Federal Regulations Title 21*, 2020). However, there are exceptions to this rule, which includes any ingredients that is traditionally thought of as food such as garlic, onion, celery and any ingredient that is derived from fruit, vegetables, meat and fish that are typically regarded as food rather than flavouring. These ingredients must be declared by their common name regardless of its form (processed, ground, granulated etc.) According to the above criteria and curated list provided by the FDA (*CFR - Code of Federal Regulations Title 21*, 2020) regarding GRAS spices (Generally Regarded As Safe) cumin, coriander and mustard are regarded as safe for consumption by the consumer in small quantities and are not required to be listed on product packaging. On the other hand, the Foodstuffs, Cosmetics and Disinfectants Act, 1972 (Act of. 54 of 1972) of South Africa states that all herbs and spices not exceeding 2% by mass either singly or in combination, are required to be listed at the end of the ingredient labelling list while mass exceeding 2% can be listed at any point on the product label (https://www.westerncape.gov.za/text/2016/August/regulations__relating_to_the_labelling_and_advertising_of_foodstuffs_-_r_1055_of_2002).

With the above FDA rules and regulations in mind, the following ingredients were detected and not listed on the packaging for the following samples detected by the trnL marker: *Allium cepa* (onions) in F_237, F_241, F_242, F_247, F_256, *Cucurbita pepo* (common pumpkin) in F_237, F_253, F_255, F_258 and F_259, *Allium sativum* (garlic) in F_239, *Pisum sativum* (pea) in F_242, F_256, F_257 and F_267, *Lens culinaris* (lentils) in F_267, *Cicer arietinum* (chickpea) in F_257 and with known allergens listed as follows: *Triticum aestivum* (wheat) in F_253, F_258, F_259, F_267 and F_269 and *Secale cereale* (barley) in F_255. The P_ITS2 primer identified the following ingredients as “pure ingredients” and not listed on the packaging for the following samples: *Allium cepa* (onions) in F_237, F_241, F_242, F_247, F_256, *Foeniculum vulgare* (fennel) in F_253, *Pisum sativum* (pea) in F_254 and F_262, *Brassica napus* (rapeseed) in F_258 and F_259, *Cicer arietinum* (chickpea) in F_257 and with known allergens listed as follows: *Triticum monococcum* (einkorn wheat) in F_247, F_264, F_266 and F_268 and *Secale cereale* (barley) in F_237, F_241, F_242, F_247, F_249, F_256 and F_266.

These food ingredients are not considered herbs or spices and can not fall under the category of flavourings. According to the Foodstuffs, Cosmetics and Disinfectants Act and the US FDA regulations, these ingredients should have been listed on their respective packaging. Although the above (except for wheat and barley) are not considered high risk allergens, in rare cases, these ingredients could still incite an allergic reaction that could cause symptoms such as hives, skin inflammation, shortness of breath/wheezing and even death in severe cases (Skypala, 2019). Stricter regulations and legislations are being enforced around the world due to the increase in cases of severe allergic reactions due to tolerances developed over time.

2.3.3.2. Case study on food anaphylaxis

A study conducted in the US reported 1500 cases of rare food anaphylaxis in a year, of which 1% of the cases that were reported were fatal (Skypala, 2019). More serious allergens such as wheat and barley can cause anaphylactic shock, which is the rapid drop in blood pressure that causes airways to narrow, obstructing normal breathing and if left untreated, could lead to death by suffocation. This drop in blood pressure is directly related to the body’s immune response towards foreign allergens, which are small amounts of organic protein naturally found in food that the immune system detects as a threat (Skypala, 2019). This usually rare occurrence has become increasingly more common as more and more people are developing allergies towards food, so companies and manufacturers must take better care to label their products correctly (Wheeler, 2016). Surprisingly, this major problem that is prevalent in the food industry isn't inherently from the deliberate inclusion of allergens but rather that they find their way into products through cross-contamination during the manufacturing process (Wheeler, 2016). Thankfully, it is possible to build an Allergen Control Programme into existing food safety systems such as Hazard Analysis Critical Control Point (HACCP) in order to identify and control allergens along with other food safety hazards (Wheeler, 2016).

An Allergen Control Policy is a legal requirement under Regulation R 146 of the Foodstuffs, Cosmetics and Disinfectants Act in South Africa which states that for all food processors or manufacturers where there is a risk of cross contamination, they are required to participate in due diligence or Good Manufacturing Practices in order to avoid or reduce the incidences of cross contamination (Wheeler, 2016). Common measures that are used to prevent cross contamination include: 1.) designing the facilities and equipment in a way that streamlines efficient cleaning practices, 2.) implementing a vendor control programme that assesses and regulates the risk of contamination of raw materials prior to receipt, 3.) dedication of certain equipment for allergen and non-allergen containing products (Wheeler, 2016). Other practical steps towards allergen regulation involve allergen residue testing of food contact surfaces and of finished products. Both in-house test kits together with external laboratory testing facilities are widely available across South Africa and can be utilized to prevent the instances of cross contamination in food processing systems (Wheeler, 2016).

If an unprecedented number of cases of food allergen reactions are reported for a particular food product, mass recalls are implemented. These recalls are not limited to smaller, more easily regulated manufacturers who have fewer resources (Wheeler, 2016). The biggest and best manufacturers have fallen victim to allergen cross contamination and incorrect packaging on products that leads to inaccurate ingredient and allergen information presented to the consumer. Although this is a common practice amongst first-world countries, this problem has received more prominence in South Africa, with the recent recall of an undisclosed muffin mix suspected to have contained undeclared nuts (Wheeler, 2016).

2.3.4. Successfully resolved species detected using the trnL and P_ITS2 marker

Supplementary Table 2.1 indicates the matches between the species resolution obtained from the sequencing reads of the trnL primer set and the list of ingredients that were provided by the packaging. **Supplementary Table 2.2** indicates similar data for the P_ITS2 primer set. The ingredient list presented in those tables includes only those ingredients that could be identified using their respective barcode (fungal or meat derived ingredients were excluded). Ingredients highlighted in green indicated a positive match (sequencing results align with ingredient composition on packaging). ASVs that constituted less than 2% of the total relative abundance were removed. All sequencing reads present were present in 3 independent PCR's and the number of reads collapsed into a final count. Overall, the trnL primer set was able to successfully detect 29.72% of the ingredients listed on the packaging while P_ITS2 primer set was able to successfully detect 32.24% of the ingredients listed on the packaging of our study group.

Some smaller mass quantity ingredients such as *Allium cepa* (onion), *Allium sativum* (garlic), *Coriandrum sativum* (coriander), *Cucurbita pepo* (common pumpkin), *Cicer arietinum* (chickpea), *Sinapis alba* (mustard) and *Cuminum cyminum* (cumin seeds) were detected in the appropriate products that indicated their presence on the packaging, demonstrating that trnL may resolve certain species accurately when exposed to heterogenous mixtures of unequal mass. Similarly, smaller mass quantity ingredients such as *Allium cepa* (onion), *Allium sativum* (garlic), *Coriandrum sativum* (coriander), *Cucurbita pepo* (common pumpkin), *Cicer arietinum* (chickpea), *Sinapis alba* (mustard), *Cuminum cyminum* (cumin seeds), *Triticum aestivum* (wheat), *Pisum sativum* (pea), *Glycine max* (soya), *Lens culinaris* (lentil) and *Capsicum frutescens* (chilli) were detected in the appropriate products that indicated their presence on the packaging, demonstrating that P_ITS2 shares the same resolution ability .

The relatively low resolution is not completely unexpected, as trnL is widely known to have poor resolution capabilities due to its length and non-variability (de Groot *et al*., 2011). However, the regions extensive reference database and highly conserved primers ensure trnL's relevance for the foreseeable future. The primer set was able to successfully detect all instances in which soya and wheat were present according to the ingredient composition for all product samples. These ingredients contributed the largest mass proportion to the products, as they provide a suitable base for all vegetarian meals and are largely accepted as an appropriate substitute for meat products by vegetarians. Some smaller mass quantity ingredients such as *Allium cepa* (onion), *Allium sativum* (garlic), *Coriandrum sativum* (coriander), *Cucurbita pepo* (common pumpkin), *Cicer arietinum* (chickpea), *Sinapis alba* (mustard) and *Cuminum cyminum* (cumin seeds) were detected in the appropriate products that indicated their presence on the packaging, demonstrating that trnL may resolve certain species accurately when exposed to heterogenous mixtures of unequal mass.

On the other hand, the 32.24% successful species resolution rate of P_ITS2 is better than trnL's resolution ability but is still not sufficient to be considered a successful optimization. Many successful multiplex PCR studies rely on preliminary assays for optimization of each of the markers within the PCR pool (Sint *et al*., 2012). These optimizations involve 1.) developing adequate concentrations of reagents and target sequences to balance sequence representation and 2.) eliminate non-specific primers by developing and designing primers that are widely different from one another to prevent marker cross-contamination (Sint *et al*., 2012). With regards to this study, further optimizations need to be implemented to develop a more balanced primer efficiency to improve the overall species resolution success of the markers utilized.

2.3.5. Traces of Tobacco detected using trnL

Surprisingly, trnL was able to resolve ASV's relating to tobacco plant species in the products as well. **Supplementary Table 2.1** shows various tobacco plant species present within the vegetarian food products: *Nicotiana tabacum* (tobacco) in F_237, F_238, F_241, F_242, F_244, F_254, F_256, F_264 and F_268, *Nicotiana sylvestris* (woodland tobacco) in F_257 and *Nicotiana tomentosa* (flowering tobacco) in F_267. A grass variant *Aegilops tauschii,* known as Tausch's goat grass was also found in samples F_237, F_258 and F_259. *Nicotiana tabacum* is a common constituent of cigarettes while *Nicotiana sylvestris* is often added to the cigarette mixture almost exclusively for its woody scent (Fatica *et al*., 2019). *Nicotiana tomentosa* is primarily cultivated in Peru and is often used as a substitute for *Nicotiana tabacum* in cigarettes (Fatica *et al*., 2019). *Nicotiana tabacum, Nicotiana sylvestris* and *Nicotiana tomentosa* were detected in trace amounts and may have been present in the food samples due to the cross-contamination of extraction samples for a different study.

During the Phenol-Chloroform extraction procedure, the vegetarian food products were co-extracted with various cigarette brands, so cross-contamination of trace amounts of DNA (as well as other reaction components) between these samples may have occurred during this step. Precautions need to be implemented in order to reduce the instance of cross-contamination among the samples by 1.) extracting different sample groups on separate days, 2.) ensuring that pipette tips are properly sterilized and are not shared amongst the different test tubes or 3.) ensure that test tubes are closed regularly between reagent mixture insertion and minimize the opening of test tubes as much as possible (Griffith, 2016). However, the detection of these tobacco plant DNA strains validates the trnL primer sets' effectiveness in identifying and resolving species from highly degraded and processed DNA samples since cigarettes are thoroughly dried during production (Griffith, 2016).

In addition to the tobacco plant species detected in the samples, *Aegilops tauschii*, known as Tausch's goat grass was detected. Interestingly, the goat grass is known to be a progenitor of the widely known wheat used for bread. It is an undomesticated plant and is considered a high quality and yield product. However, due to its large size and complexity of the genome, researchers are required to implement ordered-clone sequencing strategies in order to generate a high-quality sequencing draft for replication. Although this may have been a sequencing derivative of common wheat, it is interesting to discover traces of ancient DNA of a common ingredient in the multiplex.

2.3.6. Identification of fungal species using F_ITS2

According to the data represented in **Figure 2.4**, it is apparent that the sequencing reads were primarily allocated and classified to Kingdom Plantae and no known fungal species was found using F_ITS2 primer set.

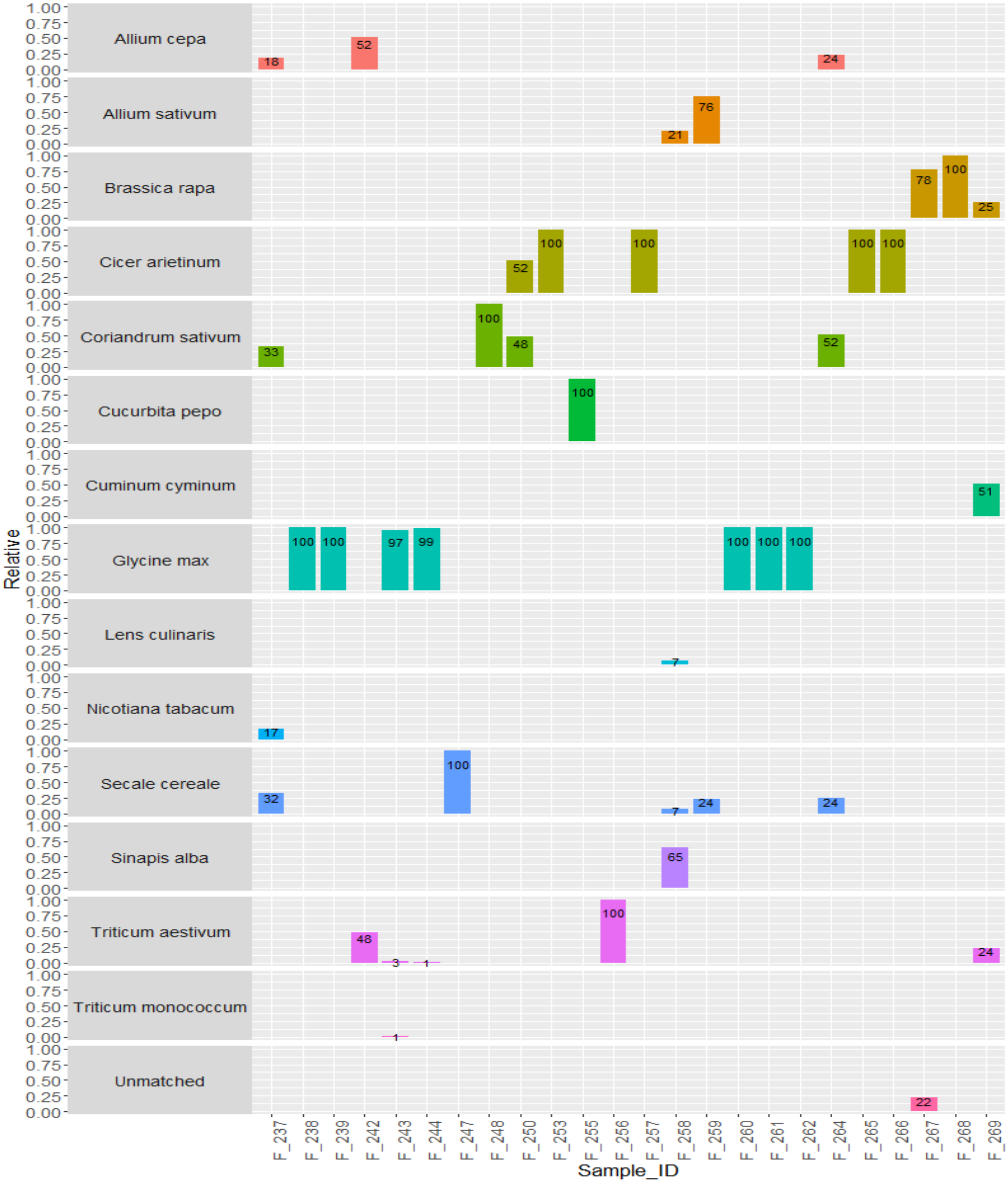

Figure 2. 4: The relative abundance of each species for each sample represented by the F_ITS2 primer set. F_2** represents the sample code allocated to each sample for discretion. The plot indicates the relative abundance of the sequencing reads depicting species information for each sample with the sequencing reads > 100 for the FITS2 primer set. The legend indicates the species represented, while the percentages show the relative abundance of the species within the sample when compared to other species present.

Amplification of untargeted DNA is a common problem metabarcoding studies. This problem may have been amplified by the fact that the primers used to target plants and fungi used an identical reverse universal primer for plant and fungi. It is apparent that this combination of primers shouldn’t be inserted into a multiplex mixture together, as either one could potentially be overshadowed by the other due to primer bias and possible competition, resulting in inaccurate results (Yao *et al*., 2010). Moreover, the primer sequence of the reverse primer was re-designed without any in silico or lab testing and the new sequence might privilege the amplification of plants and lose its potential for amplification of fungi. A better choice of primers/primer designing/testing should be conducted to ensure the viability and efficacy of the multiplexing approach proposed. Another possible solution would be to introduce novel barcoding regions that have more conserved priming sites, thereby limiting the sequencing errors present in this study. However, such novel markers can provide less taxonomic resolution, limiting the information garnered by the sequencing reads and preventing species-level analysis (Yao *et al*., 2010). Additionally, novel markers have substantially limited reference databases, making it difficult to infer species identity based on previously sequenced reads. Lastly, the lack of fungal representation indicated by the marker could suggest that no fungal DNA was present in the samples. However, further optimization assays or reproducing the experiment in singleplex would need to be conducted in order to confirm the absence of fungal DNA.

2.3.7. Identification of animal species using M_16S

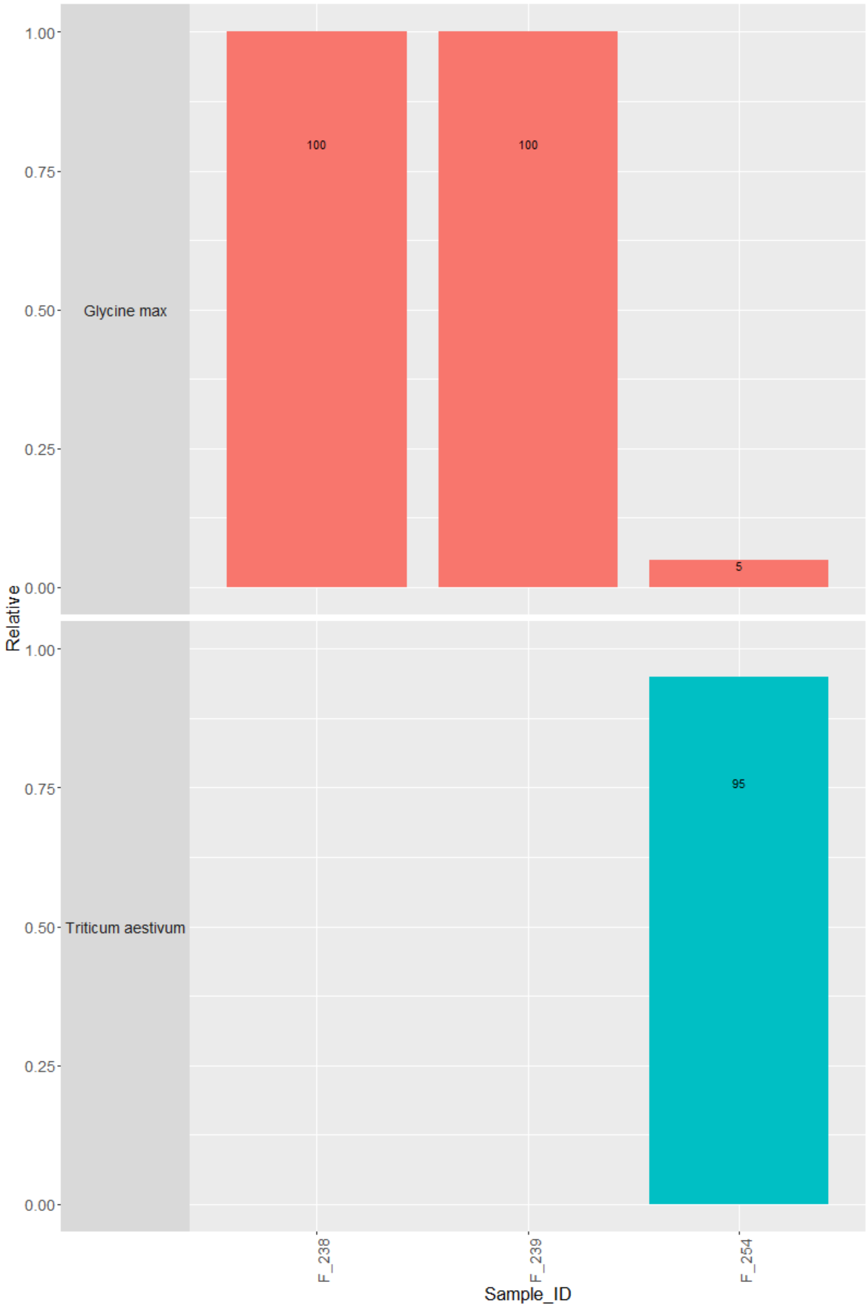

Figure 2. 5: The relative abundance of each species for each sample represented by the M_16S (mammal and bird 16S) primer set. F_2** represents the sample code allocated to each sample for discretion. The plot indicates the relative abundance of the sequencing reads depicting species information for each sample with the sequencing reads > 100 for the M_16S primer set. The legend indicates the species represented, while the percentages show the relative abundance of the species within the sample when compared to other species present.

Figure 2.5 provides a visual overview of the species composition obtained using the M_16S primer set for samples F_238, F_239 and F_254. The number of reads for each sample was 562, 1898 and 2034 respectively. The most abundant species resolution indicated by M16S primer set for F_238 and F_239 was *Glycine max* (soya) while the most abundant species in sample F_254 was *Triticum aestivum* (wheat). This unexpected representation of plant DNA could have been due to the homology between chloroplast and mitochondrial DNA (Pearson, 2013). Homology is when two sequences or structures share more similarity than would be expected by chance (Pearson, 2013). When this similarity is observed, the widely accepted explanation is that the two sequences did not arise independently and that they essentially arose from a common ancestor (Pearson, 2013). Due to the lack of mammal DNA present in the sample, the M16S primer set generated reads from the DNA of the over-represented soya and wheat present in the sample due to the similarity in structure and relation of the two target regions. This is a common challenge among metabarcoding studies and there is no consensus on how to deal with the issue of homology among target organelles (Pearson, 2013). Modifications of the homologous primer sets could be created to minimize the amplification of undesired DNA. Additionally, blocker primers can be designed to inhibit the amplification of certain problematic DNA in a complex sample for the detection of desired species (C. Liu *et al*., 2019). These blockers are known to preferentially bind to the DNA of which amplification is to be avoided. They are synthesized in the same way conventional amplification primers are, but they are modified with an additional C3 spacer at the 3' end, resulting in total inhibition of enzymatic elongation of the primer (C. Liu *et al*., 2019). Blocking primers can either compete directly with amplification primers (annealing-inhibiting blocking primers) or prevent elongation by binding onto the fragment in between the amplification primers (known as elongation-arrest blocking primers) (C. Liu *et al*., 2019).

Developing multiplexes that contain primer sets that target heterologous regions or blocker primers may minimize the instances of amplification of undesired DNA in the PCR step, however further analysis of those regions and the limitations that are associated with them will need to be extensively researched before performing preliminary assays. In conclusion, **Figure 2.5** did not indicate the presence of any animal species within the samples, even though labelling suggested the presence of milk and reconstituted egg whites in the ingredients listed in **Table 2.2.** This could indicate that the primer set isn’t effective and robust enough to detect the presence of highly degraded or minimal trace DNA within the majority-vegetarian sample set.

2.4. Conclusion

Overall, the plant, fungi and mammal DNA mini-barcodes were able to detect taxa at species level for some of the products listed on the packaging. Cytochrome B couldn’t detect any mammal DNA within the highly processed vegetarian samples, which was expected due to the nature of the product. The presence of sequencing reads relating to mammal DNA was detected in the meat-positive samples for Cytochrome B, validating its presence and effectiveness in the multiplex reaction. Mammal 16S was unable to detect animal DNA and primarily detected plant DNA at varying taxonomic levels. This phenomenon is not completely uncommon due to the Mammal 16S mini-barcodes homology with chloroplast DNA. Due to the over-representation of plant DNA in the sample, the Mammal 16S mini-barcode could have preferentially sequenced plant-based DNA because of its overwhelming presence in the samples.

Most of the reads obtained for Fungi ITS2 detected varying taxonomic levels relating to “Viridplantae” or plant-based DNA. This detection of non-specific plant DNA could have been due to the Plant ITS2 and the Fungal ITS2 mini-barcodes sharing a reverse primer. This similarity of the primer sets could have caused each mini-barcode to sequence both fungal and plant-based DNA. The trnL primer set was able to successfully detect 29.72% of the ingredients listed at species level while P_ITS2 primer set was able to successfully detect 32.24% of the ingredients listed on the packaging of our highly processed study group. Although the percentages detected for both trnL and P_ITS2 were relatively low, with further multiplex reaction testing, balancing and optimization, this percentage could increase with a better understanding of how well the primers work together in a multiplex and the further prevention of DNA metabarcoding limitations.

Surprisingly, trnL was able to detect species of plant-based DNA relating to tobacco. However, this detection could have been a result of cross-contamination between the study sample set and cigarette samples that were co-extracted. This discrepancy can be mitigated or eliminated by segregating the co-extraction procedure so that only one study sample DNA set is extracted at a time, preventing study sample cross-contamination. No other harmful adulterants were detected other than common ingredients that could insight mild food anaphylaxis symptoms such has mustard, wheat and barley. Further analysis should be conducted to understand the severity and quantity of DNA is present within the samples.

The results of this study further highlights the major limitations of this methodology with regards to highly processed standardized food control procedures. There are numerous experimental challenges such as 1.) PCR amplification biases, 2.) difficulty in the ability to assign species level annotation of sequences, 3.) the lack of existing standardized protocols that are both robust, balanced and sensitive as well as the overall 4.) cost and time implementation of the methodology. Regardless of the limitations that DNA metabarcoding currently faces, the constant innovation and improvement of new primers, decrease in overall sequencing costs and the ever-growing and expanding reference databases for sequence comparison, DNA metabarcoding may become the gold standard in the foreseeable future in rapid and cost-effective methodologies for authentication of highly processed food products and extensive quality monitoring in the food industry.

2.5. References

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution*, *9*(1), 134–147. https://doi.org/10.1111/2041-210X.12849

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., Costessi, A., Duijsings, D., Rechenmann, F., Gaspar, F. B., Barreto Crespo, M. T., Holst-Jensen, A., Birck, M., Burns, M., Haynes, E., Hochegger, R., Klingl, A., Lundberg, L., Natale, C., … Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience*, *6*(10), 1. https://doi.org/10.1093/gigascience/gix080

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

CFR - Code of Federal Regulations Title 21. (n.d.). Retrieved May 16, 2021, from https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfCFR/CFRSearch.cfm

Chauhan, A. (2020). *Food fraud – an evolving crime with profit at its heart - New Food Magazine*. https://www.newfoodmagazine.com/article/109059/food-fraud-an-evolving-crime-with-profit-at-its-heart/

Coghlan, M. L., Haile, J., Houston, J., Murray, D. C., White, N. E., Moolhuijzen, P., Bellgard, M. I., & 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. https://doi.org/10.1371/journal.pgen.1002657

Coissac, E. (2012). OligoTag: A program for designing sets of tags for next-generation sequencing of multiplexed samples. *Methods in Molecular Biology*, *888*, 13–31. https://doi.org/10.1007/978-1-61779-870-2_2

Coskun, O. (2016). Separation Tecniques: CHROMATOGRAPHY. *Northern Clinics of Istanbul*, *3*(2), 156. https://doi.org/10.14744/nci.2016.32757

D'Amato, M. E., Alechine, E., Cloete, K. W., Davison, S., & Corach, D. (2013). Where is the game? Wild meat products authentication in South Africa: a case study. *Investigative Genetics*, *4*(1), 6. https://doi.org/10.1186/2041-2223-4-6

Danezis, G. P., Tsagkaris, A. S., Camin, F., Brusic, V., & Georgiou, C. A. (2016). Food authentication: Techniques, trends & emerging approaches. *TrAC - Trends in Analytical Chemistry*, *85*(November), 123–132. https://doi.org/10.1016/j.trac.2016.02.026

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2017). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. In *bioRxiv*. bioRxiv. https://doi.org/10.1101/221499

Dopheide, A., Xie, D., Buckley, T. R., Drummond, A. J., & Newcomb, R. D. (2019). Impacts of DNA extraction and PCR on DNA metabarcoding estimates of soil biodiversity. *Methods in Ecology and Evolution*, *10*(1), 120–133. https://doi.org/10.1111/2041-210X.13086

Eisenstein, M. (2018). Microbiology: Making the best of PCR bias. *Nature Methods*, *15*(5), 317–320. https://doi.org/10.1038/nmeth.4683

Elnifro, E. M., Ashshi, A. M., Cooper, R. J., & Klapper, P. E. (2000). Multiplex PCR: Optimization and Application in Diagnostic Virology. *Clinical Microbiology Reviews*, *13*(4), 559. https://doi.org/10.1128/CMR.13.4.559-570.2000

Fatica, A., Di Lucia, F., Marino, S., Alvino, A., Zuin, M., De Feijter, H., Brandt, B., Tommasini, S., Fantuz, F., & Salimei, E. (2019). Study on analytical characteristics of Nicotiana tabacum L., cv. Solaris biomass for potential uses in nutrition and biomethane production. *Scientific Reports, 9*(1), 1–8. https://doi.org/10.1038/s41598-019-53237-8

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017a). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/s41467-017-01312-x

Frøslev, T. G., Kjøller, R., Bruun, H. H., Ejrnæs, R., Brunbjerg, A. K., Pietroni, C., & Hansen, A. J. (2017b). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*, *8*(1), 1–11. https://doi.org/10.1038/s41467-017-01312-x

Griffith, C. (2016). Surface Sampling and the Detection of Contamination. In *Handbook of Hygiene Control in the Food Industry: Second Edition* (pp. 673–696). Elsevier Inc. https://doi.org/10.1016/B978-0-08-100155-4.00044-3

Han, J., Zhu, Y., Chen, X., Liao, B., Yao, H., Song, J., Chen, S., & Meng, F. (2013). The Short ITS2 Sequence Serves as an Efficient Taxonomic Sequence Tag in Comparison with the Full-Length ITS. *BioMed Research International, 2013*, 1–7. https://doi.org/10.1155/2013/741476

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1512), 313–321. https://doi.org/10.1098/RSPB.2002.2218

Https://www.westerncape.gov.za/text/2016/August/regulations_-_relating_to_the_labelling_and_advertising_of_foodstuffs_-_r_1055_of_2002. (n.d.). *(No Title)*. Retrieved May 16, 2021, from https://www.westerncape.gov.za/text/2016/August/regulations_-_relating_to_the_labelling_and_advertising_of_foodstuffs_-_r_1055_of_2002.pdf

Johnston, R. (2018). *Vegetarian food | Smart Food Expo*. https://smartfoodexpo.ru/en/vegetarianskoe-pitanie

Kadri, K. (2020). Polymerase Chain Reaction (PCR): Principle and Applications. In *Synthetic Biology - New Interdisciplinary Science*. IntechOpen. https://doi.org/10.5772/intechopen.86491

Kang, T. S. (2019). Basic principles for developing real-time PCR methods used in food analysis: A review. In *Trends in Food Science and Technology* (Vol. 91, pp. 574–585). Elsevier Ltd. https://doi.org/10.1016/j.tifs.2019.07.037

Koen, N., Blaauw, R., & Wentzel-Viljoen, E. (2016). Food and nutrition labelling: the past, present and the way forward. *South African Journal of Clinical Nutrition*, *29*(1), 13–21. https://doi.org/10.1080/16070658.2016.1215876

Koetsier, G., Cantor, E., & Biolabs, E. (2019). *A Practical Guide to Analyzing Nucleic Acid Concentration and Purity with Microvolume Spectrophotometers*.

Kulski, J. K. (2016). Next-Generation Sequencing — An Overview of the History, Tools, and "Omic" Applications. In *Next Generation Sequencing - Advances, Applications and Challenges*. InTech. https://doi.org/10.5772/61964

Lau, J.-E. (2021). *Food Fraud - Part I - What Is Food Fraud? | Food Technology |Science Meets Food*. https://sciencemeetsfood.org/food-fraud-what-is-food-fraud/

Liu, C., Qi, R.-J., Jiang, J.-Z., Zhang, M.-Q., & Wang, J.-Y. (2019). Development of a Blocking Primer to Inhibit the PCR Amplification of the 18S rDNA Sequences of Litopenaeus vannamei and Its Efficacy in Crassostrea hongkongensis. *Frontiers in Microbiology*, *0*(APR), 830. https://doi.org/10.3389/FMICB.2019.00830

Liu, L., Wang, Y., He, P., Li, P., Lee, J., Soltis, D. E., & Fu, C. (2018). Chloroplast genome analyses and genomic resource development for epilithic sister genera Oresitrophe and Mukdenia (Saxifragaceae), using genome skimming data. *BMC Genomics*, *19*(1), 235. https://doi.org/10.1186/s12864-018-4633-x

Mafra, I., Ferreira, I. M. P. L. V. O., & Oliveira, M. B. P. P. (2008). Food authentication by PCR-based methods. *European Food Research and Technology*, *227*(3), 649–665. https://doi.org/10.1007/s00217-007-0782-x

Martin, M. (2011). *Cutadapt removes adapter sequences from high-throughput sequencing reads | Martin | EMBnet.journal*. EMBnet.Journal 17, 10. https://journal.embnet.org/index.php/embnetjournal/article/view/200/479

Matlock, B. (2019). *Assessment of Nucleic Acid Purity*. www.thermoscientific.com

Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., & Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: The polymerase chain reaction. *Cold Spring Harbor Symposia on Quantitative Biology*, *51*(1), 263–273. https://doi.org/10.1101/sqb.1986.051.01.032

Ng, J., Satkoski, J., Premasuthan, A., & Kanthaswamy, S. (2014). A nuclear DNA-based species determination and DNA quantification assay for common poultry species. *Journal of Food Science and Technology*, *51*(12), 4060–4065. https://doi.org/10.1007/s13197-012-0893-7

Pearson, W. R. (2013). An introduction to sequence similarity (“homology”) searching. *Current Protocols in Bioinformatics*, *0 3*(SUPPL.42). https://doi.org/10.1002/0471250953.bi0301s42

Piper, A. M., Batovska, J., Cogan, N. O. I., Weiss, J., Cunningham, J. P., Rodoni, B. C., & Blacket, M. J. (2019). *Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance*. *8*, 1–22. https://doi.org/10.1093/gigascience/giz092

Prado, M., Ortea, I., Vial, S., Rivas, J., Calo-Mata, P., & Barros-Velázquez, & J. (2016). Advanced DNA-and Protein-based Methods for the Detection and Investigation of Food Allergens. *Critical Reviews in Food Science and Nutrition*, *56*, 2511–2542. https://doi.org/10.1080/10408398.2013.873767

Schon, E. A., DiMauro, S., & Hirano, M. (2012). Human mitochondrial DNA: roles of inherited and somatic mutations. *Nature Reviews. Genetics*, *13*(12), 878–890. https://doi.org/10.1038/nrg3275

Simpson, F. (2018, June 9). *Sainsbury's, Tesco Investigated After Meat Traces Reported in Vegetarian Food*. https://www.businessinsider.com/sainsburys-tesco-investigated-after-meat-traces-reported-in-vegetarian-food-2018-6?IR=T

Sint, D., Raso, L., & Traugott, M. (2012). Advances in multiplex PCR: Balancing primer efficiencies and improving detection success. *Methods in Ecology and Evolution*, *3*(5), 898–905. https://doi.org/10.1111/j.2041-210X.2012.00215.x

Skypala, I. J. (2019). Food-induced anaphylaxis: Role of hidden allergens and cofactors. In *Frontiers in Immunology* (Vol. 10, Issue APR, p. 673). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2019.00673

Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., Prins, T. W., & Kok, E. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry*, *408*(17), 4615–4630. https://doi.org/10.1007/s00216-016-9595-8

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. https://doi.org/10.1111/j.1365-294X.2012.05470.x

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. https://doi.org/10.1093/nar/gkl938

UNITE - Resources. (n.d.). Retrieved October 10, 2020, from https://unite.ut.ee/repository.php

Wadapurkar, R. M., & Vyas, R. (2018). Computational analysis of next generation sequencing data and its applications in clinical oncology. *Informatics in Medicine Unlocked*, *11*, 75–82. https://doi.org/10.1016/J.IMU.2018.05.003

Wheeler, R. (2016). *Management of allergens in the food industry – still a hot topic*. https://www.foodfocus.co.za/home/Industry-Topics/food-safety/Management-of-allergens-in-the-food-industry

Yao, H., Song, J., Liu, C., Luo, K., Han, J., Li, Y., Pang, X., Xu, H., Zhu, Y., Xiao, P., & Chen, S. (2010). Use of ITS2 region as the universal DNA barcode for plants and animals. *PLoS ONE*, *5*(10). https://doi.org/10.1371/journal.pone.0013102

Zhang, G. K., Chain, F. J. J., Abbott, C. L., & Cristescu, M. E. (2018). Metabarcoding using multiplexed markers increases species detection in complex zooplankton communities. *Evolutionary Applications*, *11*(10), 1901–1914. https://doi.org/10.1111/eva.12694

2.6. Supplementary Table

Supplementary Table 2. 1: Comparison of the trnL sequencing results to the ingredient composition listed on the label. The ingredient list includes only those that could be identified using this barcode (fungal or meat derived ingredients were excluded). Ingredients highlighted in **green** indicate a positive match (sequencing results align with ingredient composition on indicated on the packaging). Ingredients in black indicate their presence on the packaging but were not detected by the primers used in this study. ASVs that constituted less than 2% of the total relative abundance were removed. The ASVs were present in all three independent PCRs. ASVs showing sequencing results that are not included in the ingredient composition listed on the label are highlighted in **red** including the percent identity (%ID), read abundance and common name).

ID	Listed Ingredients	Observed Species	Abundance	%ID	Other Species Identified	Common Name	Abundance	%ID
F_237	Chillies	-	-	-	***Allium cepa***	**Onion**	**2064**	**100**
	Garlic	-	-	-	***Coriandrum sativum***	**Coriander**	**1917**	**100**
	Maize	-	-	-	***Cucurbita pepo***	**Butternut**	**151**	**99.7**
	Mustard	-	-	-	***Cuminum cyminum***	**Cumin**	**434**	**100**
	Potato	-	-	-	***Aegilops tauschii***	**Tausch's Goatgrass**	**135**	**98.5**
	Wheat	*Triticum aestivum*	15071	100	***Nicotiana tabacum***	**Tobacco**	**170**	**99.7**
	Soya	*Glycine max*	27731	100				
F_238	**Wheat**	*Triticum aestivum*	22351	100	***Nicotiana tabacum***	**Tobacco**	**157**	**99.7**
	Onion	*Allium cepa*	731	100				
	Barley	-	-	-				
	Rape Seed	-	-	-				
F_239	Barley	-	-	-	***Brassica rapa***	**Mustard**	**957**	**99.67**
	Olive	-	-	-	***Allium sativum***	**Garlic**	**183**	**100**
	Sage	-	-	-				
	Parsley	-	-	-				
	Thyme	-	-	-				
	Onion	-	-	-				
	Wheat	*Triticum aestivum*	14005	100				
	Sugar	-	-	-				
	Rapeseed	-	-	-				
	Sunflower Seed	-	-	-				
F_241	Coconut	-	-	-	***Coriandrum sativum***	**Coriander**	**1955**	**99.7**
	Garlic	*Allium sativum*	120	100	***Nicotiana tabacum***	**Tobacco**	**162**	**99.7**
	Maize	-	-	-	***Allium cepa***	**Onion**	**498**	**100**
	Mustard	-	-	-				
	Potato	-	-	-				
	Wheat	*Triticum aestivum*	28486	100				

	Soya	*Glycine max*	35977	100				
F_242	**Garlic**	*Allium sativum*	116	100	***Coriandrum sativum***	**Coriander**	**3468**	**99.76**
	Mustard	-	-	-	***Allium cepa***	**Onion**	**1025**	**100**
	Seaweed	-	-	-	***Nicotiana tabacum***	**Tobacco**	**266**	**99.7**
	Soya	*Glycine max*	30368	100	***Pisum sativum***	**Pea**	**108**	**100**
	Wheat	*Triticum aestivum*	19127	100				
F_244	Paprika	-	-	-	***Brassica rapa***	**Mustard**	**4074**	**99.67**
	Garlic	*Allium sativum*	2063	100	***Nicotiana tabacum***	**Tobacco**	**133**	**99.7**
	Onion	*Allium cepa*	475	100				
	Pea	-	-	-				
	Sage	-	-	-				
	Sunflower Seed	-	-	-				
	Wheat	*Triticum aestivum*	20134	100				
F_246	Chillies	-	-	-	***Cuminum cyminum***	**Cumin**	**125**	**100**
	Garlic	-	-	-	***Coriandrum sativum***	**Coriander**	**174**	**99.76**
	Ginger	-	-	-				
	Mustard	-	-	-				
	Onion	*Allium cepa*	475	100				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	41224	100				
	Wheat	*Triticum aestivum*	2732	100				
F_247	Potato	-	-	-	***Allium cepa***	**Onion**	**275**	**100**
	Sunflower seed	-	-	-	***Nicotiana tabacum***	**Tobacco**	**253**	**99.7**
	Soya	*Glycine max*	19882	100				
	Wheat	*Triticum aestivum*	23687	100				
F_248	Mustard	-	-	-	***Coriandrum sativum***	**Coriander**	**274**	**100**
	Onion	*Allium cepa*	781	100				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	31031	100				
	Wheat	*Triticum aestivum*	10396	100				
F_249	Garlic	-	-	-				
	Maize	-	-	-				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	56589	100				
	Wheat	*Triticum aestivum*	6114	100				

F_250	Wheat	-	-	-	***Cuminum cyminum***	**Cumin**	**129**	**100**
	Maize	-	-	-	***Coriandrum sativum***	**Coriander**	**201**	**99.76**
	Mustard	-	-	-				
	Onions	*Allium cepa*	524	100				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	43912	100				
	Wheat	*Triticum aestivum*	6328	100				
F_251	Celery	-	-	-				
	Garlic	-	-	-				
	Onions	-	-	-				
	Sugar	-	-	-				
	Soya	*Glycine max*	87970	100				
	Sunflower	-	-	-				
	Wheat	*Triticum aestivum*	747	100				
F_252	Corn	-	-	-				
	Garlic	-	-	-				
	Lemon	-	-	-				
	Seaweed	-	-	-				
	Soya	*Glycine max*	52446	100				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	843	100				
F_253	Curry Leaves	-	-	-	***Triticum aestivum***	**Wheat**	**256**	**100**
	Soya	*Glycine max*	64202	100	***Cucurbita pepo***	**Butternut**	**195**	**99.7**
	Sunflower seed	-	-	-				
F_254	Beans	-	-	-	***Brassica rapa***	**Mustard**	**5027**	**99.7**
	Butterbeans	-	-	-	***Nicotiana tabacum***	**Tobacco**	**304**	**99.7**
	Butternut	*Cucurbita pepo*	394	100				
	Cayenne Pepper	-	-	-				
	Chickpeas	*Cicer arietinum*	5306	99.70				
	Wheat	*Triticum aestivum*	27161	100				
	Coriander	-	-	-				
	Garlic	-	-	-				
	Ginger	-	-	-				
	Onions	-	-	-				
	Pea	-	-	-				
	Red Kidney Beans	-	-	-				
	Spinach	-	-	-				

	Sweetcorn	-	-	-				
	Canola	-	-	-				
F_255	Beans	-	-	-	***Cucurbita pepo***	**Butternut**	**4401**	**100**
	Cayenne Pepper	-	-	-	***Secale cereale***	**Barley**	**29280**	**100**
	Corn	-	-	-	***Cuminum cyminum***	**Cumin**	**155**	**100**
	Wheat	*Triticum aestivum*	5705	100				
	Coriander	*Coriandrum sativum*	942	100				
	Garlic	-	-	-				
	Parsley	-	-	-				
	Ginger	-	-	-				
	Haricot bean	-	-	-				
	Lentil	-	-	-				
	Oats	-	-	-				
	Onions	-	-	-				
	Paprika	-	-	-				
	Potato	-	-	-				
	Tomato	-	-	-				
	Canola	-	-	-				
F_256	Celery	-	-	-	***Coriandrum sativum***	**Coriander**	**2233**	**100**
	Mustard	-	-	-	***Allium cepa***	**Onion**		**100**
	Potato	-	-	-	***Nicotiana tabacum***	**Tobacco**	**180**	**99.7**
	Coconut	-	-	-	***Pisum sativum***	**Pea**	**141**	**100**
	Soya	*Glycine max*	23214	100				
	Wheat	*Triticum aestivum*	11848	100				
F_257	Black Pepper	-	-	-	***Cicer arietinum***	**Chickpea**	**831**	**100**
	Chia seeds	-	-	-	***Nicotiana sylvestris***	**Woodland tobacco**	**146**	**99.7**
	Garlic	-	-	-				
	Ginger	-	-	-				
	Mustard	-	-	-				
	Potato	-	-	-				
	Rice	-	-	-				
	Rosemary	-	-	-				
	Soya	*Glycine max*	62423	100				
	Tumeric	-	-	-				
	Sunflower seeds	-	-	-				
F_258	Bay leaf	-	-	-	***Triticum aestivum***	**Wheat**	**1288**	**100**

	Black pepper	-	-	-	***Brassica rapa***	**Mustard**	**2262**	**100**
	Cloves	-	-	-	***Cucurbita pepo***	**Butternut**	**149**	**98.8**
	Coriander	*Coriandrum sativum*	2133	100	***Aegilops tauschii***	**Tausch's goatgrass**	**145**	**98.8**
	Cumin	*Cuminum cyminum*	389	100				
	Garlic	-	-	-				
	Mustard	*Sinapis alba*	23883	100				
	Oregano	-	-	-				
F_259	Bay leaf	-	-	-	***Triticum aestivum***	**Wheat**	**6390**	**100**
	Black Pepper	-	-	-	***Brassica rapa***	**Mustard**	**2079**	**100**
	Chilli	-	-	-	***Cucurbita pepo***	**Butternut**	**149**	**98.8**
	Cloves	-	-	-	***Aegilops tauschii***	**Tausch's goatgrass**	**216**	**98.8**
	Coriander	*Coriandrum sativum*	3573	100				
	Cumin seeds	*Cuminum cyminum*	815	100				
	Olive	-	-	-				
	Garlic	-	-	-				
	Mustard	*Sinapis alba*	23240	100				
F_260	**Garlic**	*Allium sativum*	437	100				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	10031	100				
F_262	Paprika	-	-	-	***Brassica rapa***	**Mustard**	**2689**	**99.67**
	Garlic	-	-	-				
	Onion	*Allium sativum*	1835	100				
	Potato	-	-	-				
	Sage	-	-	-				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	10362	100				
	White pepper	-	-	-				
F_264	Mustard	-	-	-	***Coriandrum sativum***	**Coriander**	**1766**	**100**
	Onions	*Allium cepa*	1185	100	***Cuminum cyminum***	**Cumin**	**694**	**100**
	Sunflower seed	-	-	-	***Nicotiana tabacum***	**Tobacco**	**180**	**99.7**
	Wheat	*Triticum aestivum*	18326	100				
	Soya	*Glycine max*	14835	100				
F_265	**Wheat**	*Triticum aestivum*	1288	100				

	Garlic	-	-	-				
	Soya	*Glycine max*	46999	100				
	Sunflower seed	-	-	-				
F_266	**Wheat**	*Triticum aestivum*	5622	100				
	Garlic	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	68337	100				
F_267	Black pepper	-	-	-	***Triticum aestivum***	**Wheat**	**232**	**100**
	Chickpea	*Cicer arietinum*	62546	100	***Lens culinaris***	**Lentil**	**285**	**100**
	Cumin	-	-	-	***Pisum sativum***	**Pea**	**785**	**99.7**
	Coriander	-	-	-	***Nicotiana tomentosa***	**Flowering Tobacco**	**540**	**99.7**
	Garlic	-	-	-				
	Mint	-	-	-				
	Onion	-	-	-				
	Parsley	-	-	-				
	Quinoa	-	-	-				
	Sweetcorn	-	-	-				
	Sunflower seed	-	-	-				
F_268	Black Pepper	-	-	-	***Nicotiana tabacum***	**Tobacco**	**165**	**99.7**
	Butternut	-	-	-	***Lens culinaris***	**Lentil**	**511**	**99.7**
	Carrots	-	-	-				
	Chickpea	*Cicer arietinum*	14081	100				
	Wheat	-	-	-				
	Cumin	-	-	-				
	Garlic	-	-	-				
	Onion	-	-	-				
	Quinoa	-	-	-				
	Sage	-	-	-				
	Sweetcorn	-	-	-				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	23179	100				
F_269	**Chickpea**	*Cicer arietinum*	2175	100	***Triticum aestivum***	**Wheat**	**468**	**100**
	Barley	-	-	-				
	Onion	*Allium cepa*	562	100				
	Soya	*Glycine max*	53133	100				
	Tomato	-	-	-				
	Rapeseed	-	-	-				

Supplementary Table 2. 2: Comparison of the Plant ITS2 sequencing results to the ingredient composition listed on the label. The ingredient list presented in this table includes only those that could possibly be identified using this particular barcode (fungal or meat derived ingredients were excluded). Ingredients highlighted in **green** indicate a positive match (sequencing results align with ingredient composition on packaging). Ingredients in black indicate their presence on the packaging but were not detected by the primers used in this study. ASVs that constituted less than 2% of the total relative abundance were removed. The sequencing reads were present in all three independent PCRs. ASVs showing sequencing results that are not included in the ingredient composition listed on the label are highlighted in **red** including the percent identity (%ID), read abundance and common name).

ID	Listed Ingredients	Observed Species	Abundance	%ID	Other Species Identified	Common Name	Abundance	%ID
F_237	Chillies				***Allium cepa***	**Onion**	**1687**	**100**
	Garlic				***Coriandrum sativum***	**Coriander**	**1867**	**100**
	Maize				***Secale cereale***	**Barley**	**869**	**100**
	Mustard	*Sinapis alba*	1335	100				
	Potato							
	Wheat	*Triticum aestivum*	1665	100				
	Soya							
F_238	**Wheat**	*Triticum aestivum*	1547	100				
	Onion	*Allium cepa*	582	100				
	Barley	-	-	-				
	Rape Seed	-	-	-				
F_239	Barley							
	Olive	-	-	-				
	Sage	-	-	-				
	Parsley	-	-	-				
	Thyme	-	-	-				
	Onion	*Allium cepa*	211	100				
	Wheat	*Triticum aestivum*	520	100				
	Sugar	-	-	-				
	Rapeseed	-	-	-				
	Sunflower Seed	-	-	-				
F_241	Coconut	-	-	-	***Allium cepa***	**Onion**	**1656**	**100**
	Garlic	-	-	-	***Secale cereale***	**Barley**	**1727**	**100**
	Maize							
	Mustard	*Sinapis alba*	1972	100				
	Potato	-	-	-				
	Wheat	*Triticum aestivum*	3406	100				
	Soya	*Glycine Max*	290	100				
F_242	Garlic	-	-	-	***Secale cereale***	**Barley**	**1756**	**100**

	Mustard	*Sinapis alba*	2300	100	***Allium cepa***	**Onion**	**1645**	**100**
	Seaweed	-	-	-				
	Soya	-	-	-				
	Wheat	*Triticum aestivum*	3508	100				
F_244	Paprika	-	-	-				
	Garlic	-	-	-				
	Onion	*Allium cepa*	198	100				
	Pea	*Pisum sativum*	101	100				
	Sage	-	-	-				
	Sunflower Seed	-	-	-				
	Wheat	*Triticum aestivum*	5226	100				
F_246	Chillies	-	-	-	***Coriandrum sativum***	**Coriander**	**374**	**100**
	Garlic	-	-	-				
	Ginger	-	-	-				
	Mustard	*Sinapis alba*	417	100				
	Onion	*Allium cepa*	906	100				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	451	100				
	Wheat	*Triticum aestivum*	1110	100				
F_247	Potato	-	-	-	***Secale cereale***	**Barley**	**1999**	**100**
	Sunflower seed	-	-	-	***Coriandrum sativum***	**Coriander**	**1027**	**100**
	Soya	*Glycine Max*	101	100	***Allium cepa***	**Onion**	**262**	**100**
	Wheat	*Triticum aestivum*	4584	100	***Triticum monococcum***	**Einkorn Wheat**	**121**	**100**
F_248	**Mustard**	*Sinapis alba*	470	100	***Coriandrum sativum***	**Coriander**	**600**	**100**
	Onion	*Allium cepa*	1444	100				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	267	100				
	Wheat	*Triticum aestivum*	2830	100				
F_249	Garlic	-	-	-	***Secale cereal***	**Barley**	**388**	**100**
	Maize	-	-	-				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	660	100				
	Wheat	*Triticum aestivum*	663	100				
F_250	Wheat	-	-	-	***Coriandrum sativum***	**Coriander**	**518**	**100**
	Maize	-	-	-				

	Mustard	*Sinapis alba*	515	100				
	Onions	*Allium cepa*	843	100				
	Potato	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	1225	100				
	Wheat	*Triticum aestivum*	1955	100				
F_251	Celery	-	-	-				
	Garlic	-	-	-				
	Onions	-	-	-				
	Sugar	-	-	-				
	Soya	*Glycine max*	4075	100				
	Sunflower	-	-	-				
	Wheat	*Triticum aestivum*	176	100				
F_252	Corn	-	-	-				
	Garlic	-	-	-				
	Lemon	-	-	-				
	Seaweed	-	-	-				
	Soya	*Glycine max*	1271	100				
	Sunflower seed	-	-	-				
	Wheat	-	-	-				
F_253	Curry Leaves	-	-	-	***Foeniculum vulgare***	**Fennel**	**151**	**100**
	Soya	*Glycine max*	5346	100	***Coriandrum sativum***	**Coriander**	**195**	**100**
	Sunflower seed	-	-	-	***Cuminum cyminum***	**Cumin**	**110**	**100**
F_254	Beans	-	-	-	***Pisum sativum***	**Pea**	**1602**	**100**
	Butterbeans	-	-	-				
	Butternut	*Cucurbita pepo*	3273	100				
	Cayenne Pepper	-	-	-				
	Chickpeas	-	-	-				
	Wheat	*Triticum aestivum*	3269	100				
	Coriander	-	-	-				
	Garlic	-	-	-				
	Ginger	-	-	-				
	Onions	-	-	-				
	Pea	*Pisum sativum*	1602	100				
	Red Kidney Beans	-	-	-				
	Spinach	-	-	-				
	Sweetcorn	-	-	-				
	Canola	-	-	-				

F_255	Beans	-	-	-	***Sinapis alba***	**Mustard**	**153**	**100**
	Cayenne Pepper	-	-	-				
	Corn	-	-	-				
	Wheat	*Triticum aestivum*	130	100				
	Coriander	*Coriandrum sativum*	150	100				
	Garlic	-	-	-				
	Parsley	-	-	-				
	Ginger	-	-	-				
	Haricot bean	-	-	-				
	Lentil	*Lens culinaris*	1115	100				
	Oats	-	-	-				
	Onions	-	-	-				
	Paprika	-	-	-				
	Potato	-	-	-				
	Tomato	-	-	-				
	Canola	-	-	-				
F_256	Celery	-	-	-	***Secale cereale***	**Barley**	**1259**	**100**
	Mustard	*Sinapis alba*	2438	100	***Allium cepa***	**Onion**	**2510**	**100**
	Potato	-	-	-				
	Coconut	-	-	-				
	Soya	*Glycine Max*	116	100				
	Wheat	*Triticum aestivum*	2446	100				
F_257	Black Pepper	-	-	-	***Cicer arietinum***	**Chickpea**	**283**	**100**
	Chia seeds	-	-	-				
	Garlic	-	-	-				
	Ginger	-	-	-				
	Mustard	-	-	-				
	Potato	-	-	-				
	Rice	-	-	-				
	Rosemary	-	-	-				
	Soya	*Glycine max*	1581	100				
	Tumeric	-	-	-				
	Sunflower seeds	-	-	-				
F_258	Bay leaf	-	-	-	***Brassica napus***	**Rapeseed**	**131**	**99.67**
	Black pepper	-	-	-				
	Cloves	-	-	-				
	Coriander	*Coriandrum sativum*	1185	100				
	Cumin	-	-	-				

	Garlic	-	-	-				
	Mustard	*Brassica rapa*	4770	99.67				
	Oregano	-	-	-				
F_259	Bay leaf	-	-	-	***Brassica napus***	**Rapeseed**	**117**	**99.67**
	Black Pepper	-	-	-				
	Chilli	*Capsicum frutescens*	265	98.82				
	Cloves	-	-	-				
	Coriander	*Coriandrum sativum*	853	100				
	Cumin seeds	-	-	-				
	Olive	-	-	-				
	Garlic	-	-	-				
	Mustard	*Brassica rapa*	5325	99.67				
F_260	**Garlic**	*Allium sativum*	522	100				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	3542	100				
F_262	Paprika	-	-	-	***Pisum sativum***	**Pea**	**261**	**100**
	Garlic	-	-	-				
	Onion	-	-	-				
	Potato	-	-	-				
	Sage	-	-	-				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	907	100				
	White pepper	-	-	-				
F_264	**Mustard**	*Sinapis alba*	2107	100	***Coriandrum sativum***	**Coriander**	**1573**	**100**
	Onions	*Allium cepa*	2686	100	***Secale cereale***	**Barley**	**1381**	**100**
	Sunflower seed	-	-	-	***Triticum monococcum***	**Einkorn Wheat**	**101**	**100**
	Wheat	*Triticum aestivum*	4721	100				
	Soya	-	-	-				
F_265	**Wheat**	*Triticum aestivum*	666	100				
	Garlic	-	-	-				
	Soya	*Glycine max*	2911	100				
	Sunflower seed	-	-	-				
F_266	**Wheat**	*Triticum aestivum*	3618	100	***Triticum monococcum***	**Einkorn Wheat**	**142**	**100**
	Garlic	-	-	-				
	Sunflower seed	-	-	-				
	Soya	*Glycine max*	3136	100				

F_267	Black pepper	-	-	-				
	Chickpea	*Cicer arietinum*	11213	100				
	Cumin	*Cuminum cyminum*	1212	100				
	Coriander	*Coriandrum sativum*	196	100				
	Garlic	-	-	-				
	Mint	-	-	-				
	Onion	*Allium cepa*	189	100				
	Parsley	-	-	-				
	Quinoa	-	-	-				
	Sweetcorn	-	-	-				
	Sunflower seed	-	-	-				
F_268	Black Pepper	-	-	-	***Triticum monococcum***	**Einkorn Wheat**	**107**	**100**
	Butternut	*Cucurbita pepo*	158	100				
	Carrots	-	-	-				
	Chickpea	*Cicer arietinum*	3527	99.70				
	Wheat	-	-	-				
	Cumin	-	-	-				
	Garlic	-	-	-				
	Onion	-	-	-				
	Quinoa	-	-	-				
	Sage	-	-	-				
	Sweetcorn	-	-	-				
	Sunflower seed	-	-	-				
	Wheat	*Triticum aestivum*	2072	100				
F_269	**Chickpea**	*Cicer arietinum*	4007	100				
	Barley	-	-	-				
	Onion	*Allium cepa*	2143	100				
	Soya	*Glycine max*	1119	100				
	Tomato	-	-	-				
	Rapeseed	-	-	-				

Supplementary Table 2. 3: The nucleic acid concentrations and purity ratios of 32 homogenous vegetarian samples. ID represents the sample code given to each vegetarian food product for the purpose of simplicity and anonymity with regards to brand. The description provides an overview of the types of food products sampled for this study. The weight represents how much of each sample were used for DNA extractions. [Qubit] represents the concentration of dsDNA detected with dsDNA HS Assay kit for each sample while [Nanodrop] provides the total nucleic acid concentration measured with NanoDrop™ 2000 UV/VIS Spectrophotometer in each sample. The measurement of 260/280 is used as a measurement of purity for DNA at 1.8 while 260/230Colour coding: **Green = Good**, **Yellow = Sufficient**, **Red = Bad**

ID	Description	Weight	[Qubit] ng/ul	[Nanodrop] ng/ul	260/280	260/230
F237	Meat Free Spicy Sausages	370g	283.2	1615.6	**1.95**	**1.15**
F238	Meat Free Traditional Burgers	200g	54.6	471.3	**1.72**	**0.86**
F239	Meat Free Pepper and Herb flavoured Sausages	360g	60	949	**1.77**	**1.03**
F241	Meat Free Polony Slicing Sausage	320g	224.4	1344	**2.03**	**1.39**
F242	Meat Free Vegetarian Polony Slicing Sausage	280g	144	1040	**1.86**	**0.86**
F243	Meat Free Breakfast Rashers	200g	113.6	890.9	**2.09**	**1.52**
F244	Meat Free Vegan Crumbed Schnitzels	320g	124.8	1112.9	**2**	**1.36**
F245	Meat Free Vegetarian Mince	430g	98	558.1	**2.12**	**2.32**
F246	Meat Free Asian Spiced Burgers	360g	196	1570.9	**1.81**	**0.85**
F247	Meat Free Braai-style Sausages	350g	300	1874.4	**1.93**	**1.08**
F248	Meat Free Thick Cut Chunky Strips	260g	291.2	1912.5	**1.95**	**1.12**
F249	Meat Free Chicken Style Strips	400g	272	2342.2	**1.83**	**0.91**
F250	Meat Free Golden Crumbed Schnitzel	340g	123	1086.2	**1.89**	**0.89**
F251	Meat Free Vegan Pops	390g	302	2381.3	**1.97**	**1.12**
F252	Meat Free Battered Prawn-style Prawns	290g	394	2646.6	**2.06**	**1.49**
F253	Meat Free Korma Curry	170g	307.2	1665.9	**2.11**	**1.79**
F254	Vegan Patties	380g	46	502.2	**1.7**	**0.7**
F255	Vegan BBQ Patties	360g	86.4	1121.2	**1.78**	**0.83**
F256	Meat Free Hot Dogs	350g	194	1459.2	**1.9**	**0.94**

ID	Description	Weight	[Qubit] ng/ul	[Nanodrop] ng/ul	260/280	260/230
F257	Gluten Free Nuggets	370g	288	2333.8	**1.91**	**0.97**
F258	Mushroom Biltong Plain	100g	163.2	1696.7	2.11	**1.94**
F259	Mushroom Biltong Chilli	330g	288	5814.2	2	**1.84**
F260	Meat Free Garlic and Mushroom Schnitzels	330g	61.2	579.9	**1.89**	**1.3**
F261	Meat Free Chicken-style Fillets	350g	248	2148.2	2.06	**1.79**
F262	Vegan Nuggets	360g	117	975.4	**1.9**	**1.17**
F263	Meat Free Chicken-style Pieces	320g	32.4	456.3	**1.96**	**1.36**
F264	Meat Free Original Burgers	390g	133	1375.4	**1.77**	**0.79**
F265	Meat Free Chicken-style Burgers	300g	218	1647.5	2.02	**1.32**
F266	Meat Free Chicken-style Nuggets	370g	214	2096.1	**1.99**	**1.17**
F267	Chickpea and Quinoa Falafels	180g	123.2	1011.4	**1.81**	**0.67**
F268	Chickpea and roasted Butternut Balls	360g	98.4	1217.8	**1.62**	**0.5**
F269	Vegetarian Burgers	350g	330	4183.4	**1.88**	**0.93**