The future of NGS (Next Generation Sequencing) analysis in testing food authenticity

Haynes, E., Jimenez, E., Angel Pardo, M., & Helyar, S. (2019). The future of NGS (Next Generation Sequencing) analysis in testing food authenticity. *Food Control*, *101*, 134-143. https://doi.org/10.1016/j.foodcont.2019.02.010

Published in:
Food Control

Document Version:
Publisher's PDF, also known as Version of record

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Food Control 101 (2019) 134–143

Contents lists available at ScienceDirect

Food Control

journal homepage: www.elsevier.com/locate/foodcont

Review

The future of NGS (Next Generation Sequencing) analysis in testing food authenticity

Edward Haynes[a,*], Elisa Jimenez[b], Miguel Angel Pardo[b], Sarah J. Helyar[c]

[a] Fera, York, UK
[b] AZTI, Derio, Bizkaia, Spain
[c] Institute for Global Food Security, Queen's University Belfast, Belfast, UK

ARTICLE INFO

Keywords:
Genomics
Metabarcoding
Databases
Bioinformatics
Validation
NGS limitations

ABSTRACT

The authenticity of foodstuffs is an important issue for consumers, regulators, producers and processors, as fraudulent practices can negatively affect consumer confidence and safety, as well as the operating models of legitimate businesses. This review provides an overview of the current landscape of Next Generation Sequencing (NGS) applications for food authenticity, and looks to identify the potential future developments for this technology. Specific areas highlighted include the range of NGS platforms and sequence databases available, validation of NGS, and limitations and appropriate uses of these technologies.

Many NGS platforms are available, with different properties (such as sequence read length and output) suited to different analyses. Despite this wealth of options, more platforms are being brought out frequently, and advances such as reduced error rate will enable their expanded use for food authenticity. This rapid expansion in the use of DNA sequencing has led to an equally rapid enlargement in sequence databases, and the construction of contemporaneous, authenticated databases may be a useful innovation for the application of NGS to authenticity. Such applications will require robust quality control criteria and proficiency testing schemes, both of which are being developed. Despite several caveats, for example around effective extraction and amplification of DNA, NGS is a strong candidate to become a valuable aid or even the technology of choice to achieve regulatory compliance and reputation protection in a number of food fraud situations, particularly for highly complex food matrices.

1. Introduction

1.1. Authenticity

Food authenticity is a big concern for consumers, food authorities and food producers and processors, since incorrect food labelling and other types of fraudulent practices have been demonstrated to negatively affect the confidence and even the safety of the final consumer (Barnett et al., 2016). European Union regulation (EU) No. 1169/2011 (EU, 2011) requires that consumers should be appropriately informed regarding the food they consume. This is vital in order to achieve a high level of health protection and to guarantee their right to information, as well as to protect the businesses of scrupulous producers from unfair competition. Consumers' choices can be influenced by health, economic, environmental, social and ethical considerations. In fact, the general dictionary definition of "*authenticity*" is "*the quality of being authentic, trustworthy, or genuine*", and the relevant dictionary definitions of "*authentic*" include "*not false or copied; genuine; real*" and "*having an origin supported by unquestionable evidence; authenticated; verified*". More specifically regarding food authenticity, a recently produced CEN standard defines authenticity in a food and feed context as the match between the *food product characteristics and the corresponding food product claims* (CEN WS86). These labelling requirements, which are legally specified and differ depending on the product, may include the scientific name or breed, and production method (e.g. organic, free-range, wild-caught etc.). However, other features of the product can also be included by producers to inform the consumer, including (i) ethical issues (halal, vegetarian, etc.), (ii) nutritional composition (vitamins, omega 3, etc.), (iii) the area where the product was caught or farmed (for sustainability reasons, or with particular regard to EU legislation regarding protected designation of origin (PDO), protected geographical indication (PGI), traditional specialities guaranteed (TSG) etc.), (iv) status of the product (such as whether the product has been previously frozen and defrosted) and (v) the presence of undeclared

* Corresponding author.
E-mail addresses: edward.haynes@fera.co.uk (E. Haynes), ejimenez@azti.es (E. Jimenez), mpardo@azti.es (M.A. Pardo), S.helyar@qub.ac.uk (S.J. Helyar).

https://doi.org/10.1016/j.foodcont.2019.02.010
Received 19 December 2018; Received in revised form 13 February 2019; Accepted 15 February 2019
Available online 18 February 2019

ingredients that can also represent a health risk for the consumer (allergens such as gluten, nuts, etc.).

1.2. DNA-based techniques

An example of a common food fraud is the substitution of one ingredient by a similar, cheaper one, and different analytical procedures can be used for the identification of this food adulteration including spectroscopic, chromatographic, proteomic and Polymerase Chain Reaction (PCR) based approaches (Primrose, Woolfe, & Rollinson, 2010). However, in many cases the fraud is based on the substitution of one ingredient with another that is a different breed or species, and for this DNA based methodologies have been shown to be an ideal tool to address the problem, due to the sensitivity, accuracy and ease of testing, and the stability of DNA under a range of food processing methods (Catalano, Moreno-Sanz, Lorenzi, & Grando, 2016; Pardo, 2015). The majority of methods amplify specific areas of DNA using PCR, a rapid and easy-to-use technique that permits the amplification of a small DNA segment, which is subsequently used as a molecular marker. For qualitative species identification, DNA regions within the mitochondria (animals) or chloroplast (most plants) are primarily used, even for the differentiation of closely related species (although there are exceptions), whereas nuclear markers are much more suitable for quantification and the identification of geographical origin or specific breed or landrace, among other applications (Nielsen et al., 2012; Wilkinson et al., 2012). More recently, methods applied within an authenticity context have focused on the DNA sequence at a specific site, as these are considered to have greater reliability (and hence are easier to present in a legal setting). The identification of many products can be achieved through the direct sequencing of short, standardized gene fragments (e.g. DNA Barcoding (Hebert, Cywinska, Ball, & deWaard, 2003), Forensically Informative Nucleotide Sequencing (FINS) (Bartlett & Davidson, 1992), etc.). While these gene fragments differ between taxa, for most animals a fragment (~655 base pairs) of the cytochrome c oxidase subunit 1 mitochondrial gene (COI) has been shown to provide reliable species level discrimination. For plants, a wider range of fragments are currently used (including *rbcL*, *matK*, and *ITS* regions, see Madesis, Ganopoulos, Sakaridis, Argiriou, and Tsaftaris (2014), for a review of regions and methods). There has also been a proliferation of methods developed to generate and detect products of specific oligonucleotide primers, such as oligonucleotide ligation assay (OLA) (Consolandi et al., 2007); real time-PCR (Taylor, Fox, Rico, & Rico, 2002); High Resolution Melting (Druml & Cichna-Markl, 2014); Loop-mediated isothermal amplification (LAMP) (Randhawa, Chhabra, Bhoge, & Singh, 2015; Ye et al., 2016). Finally, Digital PCR may be a promising approach for the detection of minute traces of biological adulterants in foodstuffs (Ren, Deng, Huang, Chen, & Ge, 2017; Shehata et al., 2017).

While these molecular methods are used frequently for routine analyses, they have some significant limitations. The largest of these is that each test is targeted to answer a specific question: such as which tuna species is present, or from what species does this meat originate? Hence the more difficult questions, such as the identification of all components within a complex food matrix, are highly demanding to answer. Here, Next Generation Sequencing (NGS) is beginning to have an impact. Within food safety NGS is already being applied, for example to allow in theory the accurate identification of the great majority of microbial taxa within a product, including microorganisms that are unculturable and those that are only present in small numbers (Leonard, Mammel, Lacher, Elkins, & Drake, 2015; Mayo et al., 2014).

The challenge now is to apply NGS technologies to address food authenticity questions. One of several potential uses of this is to identify the presence of transgenes to identify GMOs (Fraiture, Herman, De Loose, Debode, & Roosens, 2017; Kamle, Kumar, Patra, & Bajpai, 2017) which are increasing in scope and complexity. Moreover, the advantage of NGS is the ability to simultaneously screen multiple different genomic regions, to enable the identification of all plant, animal, fungal and microalgae ingredients in food commodities. For this reason, previously generated data including sequences from reference gene databases and peer reviewed articles are a precious source of specific and universal primers which will be required to maximise the advantage of NGS platforms. Similarly, the location of thousands of SNPs within the genome and their diagnostic evaluation, previously described and validated in a number of DNA based methodologies, will also be useful for NGS applications.

This work was commissioned as a Scientific Opinion under the EU FP7 funded FoodIntegrity project, to highlight the current state of the art in NGS applications for food authenticity, and to infer possible future trends in this fast-moving area of research. Herein we describe the current status and future trends in a number of important areas of NGS, including the features of various platforms, the range and reliability of available databases and bioinformatics solutions, the extent of validation and quality control for NGS data, limitations of the technology and crucially the way this technology is likely to be implemented in the future.

2. Next Generation Sequencing

High throughput DNA sequencing, enabled by advances in sequencing technology (NGS) has revolutionised many fields of biology, including medical microbiology, plant and animal genomics, and the study of gene transcription (e.g. Goodwin, McPherson, & McCombie, 2016). Innovation in sequencing technologies continues, and the highest throughput platform is now capable of sequencing 18,000 human genomes in a single year, at the cost of only $1000 per genome. This compares favourably to the $3 billion cost of the first human genome project, and the $10 million cost of sequencing additional human genomes before the NGS revolution (Hayden, 2014).

2.1. NGS platforms

A number of NGS platforms, using a variety of different chemistries, are now extensively available for high throughput DNA sequencing, and additional sequencing technologies are still being developed, e.g. (Ansorge, 2016; Fuller et al., 2016). The most widely used family of technologies is manufactured by Illumina (Hodkinson & Grice, 2015). These use reversible fluorescent dideoxy terminators to sequence DNA clusters amplified on the surface of disposable flow cells. Illumina provide an array of options (see Table 1) which produce a range of differing quantities and lengths of DNA sequences (or 'reads' as they are often referred to), depending on the needs of the user; i.e. from the Illumina iSeq 100, which can produce a maximum of 1.2 billion bases of sequence per run, to the HiSeq X Ten, which is a suite of ten instruments each producing up to 1.8 trillion bases. This high throughput enables population level sequencing of animal and plant genomes. Illumina continues to produce new platforms with a range of throughputs, with recent releases including the NovaSeq 5000 and 6000. These sequencer options allow the user to adjust output to match their anticipated requirements. Illumina platforms are short read sequencers, producing a maximum of 2 × 300 bases for paired-end reads, available only on the MiSeq sequencer. Alternative platforms that produce short reads are available, for example using Ion Torrent technology, which have their own advantages (e.g. cost per base) and disadvantages (e.g. higher homopolymer error rate (Divoll, Brown, Kinne, & McCracken, 2018; Laehnemann, Borkhardt, & McHardy, 2016), although this can to some extent be corrected for algorithmically (L. Song, Huang, Kang, Ren, & Ding, 2017)). Ion Torrent utilises semiconductor technology to detect H^+ ions released during the incorporation of sequentially introduced nucleotides into an elongating DNA strand. There are currently three Ion Torrent devices, the Personal Genome Machine (PGM), the Proton, and the GeneStudio S5. Short sequence reads, regardless of the technology used to produce them, are less well suited to some specific

Table 1
Data on outputs of a selection of commercially available sequencing platforms. All data is taken from the providers' system specifications. Illumina machines produce paired end reads, so total read length is the sum of both reads in the pair. QIAGEN also produces a sequencing platform, the GeneReader, though this is currently focussed towards human health applications. For reference, the human genome is approximately 3 Gb long. In practise, much more than 3 Gb of sequence must be generated to accurately sequence a single human genome, to ensure uniform coverage and sufficient depth of sequence to eliminate errors. To have an average of ten nucleotides coverage at each position on the genome, an experiment would need to produce 30 Gb of sequence data.

Provider	Sequencer	Maximum Read Length (bp)	Maximum sequence yield per run (Gb)
Illumina	HiSeq 4000	2 × 150	1,500
	HiSeq 3000	2 × 150	750
	HiSeq 2500	2 × 125	1,000
	NextSeq 500	2 × 150	120
	MiSeq	2 × 300	15
	MiniSeq	2 × 150	7.5
ThermoFisher Scientific	Ion Proton	200	10
	Ion PGM	400	2
	Ion Torrent S5	600	10-15 (though there is a trade-off between read length and output)
Pacific Biosciences	PacBio RSII	Half of data in reads > 20,000 Max length > 60,000	1
	PacBio Sequel	Half of data in reads > 20,000 Max length > 60,000	7
Oxford Nanopore	MinION	Read length = DNA Fragment length Longest reported approaching 1 Megabase	10–20
	GridION X5	Read length = DNA Fragment length Longest reported approaching 1 Megabase	50–100
	PromethION	Read length = DNA Fragment length Longest reported approaching 1 Megabase	11,000 (theoretical maximum)

applications such as resolving complex genome arrangements or sequencing long PCR amplicons.

One technology that produces longer reads is the single molecule real time (SMRT) sequencing of Pacific Biosciences (PacBio). This approach again involves fluorescently labelled nucleotides, which are incorporated into the replicating DNA molecule. A DNA template-polymerase complex is immobilised at the bottom of a well tens of nanometres in diameter, known as a zero-mode waveguide, which acts as a powerful light microscope. This allows detection of nucleotides labelled with different fluorophores when they are incorporated into the replicating DNA molecule. The primary advantage of PacBio platforms lies in their ability to produce relatively long reads (up to around 60,000 bases (60 kilobases, kb)). The most recent PacBio system is the Sequel, which has the potential to produce a million DNA sequences per run, far fewer than that produced by the highest throughput short read platforms. Initial concerns about high error rates have been addressed by circularising the DNA molecule to allow an accurate consensus sequence to be produced. However not all reads produced will be high quality reads, and these more accurate reads will be of a shorter length (Rhoads & Au, 2015).

A different long-read technology has been developed by Oxford Nanopore Technologies, in the form of the portable MinION device and benchtop PromethION (the even smaller SmidgION, designed to be used with a smart phone is also due for release). In these products DNA strands are fed through nano scale pores, and the sequence inferred from changes in potential difference across the pore. This technology still has issues surrounding error rate, but as new pore chemistries are being brought online this is improving. Additionally, nanopore technology has the potential to produce extremely long sequence reads, of the order of megabases. Long read technology has important applications in elucidating the correct arrangement of complex genomes, and for sequencing long PCR amplicons.

2.2. Databases

A consequence of the increase in throughput and availability of high throughput DNA sequencers is the need to store large amounts of sequence data. Global efforts at storing and sharing DNA sequence data have been underway for several decades. The GenBank database, managed by the United States National Center for Biotechnology Information (NCBI) has been running since 1982 (Bilofsky & Burks, 1988), and as of June 2017 contained over 260 Gigabases of DNA sequence. This is in addition to the more than 3.2 Terabases of Whole Genome Shotgun sequence processed by GenBank since 2002 (NCBI, 2018a). These data are mirrored daily among the constituents of the International Nucleotide Sequence Database Collaboration (INSDC); NCBI, the European Bioinformatics Institute (EMBL-EBI) and the DNA Data Bank of Japan (DDBJ), to ensure uniform global access to a comprehensive collection of sequence data (Benson et al., 2012).

Since 2007 a separate repository of raw NGS data has been hosted by INSDC, the Sequence Read Archive (SRA) (Leinonen, Sugawara, & Shumway, 2010). The SRA stores raw sequence data in a variety of formats, with accompanying metadata about sample origin (Kodama, Shumway, & Leinonen, 2011) and varying amounts of user-supplied information about library preparation technique (Alnasir & Shanahan, 2015). As of 2018 the SRA hosts more than 8 Petabases of open access DNA sequence data (NCBI, 2018c). This is a tremendous, freely available resource for interrogation, or for comparison of experimentally-derived sequence data. However, the database should be used judiciously, as there is limited curation of data in the SRA, especially of user-supplied metadata about sample origin. The gold standard of reliable sequence data hosted by INSDC are RefSeq sequences. Unlike the majority of GenBank entries these are curated sequences that have been assessed bioinformatically, and by expert scientific staff (O'Leary et al., 2016), and as of September 2018 represent over 84,000 species (NCBI, 2018b).

Curated, application-specific DNA sequence databases hosted by other organisations do exist. These include gene-specific databases, for example *16S* ribosomal RNA gene data for bacteria and archaea (DeSantis et al., 2006; Pruesse et al., 2007), or other barcoding genes for various groups of organisms of interest such as those curated as part of the Consortium for the Barcode of Life (Ratnasingham & Hebert, 2007). Some of these are of importance from a food authenticity point of view, including fish (Ward, 2012) and plants (Ferri et al., 2015; Hollingsworth et al., 2009). Limited authenticity-specific sequence databases have been created, an example being the JRC GMO-Amplicons database containing more than 240,000 *in silico*-generated amplicons related to genetically modified organisms (Petrillo et al., 2015). The databases to be interrogated will depend to a great extent on the targets of the assay, whether that be a barcoding gene of the ingredients

themselves (Staats et al., 2016), the microbial fingerprint of the product (Cao, Fanning, Proos, Jordan, & Srikumar, 2017; Mezzasalma et al., 2017) or even the whole genome sequences of product-associated microbes (Douillard et al., 2013). As sequencing technology reduces in price and becomes more ubiquitous, this may expand to genomic or shotgun sequencing of the ingredients themselves, which may be particularly useful for revealing novel genetic modification events (Park et al., 2015; Yang et al., 2013). Nonetheless investment is required to expand existing databases, or construct new databases, to ensure fitness for purpose for identification for authenticity purposes.

An alternative approach to comparisons with existing or historical databases would be the use of contemporaneous databases. That is, databases which are compiled from authentic samples collected at the same time as a survey or other investigation. This could be a particularly important innovation when utilising the microbial community composition of a product as a fingerprint for origin. These microbial communities will be expected to change over time, or in response to other environmental variables. The construction of contemporaneous databases could prove to be cheaper than populating large, comprehensive databases, as they could focus on specific commodities or geographical locations of interest. Additionally, this approach could benefit from analytical and technological improvements as, unlike historical database, reference material would be concurrently analysed. While these techniques and databases are being developed, conclusions about geographical origin of products inferred from NGS data should be taken with care.

3. Bioinformatic tools

The generation of large amounts of sequencing data requires the use of specialised software for analysis. There are, in general, two main options for investigators; commercial software, or freely available open-source software. There are numerous commercial software options available for the interrogation of NGS data, with examples including Bionumerics (Applied Maths), CLC Genomics Workbench (Qiagen), and SeqSphere+ (Ridom). Currently, many of these are focussed on whole genome sequencing (WGS) and typing of microbial strains, while fewer appear to be available for metagenomic applications (though One Codex is a commercially available metagenomics platform (Minot, Krumm, & Greenfield, 2015)). The advantages of commercial software include ease of use, and availability of technical support, but this does require a financial commitment from the user, and may have limited customisability compared with open source software. However, currently available tools are still targeted at the research scientist, and may not yet be intuitive or user-friendly enough for uptake by the food industry.

Open source software has source code available for inspection and modification, rather than being a so-called 'black box' analytical pipeline. Such software may be free of charge, or may require payment of a licence fee, and this may vary depending on whether it is for academic or commercial use. Much open source software is available for a wide variety of different bioinformatics applications (Roumpeka, Wallace, Escalettes, Fotheringham, & Watson, 2017). An open source model allows a full understanding of the software, as well as allowing user modifications, but may come with limited support, and require expert user input at the investigating institute. Additional issues of computations involving large scale datasets include the high computational power requirement. A potential way around this is the use of cloud based analytical platforms (e.g. CLIMB (Connor et al., 2016)). Indeed, Oxford Nanopore Technologies have a suite of easy to use analytical pipelines hosted on their EPI2ME cloud-based analysis system, including the WIMP (What's In My Pot) pipeline for identifying taxa present in a sample. This might be a model whereby rapid, easy to implement analyses of complex datasets can be put in the hands of frontline investigators or non-bioinformaticians. Uptake by commercial companies may be limited however, due to restrictions on the transfer of confidential data to external servers.

4. Validation and standardisation

Due to the relatively recent development of NGS technologies, internationally recognised validation, standardisation and accreditation efforts are less advanced than they are for other molecular biology technologies (although reference standards and approaches are being published (Hardwick, Deveson, & Mercer, 2017; Mahamdallie et al., 2018)). Quality control (QC) for NGS analyses can be divided into pre- and post-sequencing QC. Pre-sequencing QC involves measures of the starting DNA concentration and fragment size distribution, and assessment of multiple rounds of PCR and clean-up. There is some evidence that the final sequence quality is relatively robust to variation at later stages in the library preparation process, and the input DNA has the largest impact on output sequence quality (Nietsch et al., 2016). Post-sequencing QC starts with the raw sequence produced by the machine, which can be checked or trimmed on length and quality metrics (indicated by phred quality scores assigned to each base in a FASTQ file). Further QC metrics are more experiment-specific, and can involve measures of the number and quality of reads mapping to a reference, and their depth of coverage (Nietsch et al., 2016). A number of tools now exist to summarise basic post-sequencing quality metrics with some even attempting to identify and remove contaminating sequences (Zhou, Su, & Ning, 2014).

Beyond internal QC, particular challenges exist around validation and proficiency testing (PT) for NGS tests, including a lack of well-characterized PT materials or standardized comparison metrics, as well as cost and time requirements of participants (Gargis et al., 2012). These limitations are even more problematic for food products, where samples may have been subject to processing procedures or the addition of chemical preservatives which may degrade DNA or inhibit subsequent enzymatic reactions. Quality standards defined for DNA and sequence volume and quality in non-food applications may therefore not be suitable for food authenticity applications. Nonetheless, in recent years progress has been made in developing standards for healthcare related NGS fields including human (Aziz et al., 2015) and microbial (Lefterova, Suarez, Banaei, & Pinsky, 2015) sequencing, which may inform the development of food-specific standards. Plans are also being developed for the creation of an International Organization for Standardization (ISO) working group to coordinate the standardization of NGS across a number of areas including food and medical genomics (ISO, 2017a; 2017b).

A small number of NGS proficiency testing (PT) schemes exist. One such scheme is run by the Global Microbial Identifier (GMI), an initiative comprising members from government, industry and academic organisations dedicated to the expansion of microbial WGS approaches to the study of disease epidemiology. Since 2015 GMI has been running PTs for both wet lab (sequence generation) and dry lab (phylogenetic/clustering analysis) aspects of bacterial WGS. This is supported by the EU H2020 Compare project. This PT was developed after a survey of the attitudes and requirements of NGS practitioners and end users (Moran-Gilad et al., 2015). Validation exercises for metabarcoding approaches have also been undertaken (Arulandhu et al., 2017).

5. Limitations of NGS

5.1. DNA quantity and quality in food commodities

The application of NGS procedures to food authenticity is dependent on obtaining sufficient good quality DNA, a crucial step to ensure that all DNA sequences of interest present in a food sample can be amplified and identified (plant, fungal, animal and bacterial) (Ripp et al., 2014). It is now possible to extract DNA from virtually any type of ingredient or finished product, including fresh, raw, dried, powdered and highly processed materials. It is important to take into account that the kind of

material under study can have a considerable effect on the ability to accurately detect the component of interest. Each matrix may require specific procedures, and the amount of starting material could vary enormously to enable the extraction of a minimum amount of DNA necessary to work with. Furthermore, it is essential that inhibitors, such as enzymes, complex polysaccharides and divalent cations, are completely removed from purified DNA to remove any potential downstream bias.

Certain food commodities contain an extremely limited amount of DNA or only degraded DNA. This can result in an incomplete amplification of those fragments with very low representation. When dealing with metabarcoding approaches, the length of the marker of interest should not exceed 100–200 base pairs in order to allow amplification of degraded DNA (Staats et al., 2016) and sequencing on widely available NGS platforms. In the specific case of an untargeted analysis approach to complex food matrices, this issue becomes of the utmost importance, to avoid misrepresentation. Moreover, with the use of smaller DNA fragments come possible issues of decreased levels of differentiation with closely related species, which can lead to misinterpretation of species. However, this can be circumvented by taking advantage of the immense volume of sequence data that can be produced with NGS; increasing the number of shorter fragments amplified for each species can compensate for the shorter fragment length (Bybee et al., 2011). Additionally, newer platform developers (such as PacBio and Oxford Nanopore Technologies) are focusing on improving the ability to sequence longer fragments, and therefore diminishing this problem in foodstuffs where DNA of sufficient length can be extracted. Another important limitation to take into account is the possible presence of conserved fragments of DNA that might result in finding sequence matches with organisms not present (for example if the database used was incomplete, or the analysis method insufficiently robust), as well as Nuclear Mitochondrial Translocations (NUMTs) or pseudogenes (H. Song, Buhay, Whiting, & Crandall, 2008). These issues are highly variable, with factors such as taxon and genome size. However, awareness of the problem, and the application of appropriate simple measures for the avoidance of NUMT co-amplification or preferred amplification, combined with appropriate bioinformatics tools, can considerably increase the confidence in the mitochondrial origin of any mtDNA-like sequence, effectively accounting for this issue (Calvignac, Konecny, Malard, & Douady, 2011).

Quantification of the DNA intended for Next Generation Sequencing procedures is also extremely important, since different quantification strategies differentially affect the final results (Robin, Ludlow, LaRanger, Wright, & Shay, 2016). One of the most standardised quantification methods is the Qubit® dsDNA High Sensitivity Assay Kit (ThermoFisher Scientific, USA). It is expected that in the near future, considerable efforts will be devoted to the development of new more accurate and robust quantification methodologies that improve the final outcome, such as All-Food-Seq (AFS) (Liu et al., 2017). Even with these expected improvements in quantification, the inherent limitations of some food matrices (the low amounts and fragmented nature of DNA, the presence of inhibitors) means that quantification of taxa present should not be a mandatory requirement of NGS food authenticity testing.

Another field in which great progress is expected is in the amplification step. Some NGS approaches require the preparation of libraries in which DNA or RNA fragments are coupled to adapters to allow PCR amplification and sequencing. This PCR step can introduce biases since GC- or AT-rich fragments amplify with less efficiency, potentially resulting in a decreased frequency or even an absence of GC- or AT-rich sequences in the subsequent data. PCR-free procedures have been proposed, but they require a large amount of starting material. The development of robust library preparation methods or new amplification strategies able to produce a representative nucleic acid material is a crucial hurdle to obtain results that are free from amplification and sequencing bias (van Dijk, Jaszczyszyn, & Thermes, 2014).

The development of new amplification strategies for library preparation should benefit from the massive growth of public sequence databases which currently contain well curated records for tens of thousands of species (see above). These will promote the design of universal primers to produce amplified products in Next Generation Sequencing strategies. This is particularly important in the case of plant material, since poly- and aneuploid genomes are likely to contain multiple orthologous copies of the target, and therefore require a major effort in the optimization of the primer sequences. In terms of the relative quantification of ingredients present in a sample, it is still difficult to establish relationships between DNA measurement and ingredient content. Future improvements in DNA extraction and amplification procedures may overcome these barriers, enabling the accurate quantification of ingredients (and adulterants).

5.2. Appropriate use of NGS

As previously highlighted, current methodologies such as ELISA (enzyme-linked immunosorbent assay) or PCR-based tests require a prior knowledge of which ingredients might be present, hampering their use as screening tools for unsuspected ingredients. Commonly used kits are limited to the detection of a fixed panel of species. Furthermore, the cost of detecting all potential fraudulent ingredients is prohibitive to consider performing predictive controls. However, NGS can scan for thousands of species simultaneously, and these techniques will evolve to allow the identification of virtually any possible species present in a complex matrix of unlimited ingredients (Muñoz-Colmenero, Martínez, Roca, & Garcia-Vazquez, 2017; Ripp et al., 2014). Furthermore, when dealing with complex matrices, classic Sanger sequencing fails to obtain useful data (without additional time consuming cloning steps), while in NGS, each single molecule is sequenced independently facilitating assessment of complex mixtures (Burns et al., 2016).

Databases encompassing sequences generated by NGS are growing exponentially, increasingly allowing the possibility of reaching genera and species level both from microbiology as well as mammals, birds, reptiles, vegetables, algae, etc. Moreover, sequencing of several genes would also provide enough resolution to identify all organisms present to species level. The integration of large datasets and access to common databases, as well as tools to enable their simple and easy use, are critical.

NGS is already widely used in other areas of food security. The use of WGS-based characterisation of foodborne pathogens is expanding very quickly, replacing more classical techniques. Metagenomics is also allowing the detection and identification of non-culturable pathogens (Bergholz, Moreno Switt, & Wiedmann, 2014). WGS can be used to produce draft genome sequences of the bacteria responsible for foodborne alerts, which allows tracking to identify contamination sources. Therefore, NGS techniques are becoming essential in the management of the clinical characterisation of foodborne pathogens (Allard et al., 2016; Schmedes, Sajantila, & Budowle, 2016). NGS has proven to be extremely useful in understanding food spoilage, identifying bacterial communities present in food items as they undergo deterioration (Cauchie et al., 2017; Mayo et al., 2014). This ability can be exploited by food industries when performing shelf-life studies. It is important to take into account that certain problems might arise with the identification of DNA coming from dead microorganisms in heat treated food, leading to false positives in adequately processed and safe products. Furthermore, metatranscriptomics approaches towards functional genomics involving RNA sequencing will help to understand the microbial processes in spoilage. The identification of microbiota related to the environment or the ingredients composing a food commodity can also be used as a flag of their origin or handling, and could eventually be applied to management and control systems (Ottesen, White, Skaltsas, & Newell, 2009; Portillo, Franquès, Araque, Reguant, & Bordons, 2016).

There are a number of authenticity-specific applications that NGS is particularly suited to, which are highlighted in more detail below. For example, the untargeted nature of NGS makes it particularly suitable for detecting unknown targets, such as screening for unauthorised genetic modification events where the vector is unknown. Another feature of NGS, its ability to generate massive amounts of DNA sequence in parallel, makes it well suited to the identification of ingredients in complex foods. This is often based on a metabarcoding approach, where sequences are generated for one or more well characterised genomic regions that are present in all members of the taxonomic groups of interest, but with sequences that are sufficiently different to allow identification of the organism of origin. This allows the identification of all biological ingredients from a DNA extraction of a complex food matrix, e.g. a herb blend, or processed fish product.

6. Current and future applications

Currently the rate of uptake in the application of NGS techniques seems to be split between continents. In the USA, application of food focused NGS is driven by the FDA, whilst in Europe and Asia the biggest advances have been made by the food industry itself, and in particular, those companies who want to employ specialist businesses to perform a complete audit of their supply chains (often with brand image protection in mind). Several commercial sequencing companies, specialising in food testing, are currently offering NGS services to the global food supply chain with the intention of ensuring that incoming raw ingredients are the desired, unadulterated product, therefore helping to combat food fraud. The application of NGS has the power to simultaneously address multiple issues in authenticity and food safety, providing exhaustive controls which may be especially pertinent when dealing with new providers, so ensuring the maintenance of consumer trust and protecting brand reputation.

While the current use of NGS within industry is somewhat limited compared to the more established techniques, there are several areas where the application is making advances. The first of these is to the identification of GMOs and Genetically Modified Micro-organisms (GMMs). Many jurisdictions, including the European Union (EU), legally distinguish between authorized (and therefore legal) and un-authorized (and therefore illegal) GMOs (UGMOs). While some UGMOs are considered to be safe, but do not have approval, other UGMOs are likely to be in the food chain due to an accidental or deliberate release of 'experimental' GMOs from laboratories or field trials (these are considered to be unknown/unsafe). Included in this last category was the identification of GMM contamination in an animal feed additive. In 2014, viable *Bacillus subtilis* spores were found in imported vitamin B2 feed additive placed on the EU market, and identified as UGMM (Barbau-piednoir et al., 2015). This led to the first ever notification for UGMM in the European Rapid Alert System for Food and Feed (RASFF, 2014). Recently conducted research using a combination of NGS sequencing and DNA walking successfully identified samples containing low GMO concentrations, mixed GMO sources, and processed materials (Fraiture et al., 2017), demonstrating the utility of NGS within this area of authenticity testing. The future application of NGS for the identification of GMOs and GMMs looks promising.

However, one issue that still needs to be addressed is when the genetic modification is small; for example, short insertions/deletions; or a change affecting one or a few base pairs, as these changes are not easily distinguishable from changes that could occur naturally. This is the type of biotech organism that is now being generated with techniques such as CRISPR-Cas9. Recent advances using CRISPR have included button mushrooms (Waltz, 2016), apples and potatoes resistant to browning (Waltz, 2015), a herbicide-tolerant maize (Svitashev et al., 2015), and a pig resistant to the porcine reproductive and respiratory syndrome virus (Whitworth et al., 2016). Within the last five years, over 30 GMOs have been ruled as not qualifying as needing regulation (Waltz, 2016). This is because the organisms do not contain foreign DNA from 'plant pests' such as viruses or bacteria that used to be required for the gene editing process. This was relevant when the regulatory guidelines were developed in the 1980s and 1990s, when the US government developed its framework for regulating GMOs (Wolt, Wang, & Yang, 2016). However, as the ethical and safety implications of these new gene editing technologies are still under debate, and public acceptance is still uncertain, the identification of such organisms in the future may still be required (Baltimore et al., 2015; Court of Justice of the European Union, 2018).

The other main area in which NGS screening is currently used is for the analysis of complex food matrices, when a complete picture of the ingredients (and all contaminants) is required. While this is the area for which NGS is most used by industry, there are currently few published studies. However, the studies outlined below demonstrate the range and complexity of matrices that can be successfully screened. The originating plant species, the entomological source, and the most likely geographical origin of honey (including pasteurised samples) have recently been tested using NGS. The first of these studies found mixed origin for some monofloral samples, and honey with a supposed temperate source containing DNA from Asian plant species, suggesting that the honey had been fraudulently diluted with a cheaper substitute (Prosser & Hebert, 2017). A second study also identified potential authenticity issues with a monofloral honey that did not have the expected plant species as its main constituent, and a Chilean honey that contained pollen from Australian plant species (Utzeri et al., 2018). As with other methods, while pollen can be filtered out to try and confuse testing laboratories, the pollen-free DNA in the liquid fraction of honey cannot be removed by filtration, hence evidence of the original plant and insect sources will likely be detectable.

Studies have also been carried out to determine the effectiveness of NGS for identifying multiple meat species. One study showed NGS to be highly effective at simultaneously identifying multiple species with the primers used (11 mammal and bird, plus rat and human as potential contaminants). However, it should be noted that this study was only carried out on mixtures of DNA, rather than on DNA extracted from meat products, hence the effects of processing, and of other ingredients are unknown (Bertolini et al., 2015). Another study using similar methods investigated dairy products (mixed cow/goat milk samples, buffalo mozzarella cheese, goat crescenza cheese and ricotta cheese), found low levels of sheep DNA in goat milk and cheese, and suggested that this was due to the artisanal nature of the production facilities rather than fraudulent activity. While the detection levels they set were relatively low (0.2%), they stated that their methods were not quantitative. The ability for NGS methods to be quantitative is still being evaluated, however, for dairy products it is particularly unlikely due to the pasteurisation process, and the effect on the number of somatic cells in the product (Ribani et al., 2018).

Other recent studies have targeted complex food matrices. A 2016 study looked at *bacalhau* (dried cod) fish cakes in Brazil, and identified mislabelling rates of 41% within the products sampled. They also found that mislabelling was less frequently found within products possessing a governmental certification stamp (4.5%) (Carvalho, Palhares, Drummond, & Gadanho, 2017). A further study in 2017 developed a primer multiplex which could identify fish and cephalopod species (89 different species widely used for surimi production were initially tested). This was then further tested on 16 surimi products from Europe and Asia, and DNA was identified from 16 species of fish (across 13 families), and from three species of cephalopods. Mislabelling incidences were higher in samples produced in Asia than in Europe, with 37.5% of the samples mislabelled, including mislabelling of allergy related species included in the products, but not identified on the packaging. In general the products produced in Asia contained a higher number of species, including species not normally used for human consumption (Giusti, Armani, & Sotelo, 2017). In a 2016 paper, highly processed fishmeal was tested to determine its constituent species (Galal-Khallaf, Osman, Carleos, Garcia-Vazquez, & Borrell, 2016). This

is not currently an authenticity issue, as species composition is not a legal requirement for the labelling. However, as part of the effort to increase the sustainability of fisheries and aquaculture, this is likely to change in the future. The fishmeal industry relies greatly on wild capture fisheries. As a result, about 21 million tonnes (13%) of the seafood harvested from the wild is for non-human consumption (2014 figures); of this about 76% is made into fishmeal or fish oil (FAO, 2016). Fishmeal is also produced from fish processing wastes, which is a more sustainable option. Fishmeal is prepared by cooking, pressing, drying and grinding whole fishes and their bones and offal from processed fish to produce meal with natural high-quality proteins. To lay the groundwork to enable traceability and transparency for the aqua-feed industry, it is essential to be able to identify the fish species that have been utilised in its production. The study successfully amplified DNA from all seven fish-feed samples, and interestingly, of the 13 fish species that were detected, approximately 46% are classified as either over-exploited or suffered from strong decline.

Whilst not food or feed, NGS methods have also been used to look at mislabelling rates in herbal supplements (Ripp et al., 2014) and traditional Chinese medicines (plant and animal) (Coghlan et al., 2012) finding high mislabelling rates (100% and 78% respectively). For the traditional Chinese medicine samples, 68 different plant families were identified, which included genera, such as *Ephedra* and *Asarum*, which are potentially toxic. Similarly, animal species were identified that are classified by the IUCN as vulnerable, endangered, or critically endangered, including Asiatic black bear (*Ursus thibetanus*) and Saiga antelope (*Saiga tatarica*). Bovidae, Cervidae, and Bufonidae DNA was also detected in many of the samples although not declared on the product packaging. The authors highlight that while NGS methods are highly promising for carrying out untargeted screening of complex samples, caution also needs to be exercised, as the genetic profile of a sample may not represent the actual composition of a complex, highly processed item, as no methods are able to detect DNA when it has been completely degraded (for example by processing procedures). Results from NGS should therefore be regarded as a qualitative, and potentially incomplete assessment of composition rather than a quantitative measure of each ingredient (Coghlan et al., 2012). The large number of species identified in the herbal supplements were identified as occurring not only by intended or non-intended substitution, but also by possible cross-contamination with trace plant DNA occurring at any stage during the growing, harvesting, manufacturing, handling or laboratory analysis of plant material. Hence the detection of such non-target DNA is not always indicative of deliberate adulteration; and such results should be interpreted with caution, especially when legal ramifications are considered. In order to determine the relative levels of undeclared species present in such mixtures, qualitative NGS results could be followed up with standard quantitative PCR using, when available, the level of the declared species as the reference. Ripp et al. (2014) also detected the presence of fungal DNA in 14 out of 15 tested supplements. These included a large number of species which can be categorised as pathogenic, endophytic and mycorrhizal fungi naturally associated with live plant material; saprophytic fungi which had proliferated during drying and storage; and strains involved in the fermentation during manufacturing of bioactive components. This study identified nearly 120 species from some samples, and noted that while NGS is efficient at identifying species, interpretation of test results should focus on potential mycotoxin-producing fungi and human pathogens.

A further study in 2017 sequenced DNA extracted from confectionary, used as a test of the methodology to identify DNA from highly processed samples. This found several animal species including pig, cow, water buffalo, sheep, chicken, wild turkey, several fish species, a Noctuid pest of maize (*Sesamia nonagrioides),* and human (thought to be external contamination) (Muñoz-Colmenero, Martínez, Roca, & Garcia-Vazquez, 2016). These studies demonstrate the ability to extract sequencing quality DNA from highly processed matrices that have been subjected to high temperatures and/or pressure (Muñoz-Colmenero et al., 2016). It should also be noted that as shown in several of these examples, the inclusion of more than one universal locus has helped to overcome potential technical problems associated with amplification biases against some species or families (Bertolini et al., 2015; Prosser & Hebert, 2017; Ribani et al., 2018). This is particularly important for plants species.

With the authenticity of food products in the spotlight, food industries and control laboratories must make use of all available techniques to verify their products and control their ingredient providers. Combined with the regulatory environment surrounding food safety and authenticity becoming more rigorous, this is increasing the importance of routine high-throughput control techniques. NGS is a strong candidate to become a valuable aid or even the technology of choice to achieve regulatory compliance and reputation protection in a number of food fraud situations, particularly for highly complex food matrices.

The size and cost of NGS platforms are decreasing, allowing easier uptake by public health laboratories. Although currently NGS requires highly specialized equipment and substantially skilled staff to analyse the data, it is unlikely that the sequencing revolution has finished. Future developments are likely to target ease of application and data extraction, allowing widespread application beyond the current user base.

The optimal way to take advantage of the use of NGS for the benefit of the consumer would be its incorporation into routine analytical checks for authenticity/contamination at several levels of the food supply:

1. As a control laboratory: the application of these technologies will help ensure the quality and integrity of checked products. This will minimize the costs related to the entrance of non-compliant products into the food chain.
2. As a producer: the creation of NGS self-certification systems. The declaration that their materials and products meet the required standards of quality and safety would be extremely useful in added-value creation and differentiation from competitors.
3. As a retailer: for example; conducting specific studies to verify new suppliers. Also, the selection of only approved or certified suppliers certifying their product could help to confirm conformance to specification parameters. This would provide solid evidence of their real commitment to food integrity.

This way it will be possible to reach a high level of consumer protection, as well as to build trust between consumers and retailers. One criticism on the use of NGS is that it is too data intensive, and that for certain, specific questions relating to simple, unprocessed foods (e.g. is this fish fillet cod?) it is more efficient and cost effective to use a targeted approach. However, there are questions related to the composition of complex or processed foods for which the large amounts of data produced by NGS are advantageous. There are additional ways to improve the efficiency of the testing process, such as determining where commodity and geographical hotspots are, so that the types of food and feed product that are susceptible to authenticity issues could be determined (e.g., is it necessary to target all products on the market or to focus on products with certain countries of origin?).

In the near future, as NGS becomes a well-established technique, both methodologies and analysis pipelines should be harmonized among testing laboratories (Endrullat, Glökler, Franke, & Frohme, 2016; Gargis, Kalman, & Lubin, 2016). Technological advances and increased competition will continue to push the field towards lower costs, higher throughput and more user friendly options for analysis (Goodwin et al., 2016). It will also be essential that the competent testing laboratories become appropriately accredited.

Declarations of interest

None.

Acknowledgements

Funding: This work was supported by the European Union's Seventh Framework Programme for research, technological development and demonstration [grant number 613688].

References

Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., et al. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *Journal of Clinical Microbiology, 54*, 1975–1983.

Alnasir, J., & Shanahan, H. P. (2015). Investigation into the annotation of protocol sequencing steps in the sequence read archive. *GigaScience, 4*(1), https://doi.org/10.1186/s13742-015-0064-7.

Ansorge, W. J. (2016). Next generation DNA sequencing (II): Techniques, applications. *Journal of Next Generation Sequencing & Applications, 01*(S1), https://doi.org/10.4172/2469-9853.s1-005.

Arulandhu, A. J., Staats, M., Hagelaar, R., Voorhuijzen, M. M., Prins, T. W., Scholtens, I., ... Kok, E. (2017). Development and validation of a multi-locus DNA metabarcoding method to identify endangered species in complex samples. *GigaScience, 6*(10).

Aziz, N., Zhao, Q., Bry, L., Driscoll, D. K., Funke, B. H., Gibson, J. S., et al. (2015). College of American pathologists' laboratory standards for next-generation sequencing clinical tests. *Archives of Pathology & Laboratory Medicine, 139*(4), 481–493 doi:10.5858/.

Baltimore, D., Berg, P., Botchan, M., Carroll, D., Charo, R. A., Church, G., et al. (2015). A prudent path forward for genomic engineering and germline gene modification. *Science, 348*(6230), 36–38.

Barbau-piednoir, E., De Keersmaecker, S. C. J., Delvoye, M., Gau, C., Philipp, P., & Roosens, N. H. (2015). Use of next generation sequencing data to develop a qPCR method for specific detection of EU-unauthorized genetically modified *Bacillus subtilis* overproducing riboflavin. *BMC Biotechnology, 15*(1), https://doi.org/10.1186/s12896-015-0216-y.

Barnett, J., Begen, F., Howes, S., Regan, A., McConnon, A., Marcu, A., et al. (2016). Consumers' confidence, reflections and response strategies following the horsemeat incident. *Food Control, 59*, 721–730. https://doi.org/10.1016/j.foodcont.2015.06.021.

Bartlett, S. E., & Davidson, W. S. (1992). FINS (forensically informative nucleotide sequencing): A procedure for identifying the animal origin of biological specimens. *Biotechniques, 12*(3), 408–411.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Nucleic Acids Research, 41*(D1), D36–D42. https://doi.org/10.1093/nar/gks1195.

Bergholz, T. M., Moreno Switt, A. I., & Wiedmann, M. (2014). Omics approaches in food safety: Fulfilling the promise? *Trends in Microbiology, 22*(5), 275–281. https://doi.org/10.1016/j.tim.2014.01.006.

Bertolini, F., Ghionda, M. C., D'Alessandro, E., Geraci, C., Chiofalo, V., & Fontanesi, L. (2015). A next generation semiconductor based sequencing approach for the identification of meat species in DNA mixtures. *PLoS One, 10*(4), e0121701. https://doi.org/10.1371/journal.pone.0121701.

Bilofsky, H. S., & Burks, C. (1988). The GenBank genetic sequence data bank. *Nucleic Acids Research, 16*(5), 1861–1863.

Burns, M., Wiseman, G., Knight, A., Bramley, P., Foster, L., Rollinson, S., ... Primrose, S. (2016). Measurement issues associated with quantitative molecular biology analysis of complex food matrices for the detection of food fraud. *Analyst, 141*, 45–61.

Bybee, S. M., Bracken-Grissom, H., Haynes, B. D., Hermansen, R. A., Byers, R. L., Clement, M. J., et al. (2011). Targeted amplicon sequencing (TAS): A scalable next-gen approach to multilocus, multitaxa phylogenetics. *Genome Biology and Evolution, 3*(0), 1312–1323. https://doi.org/10.1093/gbe/evr106.

Calvignac, S., Konecny, L., Malard, F., & Douady, C. J. (2011). Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts). *Mitochondrion, 11*(2), 246–254. https://doi.org/10.1016/j.mito.2010.10.004.

Cao, Y., Fanning, S., Proos, S., Jordan, K., & Srikumar, S. (2017). A review on the applications of next generation sequencing technologies as applied to food-related microbiome studies. *Frontiers in Microbiology, 8*, 1829. https://doi.org/10.3389/fmicb.2017.01829.

Carvalho, D. C., Palhares, R. M., Drummond, M. G., & Gadanho, M. (2017). Food metagenomics: Next generation sequencing identifies species mixtures and mislabeling within highly processed cod products. *Food Control, 80*, 183–186. https://doi.org/10.1016/j.foodcont.2017.04.049.

Catalano, V., Moreno-Sanz, P., Lorenzi, S., & Grando, M. S. (2016). Experimental review of DNA-based methods for wine traceability and development of a single-nucleotide polymorphism (SNP) genotyping assay for quantitative varietal authentication. *Journal of Agricultural and Food Chemistry, 64*(37), 6969–6984. https://doi.org/10.1021/acs.jafc.6b02560.

Cauchie, E., Gand, M., Kergourlay, G., Taminiau, B., Delhalle, L., Korsak, N., et al. (2017). The use of 16S rRNA gene metagenetic monitoring of refrigerated food products for understanding the kinetics of microbial subpopulations at different storage temperatures: The example of white pudding. *International Journal of Food Microbiology, 247*, 70–78. http://doi.org/10.1016/j.ijfoodmicro.2016.10.012.

Coghlan, M. L., Haile, J., Houston, J., Murray, D. C., White, N. E., Moolhuijzen, P., ... 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.

Connor, T. R., Guest, M., Southgate, J., Ismail, M., Bakke, M., Poplawski, R., et al. (2016). CLIMB (the cloud infrastructure for microbial bioinformatics): An online resource for the medical microbiology community. *Microbial Genomics, 2*(9), https://doi.org/10.1099/mgen.0.000086.

Consolandi, C., Palmieri, L., Doveri, S., Maestri, E., Marmiroli, N., Reale, S., et al. (2007). Olive variety identification by ligation detection reaction in a universal array format. *Journal of Biotechnology, 129*(3), 565–574. https://doi.org/10.1016/j.jbiotec.2007.01.025.

Court of Justice of the European Union (2018). *Organisms obtained by mutagenesis are GMOs and are, in principle, subject to the obligations laid down by the GMO Directive.* [Press release].

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16S rRNA gene database and Workbench compatible with ARB. *Applied and Environmental Microbiology, 72*(7), 5069–5072. https://doi.org/10.1128/aem.03006-05.

van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research, 322*(1), 12–20. https://doi.org/10.1016/j.yexcr.2014.01.008.

Divoll, T. J., Brown, V. A., Kinne, J., & McCracken, G. F. (2018). Disparities in second-generation DNA metabarcoding results exposed with accessible and repeatable workflows. *Molecular Ecology Resources, 18*, 590–601. https://doi.org/10.1111/1755-0998.12770.

Douillard, F. P., Kant, R., Ritari, J., Paulin, L., Palva, A., & de Vos, W. M. (2013). Comparative genome analysis of *Lactobacillus casei* strains isolated from Actimel and Yakult products reveals marked similarities and points to a common origin. *Microbial Biotechnology, 6*(5), 576–587. https://doi.org/10.1111/1751-7915.12062.

Druml, B., & Cichna-Markl, M. (2014). High resolution melting (HRM) analysis of DNA – its role and potential in food analysis. *Food Chemistry, 158*, 245–254. https://doi.org/10.1016/j.foodchem.2014.02.111.

Endrullat, C., Glökler, J., Franke, P., & Frohme, M. (2016). Standardization and quality management in next-generation sequencing. *Applied & Translational Genomics, 10*, 2–9. https://doi.org/10.1016/j.atg.2016.06.001.

FAO (2016). *The State of World Fisheries and Aquaculture 2016. Contributing to food security and nutrition for all.* Retrieved from Rome.

Ferri, G., Corradini, B., Ferrari, F., Santunione, A. L., Palazzoli, F., & Alu', M. (2015). Forensic botany II, DNA barcode for land plants: Which markers after the international agreement? *Forensic Science International: Genetics, 15*, 131–136. https://doi.org/10.1016/j.fsigen.2014.10.005.

Fraiture, M.-A., Herman, P., De Loose, M., Debode, F., & Roosens, N. H. (2017). How can we better detect unauthorized GMOs in food and feed chains? *Trends in Biotechnology, 35*(6), 508–517. https://doi.org/10.1016/j.tibtech.2017.03.002.

Fraiture, M.-A., Herman, P., Papazova, N., De Loose, M., Deforce, D., Ruttink, T., et al. (2017). An integrated strategy combining DNA walking and NGS to detect GMOs. *Food Chemistry, 232*, 351–358. https://doi.org/10.1016/j.foodchem.2017.03.067.

Fuller, C. W., Kumar, S., Porel, M., Chien, M., Bibillo, A., Stranges, P. B., et al. (2016). Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. *Proceedings of the National Academy of Sciences, 113*(19), 5233–5238. https://doi.org/10.1073/pnas.1601782113.

Galal-Khallaf, A., Osman, A. G. M., Carleos, C. E., Garcia-Vazquez, E., & Borrell, Y. J. (2016). A case study for assessing fish traceability in Egyptian aquafeed formulations using pyrosequencing and metabarcoding. *Fisheries Research, 174*, 143–150. https://doi.org/10.1016/j.fishres.2015.09.009.

Gargis, A. S., Kalman, L., Berry, M. W., Bick, D. P., Dimmock, D. P., Hambuch, T., et al. (2012). Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nature Biotechnology, 30*(11), https://doi.org/10.1038/nbt.2403 doi:10.1038/nbt.2403.

Gargis, A. S., Kalman, L., & Lubin, I. M. (2016). Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *Journal of Clinical Microbiology, 54*(12), 2857–2865.

Giusti, A., Armani, A., & Sotelo, C. G. (2017). Advances in the analysis of complex food matrices: Species identification in surimi-based products using Next Generation Sequencing technologies. *PLoS One, 12*(10), e0185586. https://doi.org/10.1371/journal.pone.0185586.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next generation sequencing technologies. *Nature Reviews Genetics, 17*, 333–351.

Hardwick, S. A., Deveson, I. W., & Mercer, T. R. (2017). Reference standards for next-generation sequencing. *Nature Reviews Genetics, 18*(8), 473–484. https://doi.org/10.1038/nrg.2017.44.

Hayden, E. C. (2014). Technology: The $1,000 genome. *Nature, 507*, 294–295.

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.

Hodkinson, B. P., & Grice, E. A. (2015). Next-generation sequencing: A review of technologies and tools for wound microbiome research. *Advances in Wound Care, 4*(1), 50–58. https://doi.org/10.1089/wound.2014.0542.

Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., et al. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences, 106*(31), 12794–12797. https://doi.org/10.1073/pnas.0905845106.

ISO (2017a). ISO and food. In ISO (Ed.). https://www.iso.org/publication/PUB100297.html.

ISO (2017b). *ISO/TS 20428:2017 Health informatics – Data elements and their metadata for describing structured clinical genomic sequence information in electronic health records.* https://www.iso.org/standard/67981.html.
Kamle, M., Kumar, P., Patra, J. K., & Bajpai, V. K. (2017). Current perspectives on genetically modified crops and detection methods. *3 Biotech, 7*(3), 219.
Kodama, Y., Shumway, M., & Leinonen, R. (2011). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research, 40*(D1), D54–D56. https://doi.org/10.1093/nar/gkr854.
Laehnemann, D., Borkhardt, A., & McHardy, A. C. (2016). Denoising DNA deep sequencing data—high-throughput sequencing errors and their correction. *Briefings in Bioinformatics, 17*(1), 154–179.
Lefterova, M. I., Suarez, C. J., Banaei, N., & Pinsky, B. A. (2015). Next-generation sequencing for infectious disease diagnosis and management. *Journal of Molecular Diagnostics, 17*(6), 623–634. https://doi.org/10.1016/j.jmoldx.2015.07.004.
Leinonen, R., Sugawara, H., & Shumway, M. (2010). The sequence read archive. *Nucleic Acids Research, 39*, D19–D21. https://doi.org/10.1093/nar/gkq1019 (Database).
Leonard, S. R., Mammel, M. K., Lacher, D. W., Elkins, C. A., & Drake, H. L. (2015). Application of metagenomic sequencing to food safety: Detection of shiga toxin-producing Escherichia coli on fresh bagged spinach. *Applied and Environmental Microbiology, 81*(23), 8183–8191. https://doi.org/10.1128/aem.02601-15.
Liu, Y., Ripp, F., Koeppel, R., Schmidt, H., Hellmann, S. L., Weber, M., et al. (2017). AFS: Identification and quantification of species composition by metagenomic sequencing. *Bioinformatics, 33*(9), 1396–1398. https://doi.org/10.1093/bioinformatics/btw822.
Madesis, P., Ganopoulos, I., Sakaridis, I., Argiriou, A., & Tsaftaris, A. (2014). Advances of DNA-based methods for tracing the botanical origin of food products. *Food Research International, 60*, 163–172. https://doi.org/10.1016/j.foodres.2013.10.042.
Mahamdallie, S., Ruark, E., Yost, S., Münz, M., Renwick, A., Poyastro-Pearson, E., et al. (2018). The quality sequencing minimum (QSM): Providing comprehensive, consistent, transparent next generation sequencing data quality assurance. *Wellcome Open Research, 3*.
Mayo, B., Rachid, C. T. C. C., Alegria, A., Leite, A. M. O., Peixoto, R. S., & Delgado, S. (2014). Impact of next generation sequencing techniques in food microbiology. *Current Genomics, 15*, 293–309.
Mezzasalma, V., Sandionigi, A., Bruni, I., Bruno, A., Lovicu, G., Casiraghi, M., et al. (2017). Grape microbiome as a reliable and persistent signature of field origin and environmental conditions in Cannonau wine production. *PLoS One, 12*(9), e0184615. https://doi.org/10.1371/journal.pone.0184615.
Minot, S. S., Krumm, N., & Greenfield, N. B. (2015). One Codex: A sensitive and accurate data platform for genomic microbial identification. *bioRxiv,* 027607. https://doi.org/10.1101/027607.
Moran-Gilad, J., Sintchenko, V., Pedersen, S. K., Wolfgang, W. J., Pettengill, J., Strain, E., et al. (2015). Proficiency testing for bacterial whole genome sequencing: An end-user survey of current capabilities, requirements and priorities. *BMC Infectious Diseases, 15*(1), https://doi.org/10.1186/s12879-015-0902-3.
Muñoz-Colmenero, M., Martínez, J. L., Roca, A., & Garcia-Vazquez, E. (2017). NGS tools for traceability in candies as high processed food products: Ion Torrent PGM versus conventional PCR-cloning. *Food Chemistry, 214*, 631–636. http://doi.org/10.1016/j.foodchem.2016.07.121.
Muñoz-Colmenero, M., Martínez, J. L., Roca, A., & Garcia-Vazquez, E. (2016). Detection of different DNA animal species in commercial candy products. *Journal of Food Science, 81*(3).
NCBI (2018a). *GenBank and WGS statistics.* Retrieved from https://www.ncbi.nlm.nih.gov/genbank/statistics/on 01/10/2018.
NCBI (2018b). *RefSeq growth statistics.* Retrieved from https://www.ncbi.nlm.nih.gov/refseq/statistics/on 01/10/2018.
NCBI (2018c). *SRA database growth.* Retrieved from https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/on 01/10/2018.
Nielsen, E. E., Cariani, A., Mac Aoidh, E., Maes, G. E., Milano, I., Ogden, R., et al. (2012). Gene-associated markers provide tools for tackling illegal fishing and false eco-certification. *Nature Communications, 3*, 851. https://doi.org/10.1038/ncomms1845.
Nietsch, R., Haas, J., Lai, A., Oehler, D., Mester, S., Frese, K. S., et al. (2016). The role of quality control in targeted next-generation sequencing library preparation. *Genomics, Proteomics & Bioinformatics, 14*(4), 200–206. https://doi.org/10.1016/j.gpb.2016.04.007.
O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research, 44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189.
Ottesen, A. R., White, J. R., Skaltsas, D. N., & Newell, M. J. (2009). Impact of organic and conventional management on the phyllosphere microbial ecology of an apple crop. *Journal of Food Protection, 72*.
Pardo, M. A. (2015). Evaluation of a dual-probe real time PCR system for detection of Mandarin in commercial orange juice. *Food Chemistry, 172*(0), 377–384. https://doi.org/10.1016/j.foodchem.2014.09.096.
Park, D., Kim, D., Jang, G., Lim, J., Shin, Y.-J., Kim, J., et al. (2015). Efficiency to discovery transgenic loci in GM rice using next generation sequencing whole genome re-sequencing. *Genomics & Informatics, 13*(3), 81–85. https://doi.org/10.5808/gi.2015.13.3.81.
Petrillo, M., Angers-Loustau, A., Henriksson, P., Bonfini, L., Patak, A., & Kreysa, J. (2015). JRC GMO-amplicons: A collection of nucleic acid sequences related to genetically modified organisms. *Database-the Journal of Biological Databases and Curation, 11*. https://doi.org/10.1093/database/bav101.
Portillo, M. d. C., Franquès, J., Araque, I., Reguant, C., & Bordons, A. (2016). Bacterial diversity of Grenache and Carignan grape surface from different vineyards at Priorat wine region (Catalonia, Spain). *International Journal of Food Microbiology, 219*, 56–63. http://doi.org/10.1016/j.ijfoodmicro.2015.12.002.
Primrose, S., Woolfe, M., & Rollinson, S. (2010). Food forensics: Methods for determining the authenticity of foodstuffs. *Trends in Food Science & Technology, 21*(12), 582–590. https://doi.org/10.1016/j.tifs.2010.09.006.
Prosser, S. W. J., & Hebert, P. D. N. (2017). Rapid identification of the botanical and entomological sources of honey using DNA metabarcoding. *Food Chemistry, 214*, 183–191. https://doi.org/10.1016/j.foodchem.2016.07.077.
Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et al. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research, 35*(21), 7188–7196. https://doi.org/10.1093/nar/gkm864.
Randhawa, G. J., Chhabra, R., Bhoge, R. K., & Singh, M. (2015). Visual and real-time event-specific loop-mediated isothermal amplification based detection assays for Bt cotton events MON531 and MON15985. *Journal of AOAC International, 98*(5), 1207–1214.
RASFF (2014). *Rapid Alert System for Food and Feed) Notification details- 2014.1249 unauthorised genetically modified (Bacillus subtilis) bacteria in vitamin B2 from China, via Germany.* 2014. Retrieved from.
Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The barcode of life data system (www.barcodinglife.org). *Molecular Biology Notes, 7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2006.01678.x.
Regulation (EU) (2011). *No 1169/2011 of the European Parliament and of the council of 25 October 2011 on the provision of food information to consumers*.
Ren, J., Deng, T., Huang, W., Chen, Y., & Ge, Y. (2017). A digital PCR method for identifying and quantifying adulteration of meat species in raw and processed food. *PLoS One, 12*(3), e0173567. https://doi.org/10.1371/journal.pone.0173567.
Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics, 13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002.
Ribani, A., Schiavo, G., Utzeri, V. J., Bertolini, F., Geraci, C., Bovo, S., et al. (2018). Application of next generation semiconductor based sequencing for species identification in dairy products. *Food Chemistry, 246*, 90–98. https://doi.org/10.1016/j.foodchem.2017.11.006.
Ripp, F., Krombholz, F., Liu, Y., Weber, M., Schafer, A., Schmidt, B., et al. (2014). All-food-seq (AFS): A quantifiable screen for species in biological samples by deep DNA sequencing. *BMC Genomics, 15*(639).
Robin, J. D., Ludlow, A. T., LaRanger, R., Wright, W. E., & Shay, J. W. (2016). Comparison of DNA quantification methods for next generation sequencing. *Scientific Reports, 6*. https://doi.org/10.1038/srep24067.
Roumpeka, D. D., Wallace, R. J., Escalettes, F., Fotheringham, I., & Watson, M. (2017). A review of bioinformatics tools for bio-prospecting from metagenomic sequence data. *Frontiers in Genetics, 8*. https://doi.org/10.3389/fgene.2017.00023.
Schmedes, S. E., Sajantila, A., & Budowle, B. (2016). Expansion of microbial forensics. *Journal of Clinical Microbiology, 54*, 1964–1974.
Shehata, H. R., Li, J., Chen, S., Redda, H., Cheng, S., Tabujara, N., et al. (2017). Droplet digital polymerase chain reaction (ddPCR) assays integrated with an internal control for quantification of bovine, porcine, chicken and Turkey species in food and feed. *PLoS One, 12*(8), e0182872. https://doi.org/10.1371/journal.pone.0182872.
Song, H., Buhay, J. E., Whiting, M. F., & Crandall, K. A. (2008). Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are co-amplified. *Proceedings of the National Academy of Sciences of the United States of America, 105*(36), 13486–13491.
Song, L., Huang, W., Kang, J., Ren, H., & Ding, K. (2017). Comparison of error correction algorithms for ion torrent PGM data: Application to hepatitis B virus. *Scientific Reports, 7*, 8106. https://doi.org/10.1038/s41598-017-08139-y.
Staats, M., Arulandhu, A. J., Gravendeel, B., Holst-Jensen, A., Scholtens, I., Peelen, T., et al. (2016). Advances in DNA metabarcoding for food and wildlife forensic species identification. *Analytical and Bioanalytical Chemistry, 408*, 4615–4630. https://doi.org/10.1007/s00216-016-9595-8.
Svitashev, S., Young, J. K., Schwartz, C., Gao, H., Falco, S. C., & Cigan, A. M. (2015). Targeted mutagenesis, precise gene editing, and site-specific gene insertion in maize using Cas9 and guide RNA. *Plant Physiology, 169*(2), 931–945. https://doi.org/10.1104/pp.15.00793.
Taylor, M., Fox, C., Rico, I., & Rico, C. (2002). Species-specific TaqMan probes for simultaneous identification of (*Gadus morhua* L.), haddock (*Melanogrammus aeglefinus* L.) and whiting (*Merlangius merlangus* L.). *Molecular Ecology Resources, 2*(4), 599–601. https://doi.org/10.1046/j.1471-8278.
Utzeri, V. J., Ribani, A., Schiavo, G., Bertolini, F., Bovo, S., & Fontanesi, L. (2018). Application of next generation semiconductor based sequencing to detect the botanical composition of monofloral, polyfloral and honeydew honey. *Food Control, 86*, 342–349. https://doi.org/10.1016/j.foodcont.2017.11.033.
Waltz, E. (2015). USDA approves next-generation GM potato. *Nature Biotechnology, 33*, 12. https://doi.org/10.1038/nbt0115-12.
Waltz, E. (2016). Gene-edited CRISPR mushroom escapes US regulation. *Nature, 532*(7599), 293.
Ward, R. D. (2012). FISH-BOL, a case study for DNA barcodes. In W. J. Kress, & D. L. Erickson (Eds.). *DNA barcodes: Methods and protocols* (pp. 423–439). Totowa, NJ: Humana Press.
Whitworth, K. M., Rowland, R. R. R., Ewen, C. L., Trible, B. R., Kerrigan, M. A., Cino-Ozuna, A. G., et al. (2016). Gene-edited pigs are protected from porcine reproductive and respiratory syndrome virus. *Nature Biotechnology, 34*(1), 20–22.
Wilkinson, S., Archibaled, A. L., Haley, C. S., Megens, H.-J., Crooijmans, R. P. M. A., Groenen, M. A. M., et al. (2012). Development of a genetic tool for product regulation in the diverse British pig breed market. *BMC Genomics, 13*.
Wolt, J. D., Wang, K., & Yang, B. (2016). The regulatory status of genome-edited crops. *Plant Biotechnology Journal, 14*(2), 510–518. https://doi.org/10.1111/pbi.12444.
Yang, L., Wang, C., Holst-Jensen, A., Morisset, D., Lin, Y., & Zhang, D. (2013).

Characterization of GM events by insert knowledge adapted re-sequencing approaches. *Scientific Reports, 3*. https://doi.org/10.1038/srep02839.

Ye, J., Feng, J., Dai, Z., Meng, L., Zhang, Y., & Jiang, X. (2016). Application of loop-mediated isothermal amplification (LAMP) for rapid detection of Jumbo Flying squid *Dosidicus gigas* (D'orbigny, 1835). *Food Analytical Methods*. https://doi.org/10.1007/s12161-016-0700-6.

Zhou, Q., Su, X., & Ning, K. (2014). Assessment of quality control approaches for metagenomic data analysis. *Scientific Reports, 4*, 6957. https://doi.org/10.1038/srep06957.