JOHANNES GUTENBERG UNIVERSITÄT MAINZ

All-Food-Sequencing: Identification and quantification of food ingredients by whole-genome metagenomics

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Sören Lukas Hellmann

geb. am 13.01.1987 in Speyer

Mainz, 2026

Dean: Prof. Dr. Eckhard Thines

First examiner: Prof. Dr. Thomas Hankeln

Second examiner: Prof. Dr. Miguel A. Andrade-Navarro

Date of defense:

Table of contents

List of figures

List of tables

1 Introduction

1.1 Significance of species identification in food safety and quality control

Species identification plays a crucial role in safeguarding food quality and safety. As global food supply chains become increasingly complex, the risks associated with food fraud and mislabelling have escalated, posing significant threats to public health and economic integrity. The topic has gained increasing public interest due to scandals such as the "horse meat scandal" in Europe in 2013, where food products labelled as beef contained up to 100 % horse meat (European Parliament & Council, 2013a; O'Mahony, 2013). A recent report by the European commission observed that nearly half of the honey samples (46 %) were not complying with the provisions, e.g. by addition of syrups (European Parliament & Council, 2023). These findings were supported by an analysis commissioned by *Deutsche Berufs- und Erwerbsimkerbund und der Europäische Berufsimkerbund,* which revealed that more than 80 % of honey samples taken from German supermarkets were fraudulent (Koch, 2024). In addition to substitutions during the preparation of food products, contamination also plays an important role, highlighted by the outbreak of enterohaemorrhagic Escherichia coli in Germany in 2011 (Burger, 2012) and the outbreak of Listeria monocytogenes in Germany in 2018 (Robert Koch Institute, 2019). According to a report of the European Parliament, recent food fraud and contamination cases have damaged consumer trust in the agro-food sector as a whole (European Parliament & Council, 2013b).

Table 1: Overview of selected European food adulteration and species substitution over the last 15 years, classified by product category, year, declaration, analytical findings, and substitution pattern.

Product category	**Year**	**Declaration**	**Detection**	**Type of substitution**	**Reference**
processed meat	2013	cattle (*Bos taurus*)	horse (*Equus caballus*)	undeclared addition or replacement of horse meat in beef products	European Parliament & Council, 2013a
processed fish	2014	Atlantic cod (*Gadus morhua*)	pacific cod (*Gadus macrocephalus*)	species substitution with lower quality cod in 10.5 % of tested products; additional substitutions detected in lower frequency	Helyar et al., 2014
processed fish	2015 - 2016	dusky grouper (*Mycteroperca marginatus*), butterfish (*Pholis gunnellus*), pike-perch (*Sander lucioperca*), sole (*Solea solea*), bluefin (*Thunnus thynnus*), yellowfin tuna (*Thunnus albacares*)	hake (*Merluccius spp.*), cod (*Gadus spp.*), haddock (*Melanogrammus aeglefinus*), swordfish (*Xiphias gladius*)	species substitution of mostly high-value fish with cheaper species; 26 % of total samples mislabelled	Pardo et al., 2018
fish fillet	2020	Atlantic cod (*Gadus morhua*)	pacific cod (*Gadus macrocephalus*), haddock (*Melanogrammus aeglefinus*)	undeclared species substitution in 25 % of samples in France; German and Durch samples labelled correctly	Feldmann et al., 2021
herbs & spices	2021	oregano (*Origanum vulgare L.*), cumin (*Cuminum cyminum*), pepper (*Piper nigrum*)	olive leaves (*Olea europaea*), peanut (*Arachis hypogaea*), almond (*Prunus dulcis*), papaya (*Carica papaya*)	partial substitution: 1) shredded olive leaves instead of oregano, 2) nut shells instead of cumin, 3) papaya seeds instead of pepper	Maquet et al., 2021
honey	2023	honey	sugar syrup	addition of sugar syrups for dilution of honey	European Parliament & Council, 2023

Mislabelling of food products can lead to the consumption of allergens, toxins, or species that are otherwise unsuitable or illegal to consume (Figure 1). Ethical (e.g. halal, kosher) and lifestyle aspects (e.g. vegetarian, vegan, gluten-free) of nutrition must also be taken into account. In Germany, more than 100,000 people reportedly fall ill every year from foodborne infections, with the number of unreported cases being much higher (German Federal Institute for Risk Assessment, 2023). Food retailers and food control authorities must be able to verify the identity and quality of the goods supplied. In the interests of consumer protection and to prevent unfair competition, it is therefore necessary to have standardized methods available for the unambiguous identification of biological species.

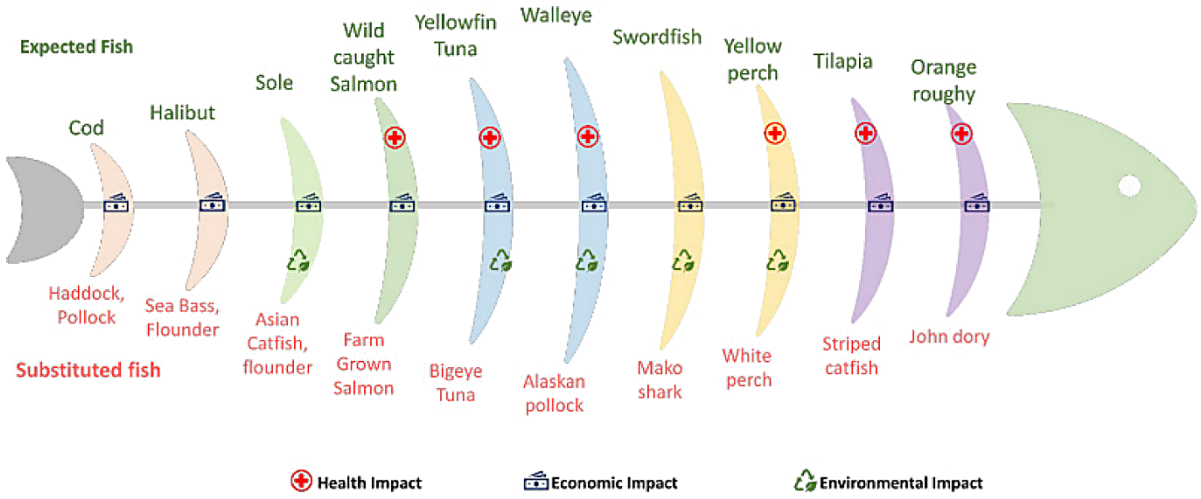

Figure 1: Commonly substituted fish species and influence on consumers (adapted from Cermakova et al., 2023)

As a response to the increasing phenomenon of adulterated and substandard food products, Interpol and Europol initiated the "Operation Opson" in 2011. A key aim of these operations is the identification of organized criminal networks behind illicit trade of counterfeit food (Europol, 2014). The "Opson" operations are carried out annually and are assisted by the food control authorities of several countries worldwide. "Operation Opson V" focused on the trade of Asian fish and the risk of species

substitution. More than 300 tons of meat and meat products, as well as over 900 tons of seafood, fish and fish products have been seized during this operation (Interpol, 2016). With a per capita consumption of 14.1 kg in 2020, fish, crustaceans, and molluscs, as well as products made from them, are popular foods in Germany. 89 % of the fishery products consumed in Germany are imported, and due to widely ramified trade routes and the great diversity of the products, they are particularly susceptible to possible fraud (Federal Office of Consumer Protection and Food Safety, 2022). Again, reliable species identification is thus necessary, especially for seafood species, due to the high commercial interest of this organismal group.

1.2 Methods for species identification in food products

Foodomics is an integrative discipline that applies molecular and computational tools to study food composition to enhance food authenticity, safety and health (Balkir et al., 2021). Within this framework, DNA-based methods have become a central pillar enabling precise species identification, traceability, and detection of adulteration in complex food matrices (Griffiths et al., 2014). As the proportion of processed and ultra-processed foods is increasing, morphological examination is becoming less reliable and is replaced by molecular approaches (Böhme et al., 2019; Cermakova et al., 2023; Danezis et al., 2016; Iammarino et al., 2017). Thus, all analytical methods recommended for animal species authentication in the official collection of examination procedures are either protein- or DNA-based (Bundesamt für Verbraucherschutz und Lebensmittelsicherheit, 2014).

DNA-based methods offer several advantages over protein-based methods: Firstly, these methods are independent of sample origin and developmental stage. While proteins are often only found in certain tissues or at a certain point in developmental time (Afzaal et al., 2022), DNA is universally present in all species, all tissues, and at all

stages (Filipa-Silva et al., 2024; Piskata et al., 2019). Secondly, DNA offers more information, facilitating insights about populations of origin, especially when looking at whole-genome information (Higgins et al., 2021; C. Zhao et al., 2023) Thirdly, DNA is more suitable for processed samples due to the higher thermostability of DNA in contrast to proteins (Prado et al., 2016; Senyuva et al., 2019; Teletchea et al., 2005). Fourthly, DNA offers another layer of information in the form of epigenetic DNA-modifications: For example, tissue-specific methylation profiles allow the differentiation of tissue types in salmon including brain, eye, liver, and muscle, as well as in veal, including heart, kidney, liver, and muscle (Rodríguez López et al., 2012).

1.2.1 Polymerase chain reaction-based methods

Polymerase chain reaction (PCR) is a fundamental technique for the amplification of DNA sequences, and it is central to genetic analysis and species identification. The technique uses a DNA polymerase to synthesize a complementary DNA strand from a template, using primers (short synthetic oligonucleotides) that flank a target region. Through repeated cycles of denaturation (melting the DNA double helix), annealing (binding of primers), and extension (synthesis of the new, complementary DNA strand), PCR exponentially increases the number of copies of the targeted DNA segment (Figure 2).

In food analysis, PCR is widely employed for species identification, as it allows for the selective amplification of species-specific markers, allowing even trace amounts of template DNA to yield detectable quantities sufficient for detection (Fanelli et al., 2021). As a targeted approach, PCR relies on the prior knowledges of DNA-sequences which will serve as the primer binding sites, which define the specificity and scope of amplification: Species-specific PCR assays are designed to detect DNA sequences unique to a single species. Primers in these assays target diagnostic nucleotide motifs that are absent in non-target species, producing a binary result: the presence of an

amplicon indicates the presence of the target species, while its absence suggests that the species is not present in the sample. Such assays are useful in authenticity testing where the goal is to confirm or exclude the presence of a declared ingredient, e.g., detection of pork DNA in kosher or halal products or identification of specific fish species in seafood mixtures (Muflihah et al., 2023; Ulca et al., 2013).

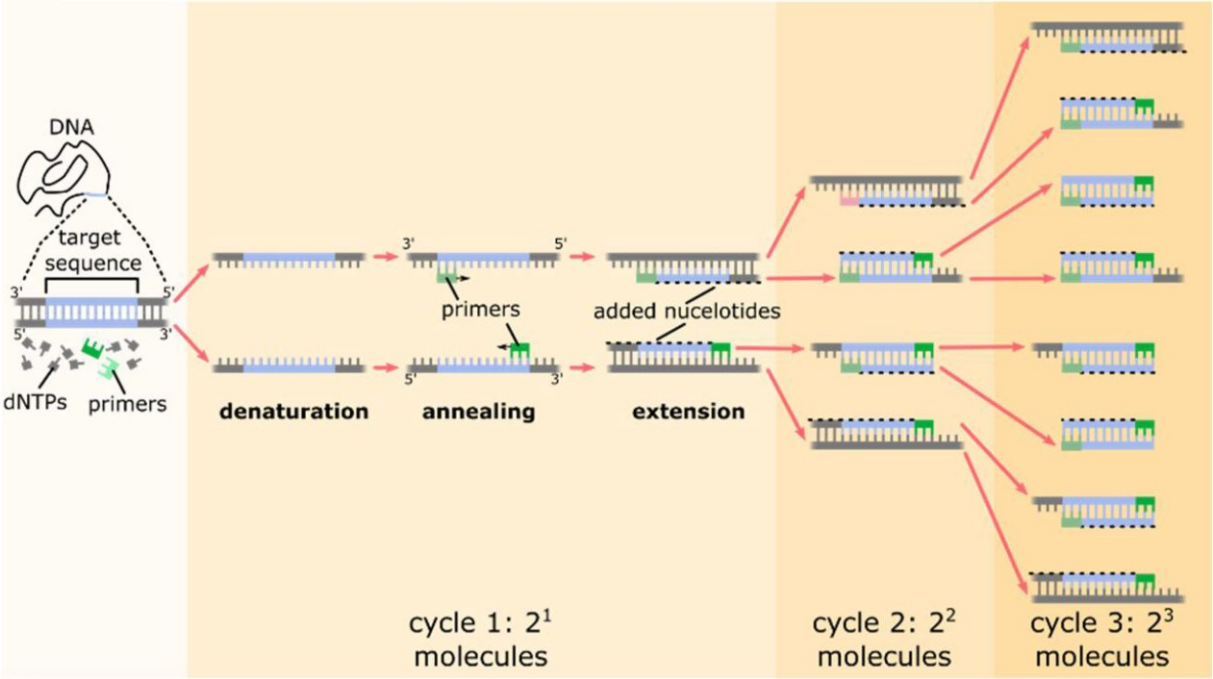

Figure 2: Principles of PCR. Each cycle consists of three steps: (1) thermal denaturation of double-stranded DNA, (2) primer annealing to complementary target sites, and (3) primer extension by a thermostable DNA polymerase. Typical reactions run for 20–40 cycles, where each cycle doubles the number of target molecules (adapted from Quan et al., 2018).

In contrast, broad-range or generic PCR assays target conserved genomic regions shared across larger taxonomic groups, enabling simultaneous screening of multiple species. These primers amplify DNA from a wide range of organisms within a kingdom or phylum, allowing for subsequent differentiation of species through secondary analyses. A commonly used PCR target for animal species identification is the *Cytochrome C Oxidase I* (*COI*) gene, located on the mitochondrial genome (mtDNA; Figure 3A). Other potential targets are mitochondrially encoded ribosomal RNA genes

(rRNA), such as *12S* and *16S*, or *Cytochrome b* (CYTB) (Fernandes et al., 2021; Heller et al., 2018). Genes located on the chloroplast genome (ctDNA), such as *Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase* (*RuBisCO*) *large subunit* (*rbcL*), *Maturase K* (*matK;* Figure 3B), and Leucine-tRNA (*trnL*) are commonly used for the identification of plant species (Mallott et al., 2018). These extranuclear target genes all share common properties, such as the lack of introns (M. L. Blaxter, 2004; Hebert, Ratnasingham, et al., 2003), and are highly effective for species-level identification due to their interspecific variability and intraspecific conservation (Hajibabaei et al., 2007). Moreover, cells typically contain several hundred to thousand copies of mt- or ct-genomes compared to only two copies of nuclear genomes in diploid organisms, with high variance between tissues (Masuyama et al., 2005; Miller et al., 2003), gender (Guevara et al., 2011) ,age (Masuyama et al., 2005; Miller et al., 2003) and diet (X. Zhang et al., 2020). In animals, elevated mitochondrial mutation rates relative to nuclear DNA (nDNA) accelerate the accumulation of interspecific differences, enabling discrimination even among related species (Allio et al., 2017). Although chloroplast genomes usually evolve more slowly than nDNA, selected plastic loci still provide sufficient interspecific divergence for species identification (Kim et al., 2024; D. R. Smith, 2015). *16S* rRNA genes are widely used for the identification of prokaryotes, whereas the *18S* rRNA gene is mostly used for detecting microbial eukaryotes (Parada et al., 2016; Walters et al., 2016). For fungi, the internal transcribed spacer (ITS) region is commonly used as the primary DNA barcode marker for taxonomic identification and community profiling (Figure 2C; Schoch et al., 2012).

After the initial PCR, the next step is to further analyse the amplification product in order to identify species present in the sample. Restriction fragment length polymorphism (RFLP) digests the PCR amplicon with restriction endonucleases and resolves the resulting fragments by gel electrophoresis. This approach can differentiate species based on variations in restriction sites and is useful for identifying genetically similar species, when diagnostic restriction-site polymorphisms are present (Haider et

al., 2012; Yao et al., 2020). However, it is typically most reliable for single-species amplicons, requires prior knowledge of the DNA sequence to select appropriate restriction enzymes, and can be less effective if the genetic diversity is low (Islam et al., 2021).

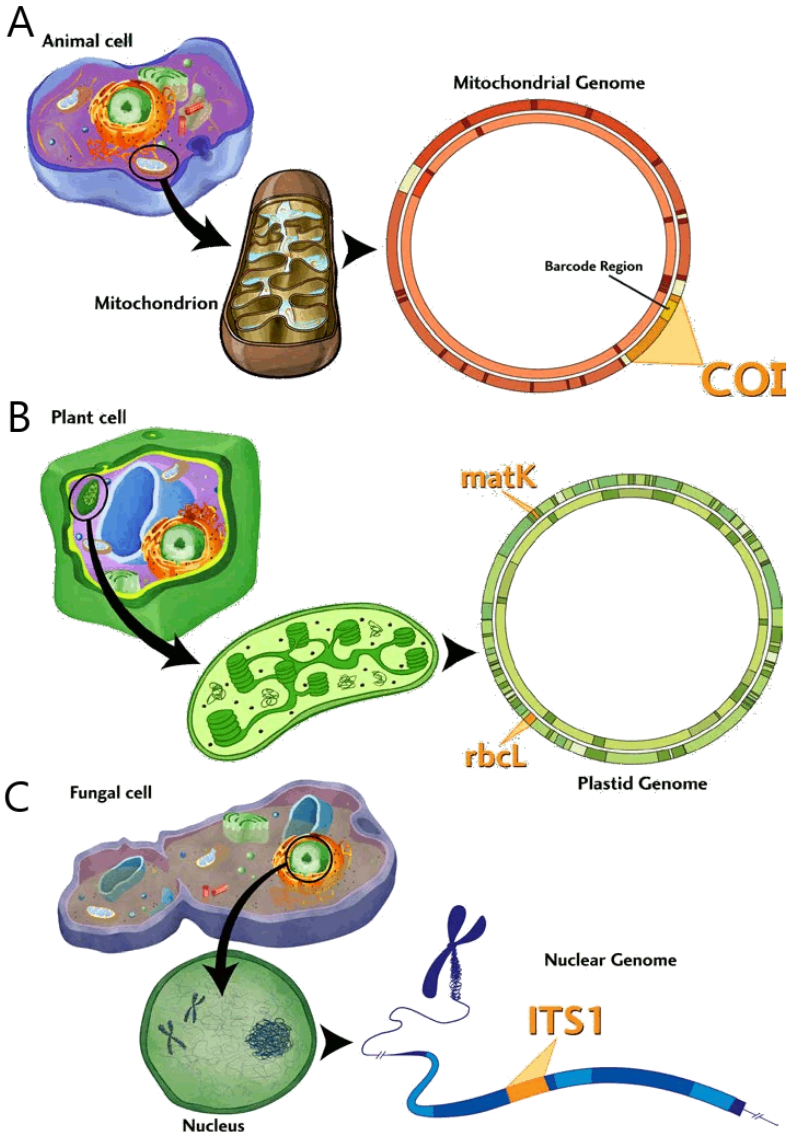

Figure 3: Typical DNA barcoding target regions across major organism groups. **A** The standard DNA barcode marker for animals is a fragment of the *COI* gene located on the mitochondrial genome. **B** The *rbcL* and *matK* genes on the chloroplast genome are commonly used in two-tier DNA barcoding approaches for plants. **C** The nuclear ITS region is the standard marker for most fungi (adapted from Centre for Biodiversity Genomics, 2021).

1.2.2 Quantitative PCR

Quantification is required in food control because regulators and operators must distinguish trace-level, technically unavoidable contamination from economically motivated adulteration, verify declared proportions in multi-ingredient products (e.g. 60 % cod, 40 % haddock), and check against legal thresholds for allergens, protected species, or genetically modified organisms. Qualitative detection alone cannot tell whether a contaminant correspond to dust from a previous production batch or to a relevant replacement of one raw material e.g. by a cheaper one. Decisions on recall, relabelling, or fines therefore need at least semi-quantitative evidence of how much of each species is present.

Quantitative PCR (qPCR), also known as real-time PCR, is able to quantify DNA amounts by detecting fluorescence emitted during each amplification cycle. This fluorescence can originate either from intercalating dyes binding to any double-stranded DNA, such as SYBR Green, or from sequence-specific probes, e.g. TaqMan (Heid et al., 1996; Higuchi et al., 1993). In species identification assays, qPCR uses species-specific probes labelled with a fluorophore (reporter dye) and a quencher. When the target DNA is present, the probe hybridizes specifically to the target strand. During the extension phase, the DNA polymerase cleaves the bound probe, separating the reporter dye from the quencher. This spatial distance allows the reporter dye to emit fluorescence, which increases proportionally with the amount of target DNA in the sample (Kubista et al., 2006). In addition to qualitative species detection, qPCR simultaneously adds quantitative data on the relative abundance of each species within a sample, thereby offering a powerful and reliable tool for routine applications in food authenticity testing and regulatory control (Singh et al., 2024).

Droplet digital PCR (ddPCR) further enhances the precision of DNA quantification by partitioning the PCR reaction mixture into thousands of water-in-oil emulsion droplets. These droplets create separate reaction chambers with the aim of obtaining only a

single DNA molecule per droplet and performing subsequent PCR independently in each droplet (Quan et al., 2018). This approach enables absolute quantification, as each reaction compartment is measured independently, and the number of positive reactions correlates directly with the original amount of template molecules in the sample without the need for calibration standards (Hou et al., 2023; Sidstedt et al., 2020). ddPCR is more tolerant to inhibitors than qPCR and offers high precision at low copy numbers (Quan et al., 2018). The method's analytical power was demonstrated in a recent study detecting seafood adulteration between closely related salmonid species: Atlantic salmon (*Salmo salar*) is often replaced by the cheaper rainbow trout (*Oncorhynchus mykiss*), which is difficult to discriminate in processed food samples. A duplex ddPCR assay targeting the single-copy nuclear *myoglobin* gene with species-specific probes enabled absolute quantification of *Salmo salar* and *Oncorhynchus mykiss* DNA, which was directly converted into corresponding meat mass fractions. Validation on gravimetrically defined mixtures showed high accuracy and reproducibility unaffected by processing treatments such as cooking, freezing, or additive inclusion. When applied to commercial "salmon" products, the method revealed mislabelling with up to 100 % substitution by rainbow trout, underscoring the robustness of ddPCR for quantitative food authentication (X. Y. Ma et al., 2023).

The most commonly used technique for identifying the species origin of a PCR amplicon is DNA sequencing of the amplified DNA, a method, which has become known as DNA barcoding (see below).

1.2.3 DNA barcoding

DNA barcoding was termed after the barcodes present on all products in supermarkets, whose pattern of black and white bars characterizes each product unambiguously. The same idea holds true for DNA barcodes, where the pattern of the four nucleotides as

components of the DNA can be assigned to the sequence of a specific species unambiguously (Hebert, Cywinska, et al., 2003; Hebert, Ratnasingham, et al., 2003).

The initial step of DNA barcoding involves PCR, which is mostly targeted at a gene of broad phylogenetic representation like genes present in mtDNA molecules. Then, DNA sequencing of the amplificate is performed, with the resulting sequence representing the barcode of the species (Figure 4). Sequencing was traditionally performed by the Sanger method, but Next-Generation Sequencing (NGS) methods have become increasingly common (Batovska et al., 2017; Tigrero-Vaca et al., 2025).

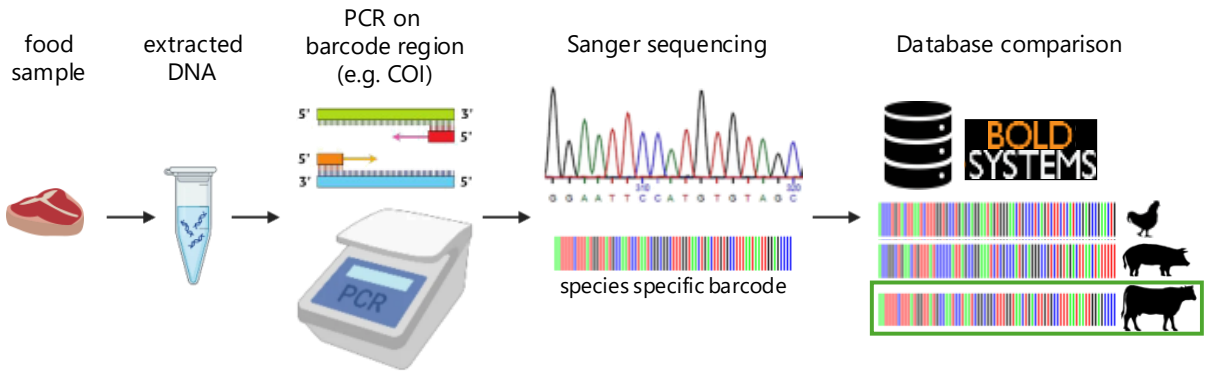

Figure 4: DNA barcoding workflow for species identification. Genomic DNA is extracted from a food sample and a standardized barcode locus (e.g., mitochondrial COI) is amplified by PCR and sequenced using Sanger technology. The resulting chromatogram is quality-checked and converted into a consensus sequence, which is queried against a curated reference database (e.g., BOLD Systems) to assign taxonomy based on sequence similarity. In routine food testing, this approach is most informative for products that are largely single-species or dominated by one biological component, for which a single, unambiguous consensus sequence can be obtained.

Following DNA sequencing, the derived sequence information is compared in a bioinformatic analysis to databases like the Barcode of Life Data System (BOLD) or National Center for Biotechnology Information (NCBI) GenBank (Ratnasingham & Hebert, 2007; Sayers et al., 2022). These databases contain millions of DNA barcodes from different target genes for reference specimen, enabling an alignment of the newly derived DNA barcode sequence with the knowledge within those databases. Matches of defined quality within those databases indicate the presence of a given species in

the sample (Ratnasingham et al., 2024). It is important to note that this barcoding procedure is only successful with pure food material derived from a single species, because Sanger DNA sequencing cannot cope with material of mixed species origin (Dawan & Ahn, 2022).

1.2.4 DNA metabarcoding

Metabarcoding extends the single-specimen barcoding to mixed samples by modifying the steps that accommodate multiple templates in parallel. The amplification of a taxonomically informative locus or multiple loci from a sample is performed using primers that are sufficiently universal to co-amplify many species at once, while at the same time adding sample-specific nucleotide index tags (Ruppert et al., 2019; Taberlet et al., 2012). Short mini-barcode targets increase amplification success in matrices with degraded DNA while preserving usable taxonomic signal, which is why COI amplicons of about 100 to 300 bp are widely used in food authentication and biodiversity studies (Andronache et al., 2025; Dobrovolny et al., 2022; Preckel et al., 2021). The indexed amplicons from multiple samples are then pooled into a single sequencing library and sequenced with NGS, capable of producing tens to hundreds of thousands of reads per amplicon. This way, every specimen's barcode is measured as a separate read rather than as a single trace chromatograms as in traditional barcoding by Sanger sequencing (Figure 5; Taberlet et al., 2012). During the subsequent computational taxonomic assignment, sequence reads are compared against reference databases, allowing identification of numerous species even from highly complex or processed matrices (Giusti et al., 2024; Staats et al., 2016). When applied to routine food surveillance, metabarcoding exhibits clear strengths that address long-standing bottlenecks: First, it enables high-throughput screening of large sample sets in a single sequencing run, since all taxa covered by the primer set are detected simultaneously, whereas classical barcoding require separate assays for each suspected ingredient or matrix (Lanubile et

al., 2024). Second, by combining markers with complementary taxonomic breadth, such as 16S rRNA, COI, and plant ITS2, the method yields multi-kingdom species profiles within one standardized pipeline, avoiding the need for multiple workflows (Giusti et al., 2024; Mottola, Piredda, et al., 2024). Third, the use of short amplicons of about 150 – 250 bp preserves detection capacity in highly processed or heat-treated foods in which DNA is degraded and conventional barcoding underperforms (Gorini et al., 2023).

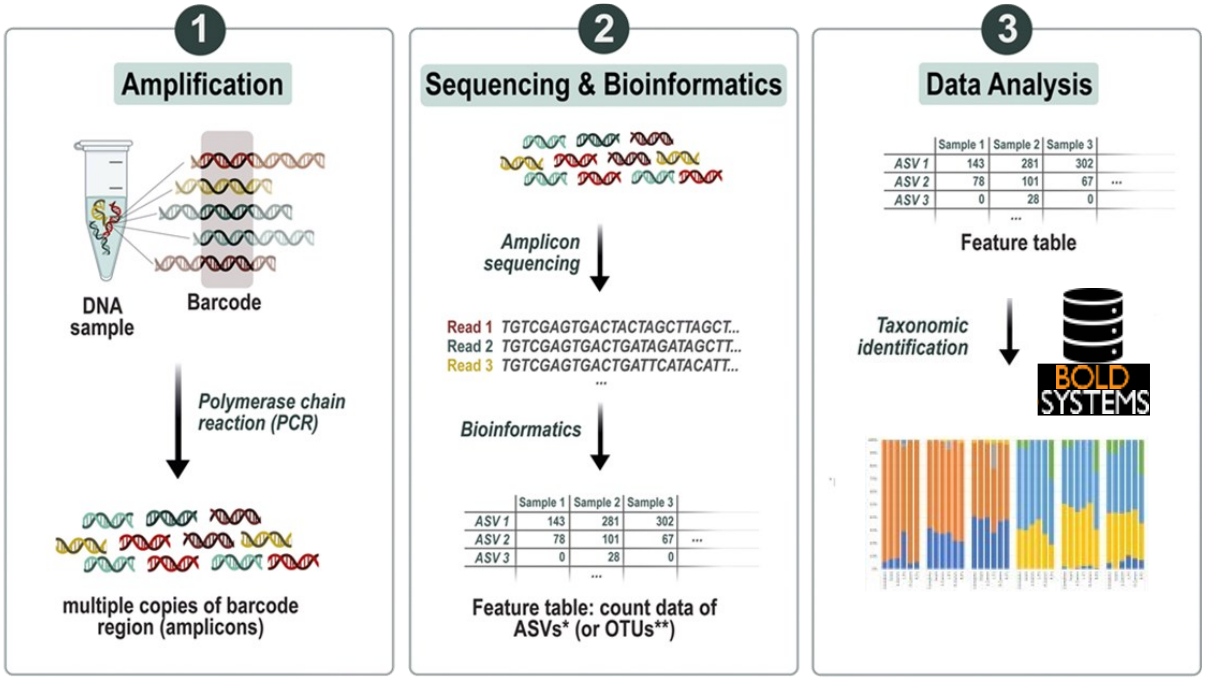

Figure 5: Overview of the metabarcoding analysis workflow. The workflow begins with PCR-based amplification of a selected marker locus (e.g., COI), followed by high-throughput sequencing to generate a set of amplicon reads. Raw reads are subsequently processed by quality filtering and either clustering to infer operational taxonomic units (OTUs) or denoising to infer amplicon sequence variants (ASVs), yielding discrete sequence features representing species entities. Subsequent to clustering or denoising, bioinformatics tools are employed to process the raw data into a feature table, summarizing the abundances of ASVs (or OTUs) across samples. Finally, taxonomic annotation is performed by comparing OTU/ASV sequences against curated reference databases (e.g. BOLD), enabling classification of OTUs or ASVs. *Amplicon Sequence Variants (ASVs): biological sequences in the sample prior to the introduction of amplification and sequencing errors. **Operational Taxonomic Units (OTUs): clusters of reads that differ by less than a fixed sequence dissimilarity thresold, most commonly 3 % (adapted from Esser et al., 2024).

In a recent survey of 62 processed seafood products from retailers in Italy, the United Kingdom, and Albania, DNA metabarcoding with COI and 12S rRNA was used to compare label claims with genetic identifications (Lorusso, Shum, et al., 2024). Amplification succeeded for all but one sample. The authors found 24 of 61 (39 %) samples were mislabelled, typically because undeclared species were present in addition, indicating that a substantial fraction of processed seafood circulating in European retail channels does not correctly represent its true biological composition. Although the authors argue that metabarcoding is ready for routine food control (Lorusso et al., 2024), several gadoids could not be resolved to species level with one marker alone, underscoring the need to combine loci for confident identifications. Moreover, relative read abundances for some taxa differed markedly between COI and 12S, revealing marker-dependent biases that limit quantitative interpretation. These observations highlight a broader regulatory constraint: detection alone is insufficient if decisions hinge on amounts.

1.2.5 Potential problems associated with targeted, PCR-based identification methods

While PCR techniques are powerful tools for species identification, they come with several potential limitations and pitfalls:

I. Barcoding efficiently identifies taxa in environmental or food-derived metagenomic samples, but often **requires multiple assays** to cover different domains of life. Because metabarcoding is marker-targeted and PCR-based, outcomes depend on locus selection and primer binding efficiency, and a single primer set can miss entire clades even with deep sequencing (Alberdi et al., 2018; Elbrecht et al., 2019; Ferreira et al., 2024). In processed food applications, this dependence on intact target loci is further aggravated by DNA fragmentation:

amplification of full-length barcode sequences from moderately or highly processed samples is often unsuccessful and can lead to PCR failure and false-negative results (Shokralla et al., 2015). For complex matrices, complementary loci need to be combined, for example 12S rRNA for animals and the chloroplast trnL gene for plants, to expand taxonomic coverage while accepting marker-specific bias (Mottola, Intermite, et al., 2024).

II. PCR is an extremely **sensitive technique** that can amplify even minute amounts of DNA. This sensitivity makes it vulnerable to contamination by extraneous DNA, which can be introduced during sample collection, preparation, or the PCR process itself. In a case study, traces of sheep (*Ovis aries*) and pork (*Sus scrofa*) were detected in minced pure beef samples, likely due to carryover from inadequate Good Manufacturing Practices and/or insufficient cleaning of shared equipment, like knives, workbenches, meat grinders (Mottola, Intermite, et al., 2024). Such contamination can lead to false-positives by amplifying DNA not intrinsic to the sample, leading to incorrect results and erroneous conclusions.

III. **Species quantification** by metabarcoding is constrained primarily by **primer targeting** and amplification efficiencies across taxa: One of the main shortcomings of metabarcoding is targeting only DNA from species for which the primers bind efficiently, and binding efficiency varies across taxa (Giusti et al., 2024; Macher et al., 2023). In a study identifying mammalian and poultry species in food samples for example, fallow deer (*Dama dama*) was not detected because a commonly used 16S primer for mammalian identification shows two mismatches to the deer sequence (Preckel et al., 2021). In a similar study analysing fish samples with 16S and COI, only four taxa were shared between two markers, while five and seven taxa could only be detected exclusively by one marker (Mottola, Piredda, et al., 2024). Additionally, quantification of components by barcode sequencing has proven problematic due to taxonomic biases induced by the varying primer binding efficiencies across taxa. In a 45-species fish mock community tested with five primer pairs (fish-specific: 2 x 12S

+ 1 x COI; vertebrate-general: 1 x 12S + 1 x 16S), detections and read proportions shifted markedly with primer choice: fish-specific 12S sets recovered the community most faithfully, whereas vertebrate-general sets produced more false-positives and false-negatives. No single primer set was able to correctly detect all 45 species in the sample, and 6 species (13.3 %) could not be amplified at all by any tested primer pair (Macher et al., 2023). Even well-performing primer pairs yielded inconsistent detection among replicates, indicating that primer-induced variability is a key factor in incomplete species recovery. In addition, read proportions for each species varied widely across utilized primers despite identical input DNA, demonstrating that primer-efficiency bias breaks read-count-to-biomass relationships (Macher et al., 2023). As organellar copy number varies strongly across species and tissues, with mitochondrial genomes differing >50-fold in humans and about 200-fold in mice (Rath et al., 2024), the read-count-to-biomass relationship is further decoupled (Shaffer et al., 2025; Shelton et al., 2022). Therefore, there are indications of a shift toward nuclear markers, for example to discriminate hard-to-separate species pairs such as domestic pig and wild boar (Adenuga et al., 2025). In addition, dominant template species can suppress amplification signal of low-abundance ingredients, impairing detection and quantification (Bruno et al., 2019). Summarizing the existing issues, metabarcoding cannot be considered a quantitative analysis tool and is better suited for qualitative species screening (Giusti et al., 2024; Kappel et al., 2023).

IV. DNA barcoding has limitations in resolving **closely related species** and providing information *beyond* species identification, which are critical for comprehensive food authentication (Valentini et al., 2009). Combining multiple DNA barcode marker genes per species must be applied to improve the accuracy and resolution of species identification by providing complementary genetic information that enhances discriminatory power and reduces misidentification (Hollingsworth et al., 2009). A study identifying fish species

using both COI and 16S markers found that a single marker was not enough to discriminate *Gadiformes* beyond genus level (*Gadus sp.* and *Merluccius sp.*), illustrating the limited species-level resolution single barcodes provide (Mottola, Piredda, et al., 2024). While the authors argue multi-marker analysis might enhance resolution, they also state that closely related species like the example of *Gadiformes* will probably remain challenging to resolve even when combining two or more markers (Mottola, Piredda, et al., 2024).

Collectively, the limitations described above are mirrored in a systematic review analysing 23 metabarcoding studies on foodstuffs of animal origin, which concludes that *"application in food authentication was proved as still very limited"* (Giusti et al., 2024). Despite its promise for identifying multiple species in complex samples, the review highlights that the approach suffers from a lack of standardized protocols for target barcode selection, primer design, and consistent quality control procedures. In addition, highly variable bioinformatic pipelines and heterogeneous detection thresholds further reduce comparability between laboratories, contribute to false-positives and false-negatives, and increase operational complexity, restricting routine use in food authentication (Giusti et al., 2024). In line with these findings, several authors emphasize that metabarcoding is not yet suitable as a quantitative method for regulatory decision-making and is currently better suited to qualitative screening for species presence (Giusti et al., 2024; Kappel et al., 2023). Together, these limitations underscore the need for alternative, unbiased approaches that can deliver robust qualitative and quantitative information on species composition in processed, multi-ingredient foods.

1.3 Whole-genome shotgun sequencing for untargeted species detection and quantification

The alternative to a targeted, PCR-based approach to species identification is the strategy to sequence the whole genomic DNA from foodstuff and to classify the generated reads by their bioinformatic assignment to whole-genome sequence databases. This whole-genome shotgun sequencing (WGS) strategy thus circumvents the structural constraints by avoiding primer dependency, thereby capturing a broader spectrum of taxa across all kingdoms of life. WGS thus provides a more holistic view of the DNA representation in the sample by sequencing entire genomes or large portions thereof, capturing not only the marker genes, but also additional genomic regions that contribute valuable information for food authentication (Figure 6; Ripp et al., 2014). Beyond identification, WGS enables quantitative inference by counting reads for each reference genome, allowing robust estimation of relative abundances across taxa (Bell et al., 2021; Haiminen et al., 2019; Rieder et al., 2023; Ripp et al., 2014). This quantitative capability is directly useful for assessing contamination levels, verifying labelling claims (especially when close to legal thresholds), and supporting regulatory compliance decisions in routine testing.

Technically, WGS involves the random fragmentation of DNA followed by high-throughput sequencing of all DNA fragments, e.g. using the Illumina Next-Generation Sequencing protocol. This method enables the simultaneous detection of bacterial, fungal, plants, and animal DNA in a single analysis (Bell et al., 2021; Haiminen et al., 2019; Ripp et al., 2014). This ability of WGS to provide a near-complete identification of species in a sample could improve the analytical process for food authentication authorities by streamlining the experimental steps required to obtain the same or even more information compared to currently applied PCR-based methods.

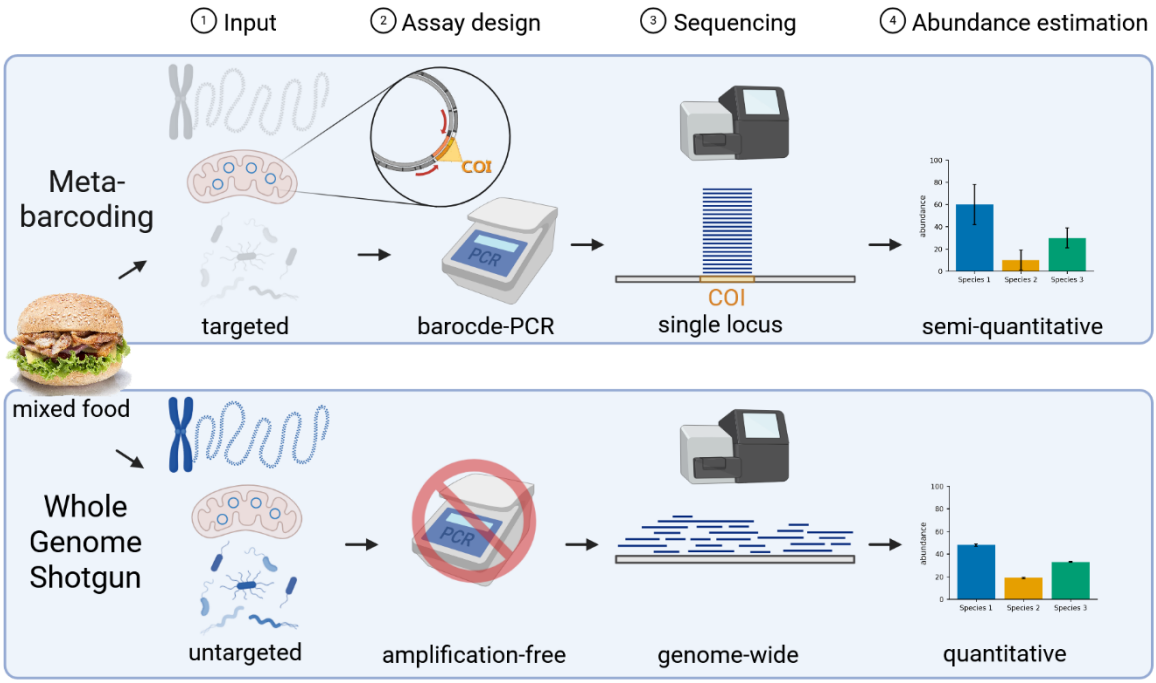

Figure 6: Conceptual comparison of metabarcoding and whole-genome shotgun sequencing (WGS) for food species identification and quantification. Metabarcoding relies on marker PCR and therefore captures sequencing information from a single targeted barcode locus, which can miss taxa that are not efficiently amplified. In contrast, WGS sequences all DNA fragments in the sample, enabling detection of both expected and unexpected species across all domains of life. Consequently, WGS is amenable to quantitative inference from read counts, whereas metabarcoding is generally only semi-quantitative due to amplification-related biases.

The principle was first demonstrated by All-Food-Sequencing (AFS): The method uses WGS reads generated by NGS and databases with reference genomes to identify and quantify species in food samples (Ripp et al., 2014). In a four-species calibration sausage, the quantitative mode (AFS-quant) returned matches at >2 % resolution (e.g., 54.8 % vs 55 % sheep), with absolute deviations across taxa of 0.24 - 1.79 %; the high-specificity mode (AFS-spec) reduced false-positive assignment to closely related water buffalo from 0.64 % to 0.07 %. In addition, the metagenomic branch of AFS classified previously unmapped reads, identified microbial signals and spike-ins of 11 plants species, among them soy, lupine, and hazelnut, and flagged undeclared taxa (Ripp et al., 2014). This highlights both the methods ability to analyse all domains of life in a single unbiased assay, as well as to identify even minor traces of allergens. A similar

approach following the same principle is Food Authentication from SEquencing Reads (Haiminen et al., 2019). FASER utilizes BLAST with WGS reads against reference genomes to identify and quantify. Analysing the same data set as Ripp et al., 2014, they were able to achieve results comparable to as AFS. Deployed on 31 factory high-protein powders, FASER confirmed the labelled chicken in all samples and revealed unexpected pork and beef signal in three samples. A comparison of WGS and metabarcoding on pollen mock communities showed that WGS achieved nearly complete species recovery and produced quantitative estimates that correlated more strongly with the actual pollen grain proportions than those obtained from metabarcoding (Bell et al., 2021), underscoring the potential of WGS for robust identification and quantification in mixed biological materials. Together these results show that one WGS assay can span all domains of life, detect undeclared ingredients, and deliver quantitative results without amplification bias in routine food authenticity testing.

1.4 All-Food-Sequencing: state at the beginning of this thesis

AFS is a WGS-based technique used for the simultaneous qualitative and quantitative analysis of species in complex food samples containing multiple components, first described by Ripp et al., 2014. The method includes DNA extraction from a homogenized food sample, followed by length-fragmentation and creation of an Illumina sequencing library. Subsequent sequencing is carried out on instruments such as Illumina Next-Seq or NovaSeq.

During the bioinformatic analysis, reads are mapped to reference genome sequences of respective species and then quantified by read counting (referred to as the "Quantitative Mapping Approach", Figure 7). AFS uses the Burrows-Wheeler Aligner (BWA) *aln* algorithm when aligning these reads to reference genomes (H. Li & Durbin,

2009). Employing an iterative mapping method, AFS starts with the analysis of fully matching sequences and classifies them into three categories: *Unique reads* that map specifically to a single reference genome, *Multimapped reads* that map with equal quality to multiple reference genomes, and *Unmapped reads* that cannot be assigned to any reference genome. *Unique reads* are directly attributed to the appropriate species. *Multimapped reads* originate from homologous genomic regions shared among multiple species, and their source cannot be definitively determined. Thus, these reads are allocated proportionately according to the distribution seen in the unique reads. *Unmapped reads* then undergo two additional rounds of mapping with a decreased requirement of sequence identity between read and reference genome each time (Ripp et al., 2014).

At the end of three successive mapping runs, the final distribution of identified species within the sample is calculated based on all matches obtained per species. Any remaining unmapped reads can be identified qualitatively in a subsequent, separate step through computationally intensive database searches using the BLAST alignment algorithm (termed "Qualitative Metagenomic Analysis Algorithm"; Altschul et al., 1990; Camacho et al., 2009). In this procedure, all unmapped reads are compared with the NCBI non-redundant nucleotide collection (nr/nt) by local BLAST. This allows for the identification of species that were not initially included in the reference genomes selection. If a reference genome is available for a newly identified species, the initial quantitative mapping step can be repeated under inclusion of this new component. In cases where no suitable reference genome is available, the unmapped reads can still help indicate the presence of certain species in food samples qualitatively. This approach makes it possible to detect even small amounts of ingredients like spices, impurities, or allergens, and also assists in determining microbiological diversity within samples (Liu et al., 2017; Ripp et al., 2014).

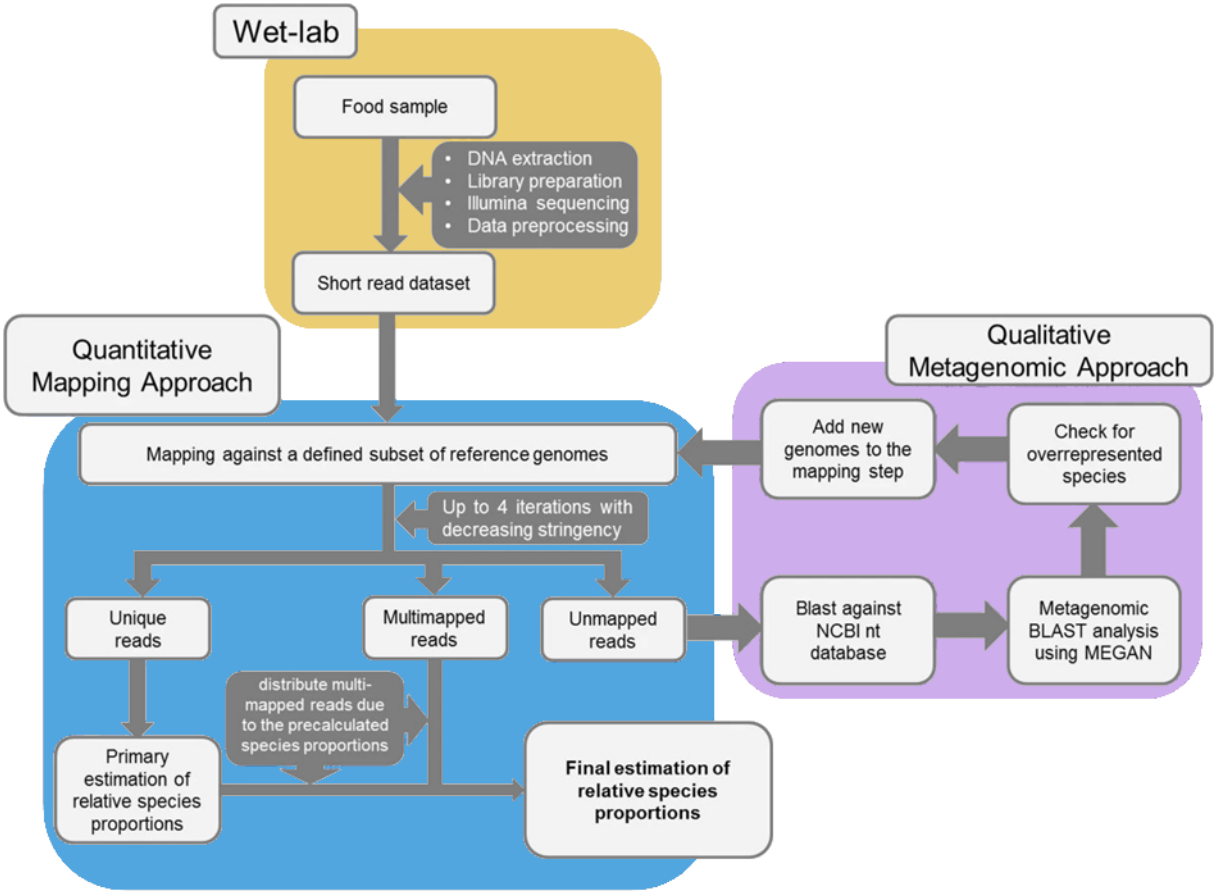

Figure 7: Overview of All-Food-Sequencing workflow (Ripp et al., 2014). In the "Wet Lab" step, DNA is extracted from food samples, libraries are prepared, and sequencing is performed. In the second step, referred to as the "Quantitative Mapping Approach," the short-read dataset is iteratively mapped against a predefined set of reference genomes. Reads that map uniquely to a reference are used to estimate the relative proportions of known species. Reads that map to multiple references are proportionally distributed according to initial abundance estimates, ensuring that every read contributes to the quantification. Any reads that cannot be aligned at progressively relaxed mapping thresholds are designated as unmapped. In the "Qualitative Metagenomic Approach" step, these unmapped reads undergo a BLAST search against a comprehensive database such as NCBI's nr/nt. The output of this search is examined with metagenomic analysis tools to identify additional or overrepresented species that were not included in the original reference set. Newly discovered genomes can be added to the reference list to refine the quantitative mapping step.

Despite its potential, AFS has certain limitations that need to be addressed to ensure its robustness and reliability in food analysis:

I. Because regulatory decisions by food control authorities are tied to validated methods and defined thresholds, AFS must be **benchmarked against established approaches**. Side-by-side tests with matched samples, certified

reference materials, and orthogonal PCR assays quantify trueness and precision, establish calibration functions, and determine when genus-level reporting is more reliable than species-level claims.

II. Previous validations of AFS were limited to six species in controlled mixtures, which is insufficient to reveal limitations that arise in larger, closely related species clades and complex matrices typical of routine food monitoring. A meaningful assessment requires **extending the taxonomic spectrum** to major product categories (mammals, poultry, fish, molluscs, and plants), including several closely related species per group that differ in genome size, ploidy state, and reference quality.

III. Equally important, the method must be tested on **real retail samples** without a priori target selection. A survey across diverse product categories should apply the untargeted AFS workflow, incorporate newly discovered taxa into the quantitative analysis, and report concordance with labels, detection of undeclared ingredients, and allergen findings.

IV. A further limitation is the method's ability to **distinguish closely related** species, which complicates the identification of taxa with high genetic similarity. Multimapped reads are distributed by heuristic rules based on uniquely mapped reads, which can inflate species present at low proportions or distribute reads across several close relatives even when only one species is present. This constrains the robustness of species-level assignments within tight clades and may necessitate more conservative genus-level reporting in such groups.

V. The iterative, alignment-based workflow is **computationally demanding**. Each BWA *aln* pass scales with read count, with the number and size of reference genomes, and with the relaxed identity thresholds applied in later rounds. Because the reference database is held in memory throughout, RAM usage rises sharply for large eukaryotic genomes. Broad reference panels therefore inflate both runtime and memory demand, often forcing preselection of targets and making wide, untargeted screens impractical on typical laboratory hardware.

1.5 Aims of this thesis

The objective of this thesis is to advance the methodology of DNA-based food analysis via AFS by directly addressing the limitations identified for the current workflow and proposing solutions based on experimental data. This objective is pursued through a series of targeted aims that encompass the comparison of AFS to established methods, improvement of the bioinformatic algorithm used in AFS, exploration of the method's limitations, and the integration of new, emerging sequencing technologies.

A key aim of this thesis is to perform a comprehensive comparison of AFS to well-established species identification and quantification methods such as qPCR and ddPCR in a validation framework that is compatible with regulatory decision-making. While qPCR and ddPCR have been widely adopted for species identification and quantification in food products due to their sensitivity and specificity, they also have notable limitations, particularly in terms of their dependence on prior knowledge of target DNA sequences. This thesis therefore conducts side-by-side tests of AFS, qPCR, and ddPCR on matched samples and certified reference materials, with a focus on regulatory decision levels. The study derives calibration functions, quantifies trueness and precision, and assesses under which conditions genus-level reporting is more reliable than species-level claims. This analysis evaluates the performance of AFS relative to these methods, focusing on parameters such as sensitivity, specificity, accuracy, and practical applicability in routine food analysis. By conducting side-by-side comparisons, the thesis aims to highlight the strengths and weaknesses of AFS, potentially positioning it as a superior alternative for comprehensive food authenticity testing.

A second central aim is to overcome the narrow validation scope of the original AFS study, which is restricted to six species in controlled mixtures. This thesis investigates these limitations by expanding the species database and incorporating more reference genomes. To evaluate performance across varied types of food products, a curated

species set is defined that covers major groups such as mammals, poultry, fish, molluscs, and plants, and incorporates several closely related species per group differing in genome size, ploidy state, and genome assembly quality. Structured calibration mixtures with varying compositions and mass fractions are used to probe cross-assignment within tight clades and to define reliable detection and quantification limits. Additionally, the research explores strategies to mitigate these limitations, such as optimizing read length and depth of sequencing to enhance resolution and discrimination power.

Third, this thesis extends AFS beyond controlled mixtures to real-world retail products, thereby testing the method under conditions that reflect routine food monitoring. A survey across diverse product categories applies the untargeted AFS workflow without *a priori* target selection. Newly detected taxa from the qualitative metagenomic step are iteratively incorporated into the quantitative mapping, and the results are evaluated in terms of concordance with labels, detection of undeclared ingredients, allergens, and other mislabelled components. This provides a systematic assessment of how well AFS captures species compositions in complex, processed matrices and where its current boundaries lie.

The initial AFS methodology employs the BWA algorithm, specifically the BWA aln algorithm for read mapping. This algorithm, however, is outdated, and newer bioinformatic approaches promise improved performance. Other algorithms like BWA mem or bowtie2 show decreased runtimes and improved accuracy. This thesis explores the bioinformatic pipeline used in AFS and improves its performance. A promising design trend in bioinformatics is the application of so-called k-mer based algorithms for read classification, which fragment sequences into shorter, overlapping sub-sequences of length k. These approaches offer higher speeds at the same or even higher sensitivity compared to traditional read mapping approaches. The integration of a k-mer-based method is expected to enhance the efficiency and accuracy of species identification and quantification in complex food samples. The thesis benchmarks these

new algorithms against the current BWA aln to quantify improvements in computational speed, mapping accuracy, and overall performance.

Advancements in sequencing technologies present new opportunities for improving AFS: Long-read sequencing provided by Third-Generation Sequencing (TGS) from both Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) generate reads that span entire genomic regions of up to megabase levels. Offering a more comprehensive view of the genetic material, these methods reduce the false-positive mapping rates, especially in hard-to-map regions such as highly repetitive sequences. Thereby, they potentially improve the accuracy of species identification, especially in complex samples and for closely related species. This thesis evaluates the feasibility and benefits of integrating long-read sequencing into the AFS workflow. By comparing the performance of long reads to the traditional short reads used in AFS, the research aims to determine whether this integration can provide more accurate and reliable results for food analysis.

Overall, the planned work addresses the key gaps that currently prevent AFS from being used as a fully established method in food analysis. Rigorous comparison with qPCR and ddPCR aligns AFS performance metrics with regulatory expectations, while the extension of the species panel and calibration designs moves validation beyond the original calibration sausage mixtures. The application of AFS to real-world retail products tests its behaviour under realistic matrix and labelling conditions, revealing both detection potential and practical limitations. Finally, the refinement of the computational workflow, with a focus on modern alignment and k-mer based approaches, improves scalability and resolution, thereby strengthening the case for AFS as a routine tool in food authenticity control.

2 Results

2.1 Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq)

Hellmann SL, Ripp F, Bikar SE, Schmidt B, Köppel R, Hankeln T

Published in: European Food Research and Technology

DOI: https://doi.org/10.1007/s00217-019-03404-y

Own contributions to this publication:

- Data analysis: Bioinformatic AFS analysis of calibration samples and Doner Kebab samples (with F. Ripp)
- Matrix calibration: Calculation of calibration to correct matrix-driven bias (with F. Ripp)
- *In silico* benchmarking: Quantification of false-positive detection rates among closely related species
- Performance evaluation: Comparison of AFS to qPCR and ddPCR results (with F. Ripp)

Experimental design, data analysis, data interpretation, and drafting of the manuscript were conducted in collaboration with F. Ripp and Prof. Dr. T. Hankeln. The project was managed by Prof. Dr. T. Hankeln.

European Food Research and Technology (2020) 246:193–200
https://doi.org/10.1007/s00217-019-03404-y

ORIGINAL PAPER

Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq)

Sören Lukas Hellmann[1] · Fabian Ripp[1,2] · Sven-Ernö Bikar[3] · Bertil Schmidt[4] · Rene Köppel[5] · Thomas Hankeln[1]

Received: 2 September 2019 / Accepted: 2 November 2019 / Published online: 16 November 2019

Abstract

Complex food matrices bear the risk of intentional or accidental admixture of non-declared species. Moreover, declared components can be present in false proportions, since expensive taxa might be exchanged for cheaper ones. We have previously reported that PCR-free metagenomic sequencing of total DNA extracted from sausage samples combined with bioinformatic analysis (termed All-Food-Seq, AFS) can be a valuable screening tool to identify the taxon composition of food ingredients. Here, we illustrate this principle by analysing regional Doner kebap samples, which revealed unexpected and unlabelled poultry and plant components in three of five cases. In addition, we systematically apply AFS to a broad set of reference meat material of known composition (i.e. reference sausages) to evaluate quantification accuracy and potential limitations. We include a detailed analysis of the effect of different food matrices and the possibility of false-positive sequence read assignment to closely related species, and we compare AFS quantification results to quantitative real-time PCR (qPCR) and droplet digital PCR (ddPCR). AFS emerges as a potent PCR-free screening tool, which can detect multiple target species of different kingdoms of life within a single assay. Mathematical calibration accounting for pronounced matrix effects can significantly improve AFS quantification accuracy. In comparison, AFS performs better than classical qPCR, and is on par with ddPCR.

Keywords Food metagenomics · Species identification · Doner Kebap · Read mapping · Next-generation-sequencing

Introduction

The determination and quantification of food ingredients is an important issue in official food control. The complexity of foodstuff, difficulties in the traceability of trading channels and the globalisation of food markets open doors for fraud and failures in correct labelling, stocking and processing procedures [1]. Possible consequences for consumers are manifold beginning with compliance of ethical aspects such as halal, kosher or vegan over health risks caused by pathogenic organisms to simple deception because of economic reasons. In fact, biological contaminants made up the vast majority of warning notices released by the German authorities [2] between 2011 and 2015. The majority of these cases were provoked by microbiological contaminations or the presence of non-declared allergenic food components. Therefore, food and drug legislation demands proper declaration of ingredients and compliance to storage and transport conditions [3, 4]. To ensure adherence to law and to maintain consumer's safety, there is a growing need for methods that allow for precise determination of food ingredients, ideally spanning all kingdoms of life including plants, animals, bacteria, fungi and

Sören Lukas Hellmann and Fabian Ripp equal contribution.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00217-019-03404-y) contains supplementary material, which is available to authorized users.

✉ Thomas Hankeln
hankeln@uni-mainz.de

[1] Institute of Organismic and Molecular Evolution, Molecular Genetics and Genome Analysis, Johannes Gutenberg University Mainz, J. J. Becher-Weg 30A, 55099 Mainz, Germany

[2] Present Address: MVZ Labor Volkmann, Karlsruhe, Germany

[3] StarSEQ GmbH, Mainz, Germany

[4] Institute of Informatics, Johannes Gutenberg University Mainz, Mainz, Germany

[5] Official Food Control Authority of the Canton Zurich, Zurich, Switzerland

perhaps also extending to viruses. A broad palette of analytical methods for analysing foodstuffs has been developed and is routinely applied at official food control laboratories, but also private and industrial control labs. Among these, DNA-based methods like PCR are probably the most widely used technologies, because of their high sensitivity and the possibility to perform quantitative measurements [5–13]. However, even when multiplexed or performed in the meta-barcoding format, PCR-based approaches have the drawback to detect only a limited range of target species and produce assay-dependent amplification biases [14–17].

We have previously shown that deep metagenomic DNA sequencing of whole-genome DNA from foodstuffs, followed by dedicated bioinformatic analysis, is in principle able to overcome these issues. DNA sequence reads obtained from food can be bioinformatically assigned to existing reference genomes for species identification, and the number of reads successfully assigned to a respective genome can be counted to give a quantitative measure of the species proportions. Importantly, such whole-genome sequencing of foodstuff DNA (termed All-Food-Seq: AFS; [18, 19]) does not require any a priori definition of possible target species. AFS can, therefore, be viewed as a screening method, which theoretically can detect an infinite spectrum of diverse species, being only limited by our current knowledge of genomes, as represented in the fast-growing public sequence databases. The "identification plus quantification" principle based on read assignment and read counting has been successfully demonstrated so far as a proof-of-principle in a limited number of foodstuff samples, i.e. sausages of pre-defined composition prepared as reference material [8, 20]. We, therefore, decided to further investigate the potential of AFS in a real-life test case, analysing different doner kebap samples obtained from snack bars. We also saw the necessity to evaluate the quantification potential of AFS in more detail. Inferring species proportions from DNA read proportions can be difficult, because it may substantially depend on the food composition and processing. As an example, high quality meat may be substituted for in a product by the addition of rind, lard or skin, which could affect DNA amounts per gram tissue, and thus the inference of species proportions within the foodstuff. To study this so-called matrix effect, we have applied the AFS method to an extended set of reference sausage samples, each containing known admixtures of different meat sources, but prepared according to different recipes [8, 20]. We compared the AFS quantification results to those obtained by qPCR and ddPCR on the same samples and evaluated the effects of matrix composition.

Materials and methods

Food samples and DNA extraction

Doner kebap samples were purchased at five snack bars distributed in the Rhine-Main area. All meat pieces were identified by eye and selected by sterile forceps for subsequent homogenization in large volume using a standard kitchen device. About 1 g of the homogenised matrix, which looked surprisingly different (ranging between an oily and granular texture), was taken for subsequent DNA isolation using the Wizard Plus Miniprep DNA purification system (Promega, Madison, USA) according to the manufacturer's protocol. DNA was quantified by Qubit fluorometry (Thermofisher Scientific, Schwerte, Germany).

Calibration sausage samples containing admixtures of cattle, chicken, pig, sheep and turkey at defined amounts were produced by a professional butchery and provided by the Official Food Control Authority of the Canton Zürich, Switzerland [8, 20]. The samples were prepared for calibration of foodstuff detection methods and reflect three different recipes of sausage production (Online Tab. S1): AllMeat sausage (Kal A-E: meat), Lyoner-style sausage (KLyo A-D: matrix of meat, rind and lard) and Poultry-Lyoner (KGeflLyo A-D: matrix of meat and skin). Total DNA was extracted out of 200 mg homogenised sausage sample using the Wizard Plus system (Promega, Madison, USA) according to the manufacturer's protocol.

Illumina library preparation and sequencing

Sequencing library preparations and sequencing were performed by a commercial provider (StarSEQ, Mainz, Germany). The Nextera DNA Library Preparation Kit (Illumina, San Diego, USA) was applied following the manufacturer's instructions. Typically, 1 ng of total DNA was used. Sequencing was carried out on an Illumina MiSeq instrument using reagent kit v.2 in 150 bp paired-end (reference sausages) and 50 bp single-end (doner kebap samples) mode, respectively. In principle, both sequencing modes deliver comparably valid results [19]. Between 200 k and 2600 k, reads were generated per sample (Online Tab. S1). Adjustments by downsampling were omitted, because our previous analysis showed that read numbers > 100 k produced consistent quantification results independent of dataset size [19]. All datasets were quality checked, trimmed and filtered using FASTQC data evaluation software [21] and trimmomatic v0.33 trimming tool [22]. Datasets have been submitted to the SRA database under the project names PRJNA271645 and PRJEB34001.

Bioinformatic analysis of main ingredients using AFS

The AFS read-mapping pipeline was executed with three rounds of iterative mapping and step-wise decreased mapping stringency, as described [18, 19]. This strategy allows for a final number of two mismatches after mapping step 3. At each round, reads that mapped against one of the provided reference genomes were cumulatively counted and reported on a 1–100% scale to reflect relative species proportions. In the doner kebap screening analysis, sequence reads were mapped against a selection of reference genomes (accession numbers: cattle: NC_037328.1, sheep: NC_040252.1, goat: NC_030808.1, pork: NC_010443.5, horse: NC_009144.3, chicken: NC_006088.5, turkey: NC_015011.2, maize: NC_024459.2 and soy: NC_016088.3). In the quantification analysis of the calibrator sausages, reference genome choice was limited to the animal species cattle, chicken, pig, horse, sheep, goat, water buffalo (accession number: NC_037545.1) and turkey. Goat and water buffalo genomes were added to test the robustness of AFS towards false-positive signals to be expected between closely related species. All evaluations were performed on a standard desktop PC (Intel(R) Core(TM) i7-8700 CPU @ 3.20 GHz, 16 GB DDR4 2667 MHz RAM, 256 GB SATA SSD, CentOS Linux release 7.6.1810).

Reads that did not match, very likely originate from species not provided as a reference during the AFS mapping step. These unmapped reads (around 3% per sample), often representing spice plants and microbiota [18], did not undergo further metagenomic analysis in the present study, since the prime goal was to evaluate the quantification properties of AFS for the main meat components.

Calculation of false-positive read assignments

To determine false-positive read assignment rates for the tested species, in particular the closely related cattle-buffalo, chicken-turkey and goat-sheep, we created in silico datasets of different proportions of reads for each species with the corresponding related species being absent. To this end, we used whole-genome shotgun datasets from the SRA (SRR8588004, SRR9663406, SRR8442931, SRR8560982, SRR6470934) and performed data pre-processing as described for the reference sausages. For each species, proportions of 1, 5, 10, 25, 50, 75 and 100% were extracted using the reformat tool from the BBMap suite [23] and complemented to 1 mio reads with the non-related plant species rice (accession number: NC_008394.4). For the cattle-buffalo species pair, we only inspected the false-positive rate of buffalo assignments given a cattle ingredient, as the opposite direction is irrelevant to food safety inspections in our opinion. To investigate the effect of sequence read length on false-positive mapping, all generated datasets were trimmed using the reformat tool to a length of 50, 100 and 150 bp, respectively. Subsequent AFS analyses were performed as described above with three mapping rounds (accepting max. two mismatches) against buffalo, cattle, chicken, goat, horse, pork, sheep and turkey genomes.

Results

AFS screening of species composition in doner kebap samples

Doner kebap samples were obtained from five snack bars in the Rhine-Main area. Their meat components were sampled, homogenised and the extracted DNA sequenced. AFS analysis revealed that samples 2 and 3 were prepared from pure beef, while samples 1, 4 and 5 consisted of beef and turkey, with the latter as the dominant component (Table 1a). Samples 1, 3 and 5 revealed measurable amounts of soybean DNA (0.5–0.8%), and sample 1 additionally contained maize DNA (1.8%). In samples 2, 3 and 5, we observed that 0.1–0.4% of sequence reads were assigned to goat and sheep. Since the latter also belong to the family of Bovidae, one could interpret the goat/sheep read assignments as candidate false-positives, produced as a consequence of the phylogenetic relatedness and the presence of conserved genomic elements. However, our detailed evaluation of possible false-positive values (see below; Online Fig. S1b, Online Tab. S2) shows that at least for samples 2 and 3, the measured values of goat and sheep are slightly higher than expected for a matrix consisting almost only of cattle. We, therefore, cannot rule out that small amounts of sheep and goat material were indeed present in these doner samples, allegedly caused by the presence of cheese matrices or due to unknown circumstances during doner production. In contrast, the 0.2 and 0.3% of chicken reads in samples 1 and 4, which are clearly dominated by turkey, may accordingly be considered false-positives.

AFS quantification of meat ingredients in reference sausages

To specifically study the quantification properties of AFS in a broad set of samples, a total of 13 reference sausage samples (Online Tab. S1), prepared according to three different standard recipes, were sequenced and analysed. Datasets were then studied to evaluate quantification accuracy, the impact of different matrices (i.e. meat, rind, lard and skin), and the probability of false-positive read assignments. AFS results were then compared to quantification data previously obtained by qPCR [8, 20] and droplet digital (dd) PCR [13] on the very same sausage samples (Table 1b–d).

Table 1 Quantification results of AFS pipeline

(a)	Beef	Turkey	Soy	Maize	Horse	Pork	Chicken	Sheep	Goat
Doner Kebab 1	8.6	88.6	0.7	1.8	0.0	0.0	0.2	0.0	0.0
Doner Kebab 2	99.5	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3
Doner Kebab 3	98.7	0.0	0.4	0.0	0.0	0.0	0.0	0.4	0.4
Doner Kebab 4	5.1	94.5	0.0	0.0	0.0	0.0	0.3	0.0	0.0
Doner Kebab 5	27.7	71.3	0.4	0.0	0.0	0.0	0.3	0.1	0.1
(b)	Beef	Pork	Sheep	Horse	Chicken	Turkey	Goat	Buffalo	Sum dev
Kal A									
Expected	1.0	35.0	9.0	55.0	0.0	0.0	0.0	0.0	
AFS	*1.4*	30.9	10.1	57.3	0.0	0.0	0.3	0.0	8.2
AFS cal.	0.3	*34.9*	10.3	*54.3*	0.0	0.0	0.3	0.0	*3.1*
qPCR	0.4	39.3	*8.9*	51.5	n.a.	n.a.	n.a.	n.a.	8.6
Kal B									
Expected	9.0	55.0	1.0	35.0	0.0	0.0	0.0	0.0	
AFS	11.2	49.5	1.3	37.8	0.0	0.0	0.1	0.1	11.0
AFS cal.	*9.6*	*54.2*	*1.0*	*35.1*	0.0	0.0	0.1	0.1	*1.7*
qPCR	23.0	51.0	1.5	24.4	n.a.	n.a.	n.a.	n.a.	29.1
Kal C									
Expected	25.0	25.0	25.0	25.0	0.0	0.0	0.0	0.0	
AFS	25.8	19.8	28.2	*25.1*	0.0	0.0	1.0	0.2	10.5
AFS cal.	23.7	22.4	29.0	23.7	0.0	0.0	0.9	0.2	10.3
qPCR	34.0	18.8	25.8	21.4	n.a.	n.a.	n.a.	n.a.	19.6
ddPCR	*25.6*	*25.3*	*24.6*	24.5	n.a.	n.a.	n.a.	n.a.	*1.8*
Kal D									
Expected	35.0	9.0	55.0	1.0	0.0	0.0	0.0	0.0	
AFS	38.2	7.7	50.8	1.3	0.0	0.0	1.7	0.3	11.1
AFS cal.	*35.3*	*9.1*	*52.1*	1.5	0.0	0.0	1.7	0.3	*5.8*
qPCR	29.9	7.6	61.7	*0.9*	n.a.	n.a.	n.a.	n.a.	13.3
Kal E									
Expected	55.0	1.0	35.0	9.0	0.0	0.0	0.0	0.0	
AFS	56.9	*1.0*	32.2	*8.5*	0.0	0.0	1.1	0.4	6.7
AFS cal.	*54.5*	1.9	*33.8*	8.4	0.0	0.0	1.1	0.4	*4.7*
qPCR	51.2	*1.0*	37.2	10.5	n.a.	n.a.	n.a.	n.a.	7.5
(c)	Beef	Pork	Sheep	Horse	Chicken	Turkey	Goat	Buffalo	Sum dev
KLyo A									
Expected	14.0	80.0	0.0	0.0	0.5	5.5	0.0	0.0	
AFS	19.3	74.6	0.0	0.0	0.7	*5.3*	0.0	0.1	11.2
AFS cal.	13.0	80.8	0.0	0.0	0.3	*5.7*	0.0	0.1	2.3
ddPCR	*14.2*	*80.2*	n.a.	n.a.	*0.4*	*5.3*	n.a.	n.a.	*0.7*
KLyo B									
Expected	36.0	58.0	0.0	0.0	2.0	4.0	0.0	0.0	
AFS	41.8	52.4	0.0	0.0	2.2	3.4	0.0	0.3	12.5
AFS cal.	*35.9*	*57.9*	0.0	0.0	*2.1*	*3.8*	0.0	0.3	*0.8*
ddPCR	34.9	60.1	n.a.	n.a.	1.4	3.7	n.a.	n.a.	4.1
KLyo C									
Expected	58.0	36.0	0.0	0.0	4.0	2.0	0.0	0.0	
AFS	66.0	28.3	0.0	0.0	*3.9*	1.4	0.0	0.4	17.0
AFS cal.	60.6	33.1	0.0	0.0	*4.1*	1.7	0.0	0.4	6.3
ddPCR	*58.4*	*36.8*	n.a.	n.a.	2.9	*1.9*	n.a.	n.a.	*2.4*

Table 1 (continued)

(c)	Beef	Pork	Sheep	Horse	Chicken	Turkey	Goat	Buffalo	Sum dev
KLyo D									
Expected	80.0	14.0	0.0	0.0	5.5	0.5	0.0	0.0	
AFS	82.8	11.3	0.0	0.0	5.0	*0.4*	0.0	0.5	6.8
AFS cal.	77.7	*15.7*	0.0	0.0	*5.3*	0.7	0.0	0.5	4.9
ddPCR	*79.2*	15.9	n.a.	n.a.	4.3	*0.6*	n.a.	n.a.	*4.0*
(d)	Beef	Pork	Sheep	Horse	Chicken	Turkey	Goat	Buffalo	Sum dev
KGeflLyo A									
Expected	0.5	5.5	0.0	0.0	14.0	80.0	0.0	0.0	
AFS	*0.6*	5.1	0.0	0.0	29.8	64.5	0.0	0.0	31.9
AFS cal.	0.2	*5.6*	0.0	0.0	*10.0*	*84.1*	0.0	0.0	*8.5*
ddPCR	0.8	9.4	n.a.	n.a.	24.6	65.1	n.a.	n.a.	29.7
KGeflLyo B									
Expected	2.0	4.0	0.0	0.0	36.0	58.0	0.0	0.0	
AFS	*2.0*	3.5	0.0	0.0	56.9	37.6	0.0	0.0	41.9
AFS cal.	2.2	*3.9*	0.0	0.0	*40.3*	*53.5*	0.0	0.0	*9.1*
ddPCR	2.1	5.0	n.a.	n.a.	41.7	51.2	n.a.	n.a.	13.6
KGeflLyo C									
Expected	4.0	2.0	0.0	0.0	58.0	36.0	0.0	0.0	
AFS	3.4	1.4	0.0	0.0	75.7	19.5	0.0	0.0	35.4
AFS cal.	4.4	*1.8*	0.0	0.0	*61.2*	*32.6*	0.0	0.0	*7.2*
ddPCR	*4.2*	2.4	n.a.	n.a.	62.6	30.8	n.a.	n.a.	10.4
KGeflLyo D									
Expected	5.5	0.5	0.0	0.0	80.0	14.0	0.0	0.0	
AFS	3.9	*0.4*	0.0	0.0	89.2	6.5	0.0	0.0	18.4
AFS cal.	*5.2*	0.7	0.0	0.0	*76.4*	*17.8*	0.0	0.0	*7.9*

Raw results obtained by AFS analysis as well as calibrated AFS values obtained by linear regression are compared to PCR-based quantification for (a) Doner Kebab samples, (b) Kal A-E, (c) Klyo A-D and (d) KGeflLyo A-D. qPCR data were obtained from [8, 20], ddPCR data from [13]. "Sum dev" represents the sum of % deviation from expected proportions. Best results for each sausage are italics. (*n.a.* not analysed)

AFS quantification accuracy

Our sample set covered expected species proportions from 0.5 to 80%. Minimal and maximal expected components varied between the three different sample types (meat-only samples Kal A-E 1–55%, mixed-matrix samples KLyo A-D and KGeflLyo A-D 0.5–80%). It turned out that even the low concentrations of ingredients could be detected by AFS with high accuracy ($0.5 \pm 0.1\%$; $1 \pm 0.1\%$; $2 \pm 0.4\%$; $4 \pm 0.2\%$; $5.5 \pm 0.5\%$). As species concentrations increased, absolute deviations of measured values also increased to a maximum of 20.9% for the 9–36% interval and 20.4% for the 55–80% interval, respectively (Table 1b–d). To compare the performance of AFS for the different sausage types, we summed up the individual species deviations for each sausage individually. Results showed that, omitting any calibration calculations (see below), Kal A-E samples were quantified with overall best results (ranging between 6.7 and 11.1% deviation), followed by KLyo A-D (6.8–17.0%) and KGeflLyo A-D (18.4–41.9%).

Evaluation of false-positive read assignments between related species

The species assignment of sequence reads in AFS is based on classical read-mapping algorithms involving sequence alignment [18]. This implies the potential danger of a misclassification if a read contains highly conserved DNA sequences, often present in the genomes of phylogenetically closely related taxa. Of course, such false-positive assignments could have, if present, an eminent effect on detection accuracy. To evaluate the potential of such false-positive read assignment within AFS, we intentionally included in the read-mapping step the reference genomes of species, which are not present in the sausages, but which are evolutionarily close to the real food components. Specifically, we added the genome of the water buffalo (*Bubalus bubalis*), which shared a common ancestor with cattle 13 mio years ago, and the goat (*Capra hircus*), which diverged from sheep 10 mio years ago ([24]; Online Fig. S1a). False-positive signals of buffalo and goat from sausages containing cattle or

sheep as real ingredients ranged between 0.0 and 1.7% and depended on the amount of the corresponding real ingredient species (Table 1b–d). For example, the maximal value of 1.7% false goat reads was obtained for the Kal D sausage containing 55% sheep.

To systematically specify the chance of false-positive read assignments between species pairs in AFS, we simulated read datasets with varying, known amounts of reads from the species in our study and mapped them to the respective reference genomes. The amount of false-positive reads in fact scaled linearly with the real ingredient proportions (Online Fig. S1b, Online Tab S2), allowing us to define threshold values for the respective species pairs. Interestingly, but not unexpectedly, the short 50 bp reads produced markedly higher false-positive values than 100 bp and 150 bp reads. For example, a 100% sheep dataset produced 5.1% false-positive goat assignments with 50 bp read length, but only 2.7% with 150 bp reads (Online Fig. S1b). Some minor 'asymmetric' quantification results (i.e. chicken against turkey genome versus turkey against chicken genome) could be noted and are probably caused by different qualities of the respective reference genomes. Notwithstanding, these calculated values can now be applied by the AFS user to objectively assign quantification values as potential false-positives, as done above in the case of the doner kebap samples.

Matrix effects and their possible correction by linear regression

Different types of food matrices can bias quantification analyses, because different tissues often contain varying concentrations of DNA, and cellular DNA may also be extracted from them at different efficiencies. To study this effect, we included three types of sausage matrices: the Kal samples, consisting only of pure meat, the KLyo samples, in which pork material was represented by three tissues (meat, rind and lard at a ratio of 1:4:15) and the KGeflLyo sausages, containing chicken material as a 1:1 mixture of meat and skin (Online Tab. S1).

Specifically for the KGeflLyo sausages with their partial replacement of chicken meat by skin (Online Fig. S2), the chicken component showed a substantial overrepresentation on the DNA level, thus severely compromising the quantification results for this matrix type (independent of whether AFS or PCR methods were applied; comp. Table 1). While samples containing meat-only chicken showed minimal deviations of 0.1–0.5% from expected values, the meat/skin matrix led to an almost proportional overestimation of chicken by 9–20% (Table 1d; Online Fig. S2). A second, but milder effect was noticed for pork as an ingredient, which was systematically underestimated by 2.7–7.8% in the KLyo A-D, 0.1–0.6% in the KGeflLyo A-D and 0–5.5% in the Kal A-E samples, respectively.

Assuming that the observed effects represent systematic errors, we decided to normalise our measurements by applying linear regression. We did this for every sample type and species separately to consider both matrix-specific and species-dependent effects (see calibrated AFS values in Table 1). In fact, the improvement of the quantification values turned out to be massive, showing that AFS (very much like the PCR methods; comp. [8, 13]) will benefit from the establishment of such matrix calibration factors. Indeed, we were able to correct efficiently for most of the systematic error over a broad range of expected values. Note, however, that in some cases (e.g. Kal A and E), deviation slightly increased after the normalisation procedure for the very low expected values of 0.5 and 1%, respectively (Table 1b).

Limits of detection and precision

Using normalised values gained by linear regression, we calculated the limit of detection (LoD) of AFS at a confidence level of 95%, applying the procedure described by [25]. The LoD describes the lowest quantity of an analyte that can be reliably detected above the observed background noise. In the case of read-mapping approaches, LoD will depend on genome relatedness and resulting chance for false-positive read assignment, which in turn partly depends on read length (see above). If closely related genomes (e.g. sheep vs. goat and cattle vs. buffalo) are included in an AFS mapping procedure using 150 bp reads, the method produces a LoD of 1.6%. If only distant species are tested for, the LoD decreases to 1.0%.

To also infer the random error produced by AFS, and thus the precision of the method, we calculated 95% confidence intervals for every instance of the expected species proportions between 0.5 and 80% (Online Tab. S3). Proportion components below 2% are measured with about 50% uncertainty. Measurement error decreased to about 10% for proportions between 2 and 36%, and 4% for proportions above 36%. Overall, CIs turned out to be close to the expected values and, therefore, are an excellent indication of high AFS precision over the entire range of expected values from 0.5 to 80%.

Discussion

Classical DNA-based species identification in food is routinely performed as a targeted approach using PCR-based methods, which can detect only a certain range of taxa, for which the PCR primers ideally fit [5–8, 10–13]. AFS in contrast analyses the complete DNA of a foodstuff without amplification and is, therefore, a non-targeted, whole-genome screening approach [18]. To investigate the potential of AFS to detect unforeseen species components, we chose

to study a real-case food control scenario and sequenced the meat from five doner kebap samples from the Rhine-Main area. According to German food legislation, snacks sold under the label "doner kebap" are expected to consist only of sheep and/or beef [26]. However, occasional surveys conducted by food authorities [27] or even occasioned by broadcasting stations [28] have already pointed at a considerable heterogeneity of animal species components in doner kebap samples from Germany, which very often contained unlabelled poultry (chicken, turkey) and in rare cases even pork. Using AFS, we found that three of our five samples indeed contained turkey meat, two samples even at a major extent (90% or more). None of the samples, however, was openly advertised to the consumer as "poultry doner". In addition, AFS detected in four cases soy as an unexpected and unlabelled ingredient, which may be critical for consumers suffering from allergy towards soybeans. Soybean DNA may originate from the usage of spice coating (panada). One sample additionally contained maize DNA, the origin of which is unclear. AFS thus confirmed the results previously obtained by other labs in doner kebap species screens and thus should function well as a method in routine food screening.

The performance of AFS for the quantification of species in different types of food matrices has not yet been investigated systematically. The main focus of the current study was, therefore, to explore the quantification potential of AFS to infer species proportions of reference sausages, which have previously been used in the field to evaluate PCR-based quantification methods. To directly compare AFS to quantification results obtained by qPCR [8] and ddPCR [13], we calculated for simplicity the sum of the % deviation (measured vs. expected) for each sausage sample (Table 1b–d). Results showed that AFS data—very much like the qPCR and ddPCR data—need to be calibrated for matrix-dependent biases to generate the most accurate results. Indeed, in 8 of 12 cases "AFS-cal" produced the best results, while ddPCR turned out to be clearly superior in 1 case (Kal C) and slightly better in 3 cases (KLyo A, C, D). AFS readily identifies and quantifies proportions of species over a broad % range. Most importantly, it works at the 1% level, a value often approximatively taken by food authorities to distinguish problematic species amounts from trace amounts, e.g. originating by unavoidable contamination.

Very much like for other DNA-based methods, the limitations of AFS are set by sequence similarities between closely related genomes and by the so-called matrix effect, which ultimately determines the extent to which species proportions in food can be indirectly inferred from DNA proportions. Our theoretical evaluation of possible wrong read assignments between closely related taxa provides the applicant of AFS with a means to readily distinguish between true and false quantification results. Food consisting of species, which have diverged at minimum 10 mio years ago (e.g. sheep–goat or cattle–buffalo), may thus be analysed without much problems. If AFS is performed for other, possibly closer taxa, the limits of false-positives can easily be determined by the procedure, which we have outlined in the methods section.

As previously noticed for PCR-based quantification methods [8, 13], the AFS requires mathematical calibration for matrix effects to achieve best results (see above). Theoretically, for instance, it should be necessary for AFS to take into account that birds have only 1/3 the genome size of mammals. In practice, this consideration proved to be not useful at all for quantifying food containing a mixture of bird and mammalian material by AFS (data not shown). The possible reason is that chicken meat may contain more DNA per gram tissue than, e.g., pork [29], thus compensating for the smaller genome size. It will be almost impossible to define the DNA amounts for all conceivable tissues from food-relevant species. However, the application of food matrix reference material, as done in the present study, facilitates a guided calibration of matrix effects and thus efficiently circumvents this problem.

In conclusion, we confirm here that AFS is a potent additional screening and quantification tool in the repertoire of foodstuff analysis. We have calculated that AFS sequencing reagent costs (50 libraries prepared in parallel, 500 k reads each, all loaded on 1 Illumina MiSeq flowcell) currently would amount to appr. 90 EUR per sample (see [19] for high-multiplex estimations). The computer skills required match those of a typical bioinformatics master student, and routine screening of 1 mio reads against up to 10 eukaryotic genomes can be performed on a laptop PC requiring a computation time of appr. 20 min (see Materials and Methods for hardware used). We like to point out that, in contrast to standard PCR analytics and depending on the desired depth of analysis, the AFS can go well beyond the mere identification of animal and plant species into the world of food microbiota, even including viruses [18]. Ideally, AFS would screen for the ever-growing number of sequenced animal, plant, fungal and bacterial genomes in one single analysis on standard computers. However, due to the usage of algorithms involving read mapping and sequence alignments, the screening power of AFS is currently limited to 20–30 species with large eukaryotic genomes in 1 analysis. We, therefore, investigate the applicability of novel non-alignment-based, memory-efficient algorithms for AFS. At the same time, the identification and quantification of microbiota from foodstuff by AFS is a goal worth of pursuing in future.

Acknowledgements TH and SLH gratefully acknowledge funding by the Federal Office for Agriculture and Food (project ID: 2816503814), Johannes Gutenberg University Center for Computational Sciences (CSM) and the Ministry of Justice and for Consumer Safety Rhineland-Palatinate.

Compliance with ethics standards

Conflict of interest The authors declare that they have no conflict of interest.

Research involving human participants and/or animals This article does not contain any studies with human and animal subjects.

References

1. German Federal Office for Risk Assessment: Food safety and globalisation—challenges and opportunities (Bundesamt für Risikobewertung) (2019) https://www.bfr.bund.de/en/press_information/2014/13/food_safety_and_globalisation___challenges_and_opportunities-190341.html. Accessed 28 Aug 2019
2. German Federal Office of Consumer Protection and Food Safety (Bundesamt für Verbraucherschutz und Lebensmittelsicherheit) (2019) https://www.bvl.bund.de/DE/Home/homepage_node.html. Accessed 28 Aug 2019
3. German Drug Law (Bundesministerium der Justiz und für Verbraucherschutz: Gesetz über den Verkehr mit Arzneimitteln) (2019) https://www.gesetze-im-internet.de/amg_1976/AMG.pdf. Accessed 28 Aug 2019
4. Swiss Food Legislation (Schweizerisches Bundesgesetz über Lebensmittel und Gebrauchsgegenstände (Lebensmittelgesetz, LMG) vom 20 (2014) https://www.admin.ch/opc/de/official-compilation/2017/249.pdf. Accessed 28 Aug 2019
5. Brodmann PD, Moor D (2003) Sensitive and semi-quantitative TaqMan™ real-time polymerase chain reaction systems for the detection of beef (Bos taurus) and the detection of the family Mammalia in food and feed. Meat Sci 65:599–607. https://doi.org/10.1016/S0309-1740(02)00253-X
6. Zhang C-L, Fowler MR, Scott NW et al (2007) A TaqMan real-time PCR system for the identification and quantification of bovine DNA in meats, milks and cheeses. Food Control 18:1149–1158. https://doi.org/10.1016/J.FOODCONT.2006.07.018
7. Köppel R, Ruf J, Zimmerli F, Breitenmoser A (2008) Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, chicken and turkey. Eur Food Res Technol 227:1199–1203. https://doi.org/10.1007/s00217-008-0837-7
8. Köppel R, Ruf J, Rentsch J (2011) Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, horse and sheep. Eur Food Res Technol 232:151–155. https://doi.org/10.1007/s00217-010-1371-y
9. Köppel R, Eugster A, Ruf J, Rentsch J (2012) Quantification of meat proportions by measuring DNA contents in raw and boiled sausages using matrix-adapted calibrators and multiplex real-time PCR. J AOAC Int 95:494–499. https://doi.org/10.5740/jaoacint.11-115
10. Ulca P, Balta H, Çağın İ, Senyuva HZ (2013) Meat species identification and Halal authentication using PCR analysis of raw and cooked traditional Turkish foods. Meat Sci 94:280–284. https://doi.org/10.1016/j.meatsci.2013.03.008
11. Floren C, Wiedemann I, Brenig B et al (2015) Species identification and quantification in meat and meat products using droplet digital PCR (ddPCR). Food Chem 173:1054–1058. https://doi.org/10.1016/j.foodchem.2014.10.138
12. Song K-Y, Hwang HJ, Kim JH (2017) Ultra-fast DNA-based multiplex convection PCR method for meat species identification with possible on-site applications. Food Chem 229:341–346. https://doi.org/10.1016/j.foodchem.2017.02.085
13. Köppel R, Ganeshan A, Weber S et al (2019) Duplex digital PCR for the determination of meat proportions of sausages containing meat from chicken, turkey, horse, cow, pig and sheep. Eur Food Res Technol 245:853–862. https://doi.org/10.1007/s00217-018-3220-3
14. Markoulatos P, Siafakas N, Moncany M (2002) Multiplex polymerase chain reaction: a practical approach. J Clin Lab Anal 16:47–51. https://doi.org/10.1002/jcla.2058
15. Berry D, Ben Mahfoudh K, Wagner M, Loy A (2011) Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl Environ Microbiol 77:7846–7849. https://doi.org/10.1128/AEM.05220-11
16. Tedersoo L, Anslan S, Bahram M et al (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. MycoKeys 10:1–43. https://doi.org/10.3897/mycokeys.10.4852
17. Sze MA, Schloss PD (2019) The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data. mSphere 4:e00163–19. https://doi.org/10.1128/msphere.00163-19
18. Ripp F, Krombholz C, Liu Y et al (2014) All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. BMC Genomics 15:639. https://doi.org/10.1186/1471-2164-15-639
19. Liu Y, Ripp F, Koeppel R et al (2017) AFS: identification and quantification of species composition by metagenomic sequencing. Bioinformatics 33:btw822. https://doi.org/10.1093/bioinformatics/btw822
20. Eugster A, Ruf J, Rentsch J, Köppel R (2009) Quantification of beef, pork, chicken and turkey proportions in sausages: use of matrix-adapted standards and comparison of single versus multiplex PCR in an interlaboratory trial. Eur Food Res Technol 230:55–61. https://doi.org/10.1007/s00217-009-1138-5
21. FASTQC (2019) A quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 28 Aug 2019
22. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
23. BBTools (2019) A suite of fast, multithreaded bioinformatics tools designed for analysis of DNA and RNA sequence data. https://sourceforge.net/projects/bbmap/. Accessed 28 Aug 2019
24. Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. Mol Biol Evol 34:1812–1819. https://doi.org/10.1093/molbev/msx116
25. Armbruster DA, Pry T (2008) Limit of blank, limit of detection and limit of quantitation. Clin Biochem Rev 29(Suppl 1):S49–S52
26. Marking of "doner kebab" and "similar" products by bulk delivery (2019) (Kenntlichmachung von „Döner Kebab"und „ähnlichen"Erzeugnissen bei loser Abgabe. Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit). https://www.lgl.bayern.de/downloads/lebensmittel/doc/merkblatt_doener_kebab.pdf. Accessed 28 Aug 2019
27. The composition and labelling of doner kebabs (2019) LACORS. https://www.ihsti.com/lacors/ContentDetails.aspx?id=21001. Accessed 28 Aug 2019
28. Frequently minced meat and additives in doner kebab (2019) (Häufig Fleischbrät und Zusatzstoffe im Döner. Norddeutscher Rundfunk). https://www.ndr.de/ratgeber/verbraucher/Haeufig-Fleischbraet-und-Zusatzstoffe-im-Doener,doener164.html. Accessed 28 Aug 2019
29. Cai Y, Li X, Lv R et al (2014) Quantitative analysis of pork and chicken products by droplet digital PCR. Biomed Res Int 2014:810209. https://doi.org/10.1155/2014/810209

2.2 All-Food-Seq: Next-Generation Sequencing-basiertes Screeningverfahren zur quantifizierbaren Spezies-identifikation in prozessierten Lebensmitteln

Hellmann SL, Kobus R, Schmidt B, Bikar SE, Köppel R, Hankeln T

Published in: 8. Fachtagung Gentechnik - Neue molekularbiologische Techniken und deren Herausforderungen für die Analytik - Band 12 Gentechnik für Umwelt- und Verbraucherschutz

ISSN: 1866-7775
ISBN: 978-3-96151-077-1

Own contributions to this publication:

- Conceptualization of scope
- Writing: Original draft and review
- Visualization: design and refinement of all schematics

Scope and drafting of the manuscript were conducted in collaboration with Prof. Dr. T. Hankeln. The project was managed by Prof. Dr. T. Hankeln.

6 All-Food-Seq: Next Generation Sequencing-basiertes Screeningverfahren zur quantifizierbaren Speziesidentifikation in prozessierten Lebensmitteln

Sören Lukas Hellmann[1], Robin Kobus[2], Bertil Schmidt[2], Sven Bikar[3], René Köppel[4] und Thomas Hankeln[1]

[1] *Institut für molekulare und organismische Evolutionsbiologie, AG Molekulare Genetik & Genomanalyse, Johannes Gutenberg-Universität, Mainz*

[2] *Institut für Informatik, AG Hochleistungsrechnen, Johannes Gutenberg-Universität Mainz*

[3] *StarSEQ GmbH, Mainz*

[4] *Kantonales Labor Zürich, Gesundheitsdirektion Kanton Zürich, Schweiz*

6.1 DNA-basierte Speziesidentifikation: ein unverzichtbares Werkzeug der Lebensmittelüberwachung

Regelmäßig berichten die Nachrichten von Lebensmittelskandalen: falsch deklarierte Rezepturen oder Austausch teurer durch günstigere Zutaten werden mit Regelmäßigkeit aufgedeckt [1–5]. Neben einer Täuschung des Verbrauchers kann dies erhebliche gesundheitliche Auswirkungen haben, wenn der Verbraucher Nahrungsmittelunverträglichkeiten oder Allergien hat oder es zur Aufnahme gesundheitsschädlicher Substanzen kommt. Auch ethische Aspekte der Ernährung (z. B. halal, koscher, vegan) gilt es zu beachten. In Deutschland erkranken jährlich über 200.000 Menschen an durch Lebensmittel übertragenen Mikroorganismen [6], die z.B. durch einen Mangel an Hygienemaßnahmen bei der Verarbeitung in die Produkte gelangen können [7]. Lebensmittelhändler selbst sowie Behörden der Lebensmittelüberwachung müssen die Möglichkeit haben, die Identität und Qualität der gelieferten Waren zu überprüfen. Im Sinne des Verbraucherschutzes und zur Verhinderung von unlauterem Wettbewerb ist es daher erforderlich, standardisierte Methoden für die eindeutige Identifizierung biologischer Arten zur Verfügung zu haben. Für prozessierte Lebensmittel haben sich dafür DNA-basierte Nachweisverfahren als vorteilhaft erwiesen, da die Struktur von Proteinen je nach Grad der Verarbeitung oft zerstört wird und ein Nachweis auf Proteinebene sich somit schwierig gestalten kann. Des Weiteren zeichnen sich DNA-basierte Methoden wie die am häufigsten eingesetzte Polymerase-Kettenreaktion (PCR) durch ihre hohe Sensitivität, Spezifität und Quantifizierbarkeit (quantitative PCR, qPCR) aus [8–17]. Diese PCR-Verfahren beruhen zum Nachweis einer Spezies in der Regel auf der Amplifikation von allgemein hoch-konservierten Genabschnitten, die charakteristische artspezifische Basensubstitutionen aufweisen. Die Zielsequenzen für die PCR stammen oftmals aus mitochondrialer (*cytB*, *cox1*, *16S rDNA*) oder plastidärer DNA (*rbcL*, *matK*), da diese Organellen-DNAs in hoher Kopienzahl vorhanden und selbst nach starker Prozessierung im Gewebe effizient nachweisbar sind.

Typischerweise werden die PCR-Amplikons für den Artennachweis mit der klassischen Sanger-Technik sequenziert. Die dann sichtbaren charakteristischen Sequenzaustausche werden als artspezifischer „DNA-Barcode“ bezeichnet [18–20]. Für viele hunderttausende tierische, pflanzliche und Pilz-Spezies sind solche DNA-Barcodes bereits in entsprechenden Sequenzdatenbaken gesammelt worden, so dass die aus einem Lebensmittel erhaltenen Barcode-Sequenzen einfach durch Datenbankabgleich identifiziert werden kann [21]. Eine Abwandlung der beschriebenen Methodik kommt zum Tragen, wenn nicht homogenes biologisches Material aus einer Spezies, sondern komplexe Lebensmittelgemische bestehend aus mehreren Arten analysiert werden sollen. Beim sogenannten „Meta-Barcoding“ erfasst ein Primerpaar beispielweise das *cytB*-Gen von Tieren. Die resultierenden Amplifikate stellen dann ein Gemisch der unterschiedlichen Tierarten in der Probe dar. Sie werden hoch-parallel durch Next-Generation-Sequencing entschlüsselt. Die im Gemisch vorhandenen unterschiedlichen artspezifischen Sequenzen können danach durch einen Datenbankabgleich identifiziert werden (**Abbildung 1**).

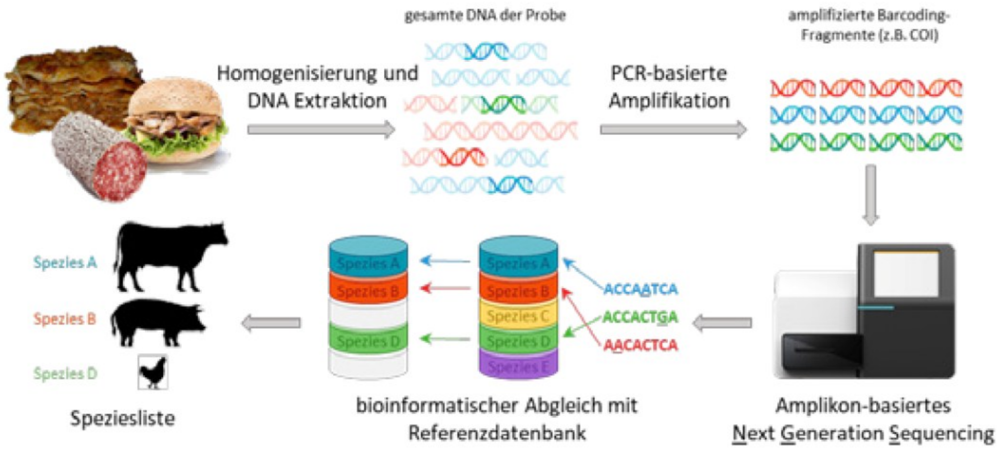

Abbildung 1: Schematischer Aufbau einer „Meta-Barcoding“-Analyse: Die zu analysierende Probe wird homogenisiert und eine Gesamt-DNA-Extraktion durchgeführt. Mittels PCR werden Genabschnitte mit speziesspezifischen Sequenzaustauschen (z.B. aus dem Gen der mitochondrialen COI) amplifiziert, sog. DNA-Barcoding Fragmente. Für eine umfassende Analyse multipler Spezies sind parallele PCR-Ansätze mit unterschiedlichen Primersystemen notwendig. Die DNA-Barcoding Fragmente werden anschließend hoch-parallel durch NGS-Methoden sequenziert. In einer bioinformatischen Analyse werden die sequenzierten Fragmente mit bestehenden Referenzdatenbanken abgeglichen und somit das Artenspektrum der metagenomischen Probe abgeleitet.

Qualitativ kann man die Artzusammensetzung des Lebensmittels auf diese Weise hochspezifisch ermitteln. Für die gleichzeitige Erfassung z.B. von Pflanzen, Tieren und Mikroorganismen sind aber parallele Ansätze mit jeweils anderen PCR-Primersystemen erforderlich. Prinzipiell ist also das Next-Generation Sequencing (NGS)-basierte Meta-Barcoding durch den Einsatz dieser Primersysteme auf ein Spektrum an identifizierbaren Arten beschränkt.

Die Barcode-typischen Markergene aus Organellen-DNA schwanken zudem in der Kopienzahl je nach Gewebe und auch zwischen den Arten sehr stark. Eine Quantifizierung von Art-Anteilen durch einfaches Auszählen der Meta-Barcode-Sequenzen gilt daher durch Schwankungen in der Anzahl der zur Verfügung stehenden PCR-Matrizen, durch die variable Bindungsspezifität der Primer und durch Assay-abhängige Amplifikationsbias als eher problematisch [22–29].

Eine Alternative zum PCR/Barcode-basierten Speziesnachweis stellt die Sequenzierung der Gesamtheit aller genomischer DNA eines Lebensmittels dar („whole-genome shotgun metagenomics“). Die artspezifischen Unterschiede der erhaltenen Sequenzabschnitte sollten dabei eine Bestimmung der Artzusammensetzung erlauben. In der Tat bestehen die Genome von Eukaryoten überwiegend aus nicht-funktionellen Bereichen, die weitgehend ohne selektiven Druck während der Evolution speziesspezifische Mutationen anhäufen und so für eine Artbestimmung bestens geeignet sind. Der Genomanteil zwischenartlich konservierter Gen-Exons ist dagegen niedrig (z.B. 1.2 % bei Säugetieren [30]). Lebensmittelrelevante Spezies wie Schwein, Huhn u.a. zeigen zudem innerhalb einer Art nur wenige Polymorphismen von etwa 0,5 – 5 pro 1.000 Nukleotiden [31–34]. Diese geringen Werte sollten eine Differenzierung der Spezies durch gesamt-genomische Sequenzierung nicht negativ beeinflussen; sie könnten hingegen für eine Bestimmung von Populationen und geografischer Herkunft des Materials ausgewertet werden.

Eine weitere Überlegung ist, dass der DNA-Anteil jeder Spezies in einem komplexen Lebensmittel in etwa proportional zum Gewichtsanteil der entsprechenden Spezies in der Probe sein sollte. Dies würde zusätzlich zur Artidentifikation eine Quantifizierung von Speziesanteilen in Nahrungsmittelgemischen ermöglich, indem man die Anteile der Arten in dem Sequenzdatensatz durch Auszählen der sogenannten „Sequenz-Reads“ bestimmt. Zudem kann ein solcher Ansatz der metagenomischen Gesamtsequenzierung von Lebensmittel-DNA alle in der Probe vorhandenen Spezies, egal ob eukaryotisch, prokaryotisch oder gar viral, in einem einzigen Experiment bestimmen. Daher entfällt die Notwendigkeit von multiplen Assays, um Bestandteile aller Domänen des Lebens identifizieren zu können. Eine zusätzliche DNA-Amplifikation über Primer entfällt ebenso, wie das hieraus resultierende Bias. Da die erhaltenen genomischen DNA-Sequenzen quasi ohne vorherige Erwartungen betrachtet werden, können auch unerwartete „exotische“ Speziesanteile detektiert werden, insbesondere auch solche, die ein mögliches Gesundheitsrisiko für den Konsumenten darstellen.

6.2 All-Food-Sequencing: Gesamt-genomisches NGS-basiertes Lebensmittel-Screening

Das All-Food-Sequencing (AFS) ist ein DNA-basiertes Screeningverfahren zur gleichzeitigen qualitativen und quantitativen Artendiagnose in komplexen Lebensmitteln, die aus Anteilen mehrerer Spezies bestehen [35]. Die Methode umfasst die ungezielte Sequenzierung der gesamten Genom-DNA eines Lebensmittels mittels NGS. Daher wird im Gegensatz zu anderen Methoden wie PCR oder Meta-Barcoding keine Amplifikation bestimmter genomischer Regionen und somit auch keine a priori-Informationen in Form von z.B. Amplifikationsprimern benötigt. Eine zu prüfende Lebensmittelprobe wird homogenisiert und die gesamte DNA extrahiert. Die gewonnene DNA wird längenfragmentiert und eine Illumina-Sequenzierbibliothek erstellt; die Sequenzierung erfolgt auf z.B. auf einem Illumina HiSeq-, NextSeq- oder MiSeq-Gerät. Bereits ca. 200k erhaltene Sequenz-Reads einer Leselänge von je 50 bp sind ausreichend, um die Hauptkomponenten des Lebensmittels mit einem Anteil von mindestens 1 % zuverlässig zu quantifizieren [36]. Besser werden für eine Analyse mit der AFS jedoch etwa 1 Mio. Reads erstellt, um auch Bestandteile in geringeren Anteilen zuverlässig detektieren zu können. Die bioinformatische Arbeit, bei der die Reads identifiziert und gezählt werden, läuft folgendermaßen ab (**Abbildung 2**): zunächst werden die Reads hochspezifisch an eine Palette zur Verfügung stehender Referenz-Genomsequenzen kartiert und somit einer Art zugeordnet sowie dabei auch ausgezählt („quantitatives Mapping“). Reads, die bei diesem Schritt der Auswahl an Referenzgenomen nicht zugeordnet werden können („Unmapped Reads“), werden optional durch massive Datenbank-Suchen mit dem Alignment-Algorithmus BLAST identifiziert („qualitative Metagenomik“).

Beim Kartieren der Reads an Referenzgenome greift AFS auf den etablierten Mapping-Algorithmus BWA zurück [37]: Im quantitativen Mapping werden die Reads einer Lebensmittelprobe iterativ in drei Durchgängen mit abnehmender Stringenz an eine Auswahl von Referenzgenomen kartiert. Zunächst werden nur vollständig sequenzidentische Reads ausgewertet. Resultierende Treffer werden in drei Kategorien eingeteilt: 1) Unique Reads, die spezifisch an ein Referenzgenom kartieren, 2) Multimapped Reads, die mit gleicher Güte an mehrere Referenzgenome kartieren und 3) Unmapped Reads, die keinem Referenzgenom zugeordnet werden konnten (**Abbildung 2**). Die Unique Reads werden direkt der entsprechenden Spezies zugeordnet. Multimapped Reads entstehen durch zwischen mehreren Spezies konservierte Genomregionen und ihre Herkunft ist nicht eindeutig bestimmbar. Daher werden diese Reads anteilig im Verhältnis entsprechend der bei den Unique Reads beobachteten Distribution auf die beteiligten Spezies verteilt. Zuletzt werden die Unmapped Reads als Eingabe für zwei weitere Mapping-Durchgänge genutzt, wobei jeweils die geforderte Sequenzidentität zwischen Read und Referenzgenom um 1 % gesenkt wird.

Am Ende der drei Durchgänge wird anhand aller pro Spezies erzielten Treffer die finale Distribution der identifizierten Spezies berechnet. Die verbleibenden Unmapped Reads können in der qualitativen Metagenomik durch optionale massive BLAST-Analysen zugeordnet werden [38, 39]. Hierbei werden alle Unmapped Reads per lokalem BLAST mit der NCBI nonredundant nucleotide collection (nr/nt) abgeglichen. Somit können Spezies identifiziert werden, die nicht bei der initialen Auswahl der Referenzgenome berücksichtigt wurden.

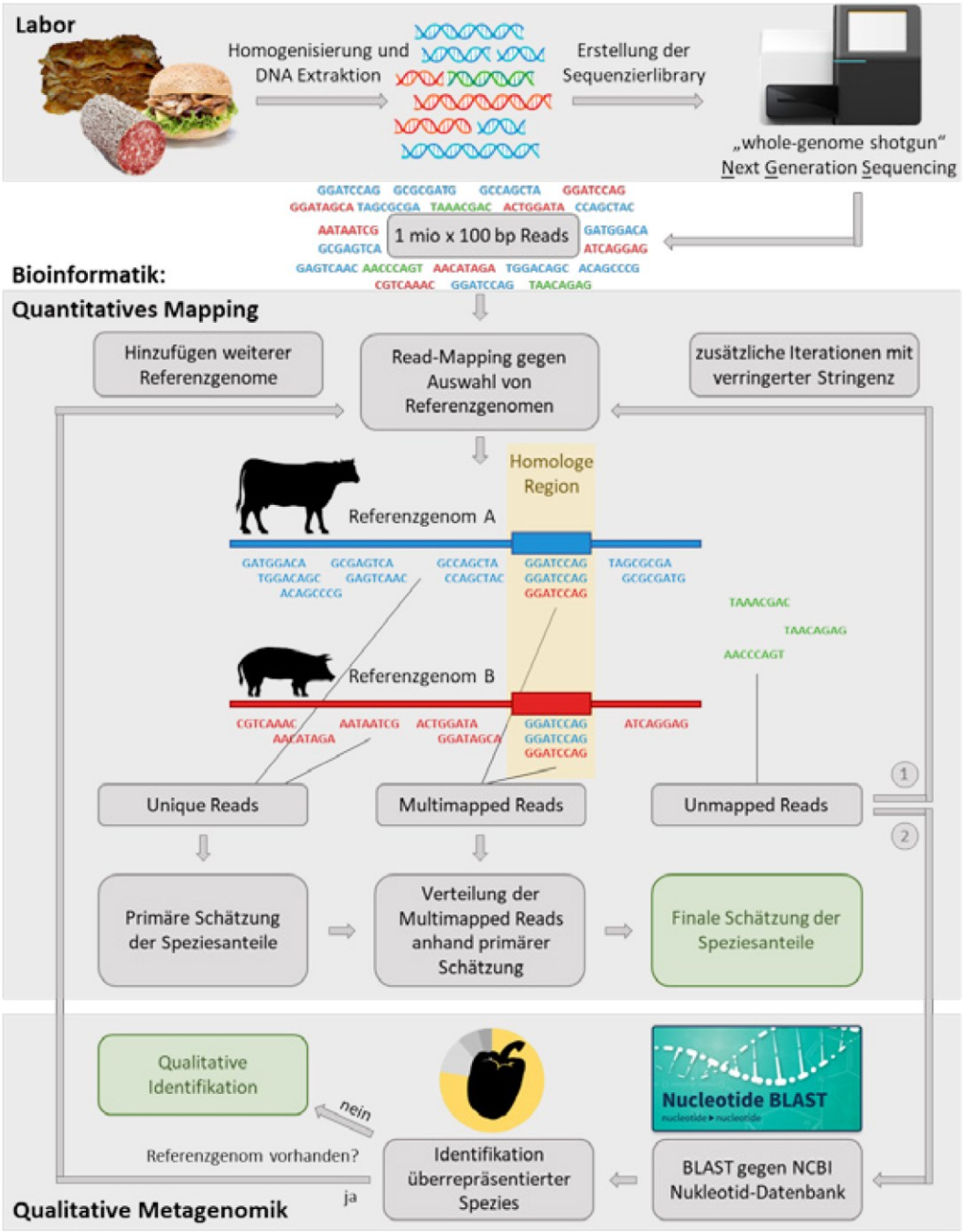

Abbildung 2: Workflow der AFS. Mit gängigen molekularbiologischen Methoden wird die gesamte DNA einer Lebensmittelprobe extrahiert, eine NGS library erstellt und auf einem Illumina-Gerät sequenziert. Im quantitativen Mapping werden in drei Iterationen die sequenzierten Reads an ausgewählte Referenzgenome kartiert und in Klassen unterteilt: Unique Reads treffen spezifisch an nur einem Referenzgenom, Multimapped Reads treffen aufgrund konservierter Sequenzen an mehreren Referenzgenomen. Multimapped Reads werden anhand des Verhältnisses von Unique Reads auf Spezies aufgeteilt. Unmapped Reads können keinem Referenzgenom zugeordnet werden und zeigen in der Regel zusätzliche in der Probe enthaltenen Spezies oder mikrobiologische Belastungen an. In der qualitativen Metagenomik wird die Speziesherkunft jedes ungemappten Reads durch BLAST-Analysen ermittelt. Sofern ein Referenzgenom für neu entdeckte Spezies vorhanden ist, kann das quantitative Mapping um entsprechende Spezies erweitert werden.

Wenn für eine dabei neu identifizierte Spezies ein Referenzgenom existiert, so kann der initiale quantitative Mappingschritt unter Einbeziehung dieser neu identifizierten Komponente wiederholt werden. Bei Fehlen eines passenden Referenzgenoms kann mit den Unmapped Reads zumindest eine qualitative Aussage über die Anwesenheit der Spezies im Lebensmittel getroffen werden. Durch diese qualitative Metagenomik sind selbst in Spuren vorhandene Zutaten wie Gewürze, Verunreinigungen oder Allergene in der Probe detektierbar. Auch eine Bestimmung des mikrobiologischen Artenspektrums ist auf diese Weise möglich.

6.3 AFS liefert exaktere Resultate als qPCR

Die Performanz der AFS-Methode wurde anhand von Wurstproben mit exakt bekannten Zusammensetzungen getestet, sogenannten Kalibrator-Würsten. Diese Proben wurden von einer professionellen Metzgerei zu Versuchszwecken nach drei Rezepturen erstellt: AllMeat (Rind, Schwein, Schaf, Pferd in variablen Anteilen), Lyoner (Rind, Schwein und in geringerem Umfang Huhn, Truthahn) sowie Geflügel-Lyoner (Huhn, Truthahn und in geringerem Umfang Rind, Schwein) [9, 10]. Viele Lebensmittel enthalten auch pflanzliche Komponenten, die bei Personen zu allergischen Reaktionen führen können. Daher wurden insgesamt 11 Pflanzen mit Allergiepotential in unterschiedlichen Mengen in die Wurstrezepturen eingearbeitet.

Die so konzipierten Proben wurden per AFS analysiert und die Ergebnisse mit denen konventioneller Methoden wie qPCR und droplet digital PCR (ddPCR) verglichen (**Abbildung 3**): In 9 von 13 Fällen lieferte AFS die besten Resultate, bei 3 Proben war eine ddPCR leicht besser und nur in einem Fall lieferte eine ddPCR eindeutig bessere Ergebnisse als AFS [40]. Für keine der betrachteten Proben konnte eine qPCR das beste Ergebnis liefern. Selbst Zutaten mit einem Anteil von 0,5 – 1 % konnten zuverlässig durch AFS detektiert werden. Noch keine quantitative Analytik war zum Zeitpunkt der Datenauswertung für die potenziell allergenen Pflanzenspezies möglich, da für diese nur ein einziges Referenzgenom existierte. Daher konnten diese Bestandteile nur durch qualitative Metagenomik untersucht werden. Alle Pflanzenspezies konnten dabei identifiziert werden [35]. Eine quantitative Aussage ist bei dieser BLAST-Analyse aufgrund der ungleichmäßigen Repräsentation von Spezies in der Datenbank nicht möglich. Zwar wurde eine Pflanzenart sogar mit nur einem Read identifiziert, vor allem bei Allergenen gehen wir jedoch davon aus, dass eine qualitative Aussage als Alarmsignal ausreichend ist.

Zur Einschätzung der Gefahr von möglichen falsch-positiven Resultaten bei der AFS-Analyse wurden zusätzlich zu den in den Würsten real enthaltenen Spezies-Komponenten zusätzlich nah-verwandte Spezies getestet.

Eine enge phylogenetische Beziehung von Arten bedingt eine höhere Sequenzähnlichkeit aufgrund konservierter DNA-Sequenzen, die wiederum zu uneindeutigen Read-Zuordnungen und somit zu qualitativ und quantitativ falschen Aussagen führen könnten. Es konnte gezeigt werden, dass phylogenetische Distanzen von ca. 10 Mio. Jahren (z.B. Schaf-Ziege oder Rind-Wasserbüffel) nur zu geringen Falsch-Zuordnungen bei der AFS führen und diese lebensmittelrelevanten Speziespaare somit sicher unterschieden werden können [40]. Durch in silico-Simulationen haben wir gezeigt, dass eine Erhöhung der Read-Leselänge bei der Sequenzierung auf 150 bp falsch-positive Treffer erwartungsgemäß verringert, da längere Reads häufiger diagnostische Unterschiede zwischen den Spezies enthalten. Durch diese technische Modifikation kann die falsch-positive Detektionsrate mit nur geringen Mehrkosten deutlich gesenkt werden.

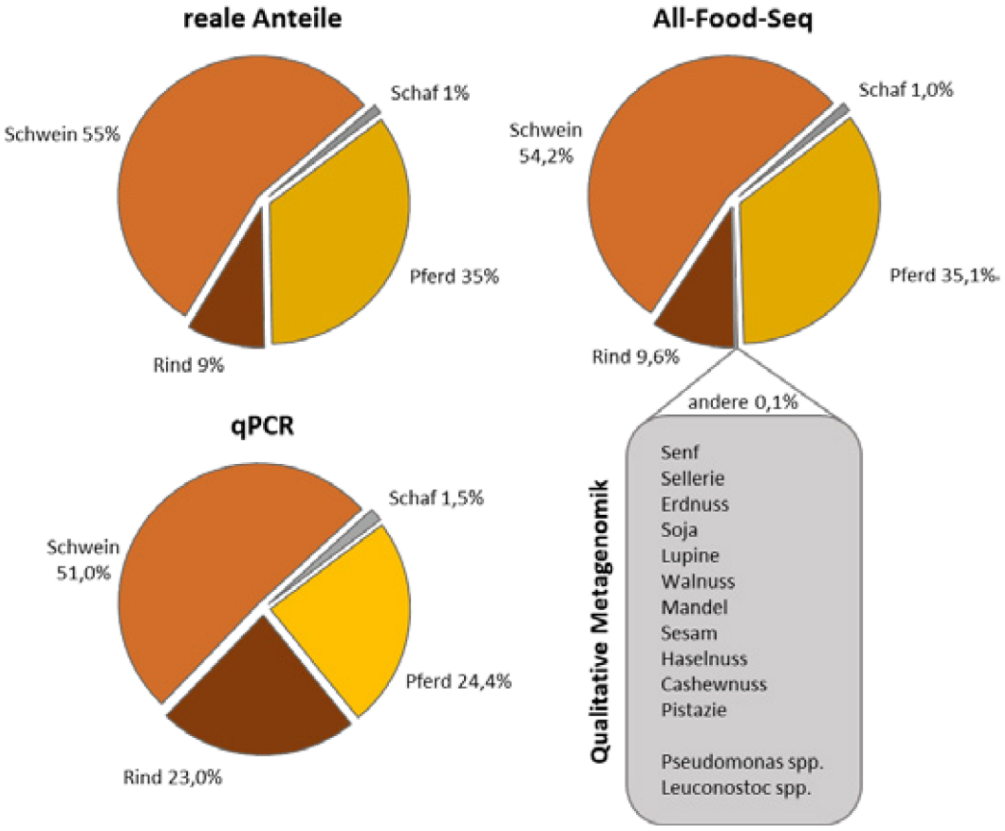

Abbildung 3: Vergleich der All-Food-Seq und qualitativer PCR am Beispiel des AllMeat-Kalibrators B. Die Ergebnisse der qualitativen PCR stammen aus [8].

6.4 AFS identifiziert „exotische“ Komponenten in der Praxis

Das Potential der AFS zur Detektion unerwarteter Komponenten sollte in der Praxis getestet werden, indem reale Lebensmittelproben sequenziert wurden. Diese umfassten fünf Döner Kebab von Imbissen aus dem Rhein-Main-Gebiet. Laut dem Bayerischen Landesamt für Gesundheit und Lebensmittelsicherheit (LGL) dürfen unter dem Namen Döner Kebab verkaufte Gerichte ausschließlich Fleisch von Lamm/Schaf oder Kalb/Rind enthalten; Hähnchen- oder Putenhaltige Gerichte müssen hingegen einen entsprechenden Namen gemäß der enthaltenen Geflügelart tragen [41]. Mehrere Studien berichten jedoch von Fehldeklarationen von Döner Kebab-Snacks, wobei häufig nicht deklariertes Geflügel oder in seltenen Fällen auch Schweinefleisch enthalten waren [42–44]. Bei der Untersuchung mit der AFS lag daher der Fokus auf der Identifikation der Fleischkomponenten, weshalb nur diese extrahiert und analysiert wurden. In drei von fünf Proben konnte Fleisch vom Truthahn mit je einem Anteil von über 70 % nachgewiesen werden, obgleich keine der Proben als Geflügeldöner verkauft wurde [40]. In drei Proben konnten geringe Mengen an nicht deklarierter Soja nachgewiesen werden, die womöglich aus einer Gewürzmischung der Marinade des Fleisches stammen könnte und für Konsumenten besonders im Hinblick auf Allergien von großer Relevanz ist. In einer Probe wurden außerdem nennenswerte Mengen an Mais gefunden, dessen Ursprung wir in der Salatbeilage des Döner Kebab vermuten.

Des Weiteren wurde eine Paella mit Huhn analysiert. Bei diesem Gericht handelt es sich um ein rezeptorisch komplexes Gericht mit vielen Zutaten in zum Teil geringen Mengenanteilen. Zur Vereinfachung dieser komplizierten Aufgabe wurden bei der Analyse mit der AFS nur die „Nicht-Reis“-Komponenten untersucht. Die meisten der Komponenten wie Huhn, Alaska Seelachs, Miesmuscheln, Tomaten und Paprika konnten zuverlässig detektiert werden [unveröffentlichte Ergebnisse]. Diese Probe zeigt aber auch eine Schwäche der AFS auf: für die im Gericht enthaltenen Erbsen sowie Zwiebeln existieren bis dato keine Referenzgenome, die jedoch eine zwingende Voraussetzung für die quantitative Analyse sind. Durch die daher notwendigerweise fehlerhafte Quantifizierung wurde der Anteil der real vorhandenen Spezies (mit Referenzgenom) überschätzt, sodass eine exakte mengenmäßige Aussage in diesem Fall nicht möglich war. Durch den zweiten Analyseschritt der AFS, die Metagenomik per BLAST-Analyse, war es indessen möglich, diese Spezies zumindest qualitativ zu identifizieren. Auf diese Weise konnte im Gericht entgegen der deklarierten Rezeptur Hefe detektiert werden (ein Befund, den der Hersteller auf Nachfrage bestätigte). Außerdem wurden geringe Mengen an *Nepetoideae* identifiziert. Dieser Pflanzen-Unterfamilie gehören diverse Gewürzpflanzen wie Basilikum, Oregano, Rosmarin, Thymian an. Auch geringe Spuren von Fadenwürmern wurden identifiziert. Da Seelachs häufig von diesen Parasiten befallen ist, erscheint dieser Weg in das Lebensmittel am naheliegendsten.

Im tiefgefrorenen Produkt stellen diese zwar keine Gesundheitsgefahr für den Menschen dar, sie sind jedoch Indiz für eine verspätete Entfernung des Bauchlappens bei der Bearbeitung des Fischs [45, 46]. Interessanterweise konnte die AFS die im Rezept angegebenen Shrimps nicht bestätigen, dafür wurde unerwartet Tintenfisch identifiziert. Wir vermuten, dass es sich hierbei um eine mögliche Änderung der Rezeptur oder eine Kontamination in der Produktionslinie handelt; diese Beobachtung hat der Hersteller jedoch nicht kommentiert. Diese Beispiele demonstrieren, dass die AFS als Screening-Methode sinnvoll einsetzbar ist. Ergänzt durch etablierte Methoden wie die gezielte PCR-Detektion hat die AFS ein großes Potential zur routinemäßigen Analyse von komplex zusammengesetzten Lebensmitteln.

6.5 Bakterien- und Phagen-Detektion ohne Mehraufwand

Die AFS basiert wie beschrieben auf einer Sequenzierung der gesamten DNA einer Probe. Dies bietet zeitgleich das Potential, das mikrobielle Spektrum ohne finanziellen Mehraufwand bedingt durch eine separate Typisierung testen zu können. Diese Stärke der AFS zeigte sich eindrucksvoll an den Geflügel-Lyoner-Proben: In diesen wurden die Bakterien *Brochothrix spp.*, *Pseudomonas spp.* und *Psychrobacter spp.* identifiziert [47]. Vertreter aller drei Gattungen sind dafür bekannt Lebensmittel zu verderben [48, 49]. Des Weiteren wurde in diesen Lyonerwürsten *Brochothrix* phage BL3 nachgewiesen [47]. Dieser Bakteriophage wird in der Lebensmittelindustrie eingesetzt, um die Haltbarkeit von verpacktem Fleisch zu verlängern [50, 51]. Somit zeigt AFS eindrucksvoll, dass selbst die Detektion von Viren und auch bioprozessierten Lebensmitteln möglich sind.

6.6 Limitation durch Referenzgenome

Bedingung für die Funktionsweise der AFS ist das Vorhandensein von Referenzgenomen für möglichst viele zu testende Spezies. In der Vergangenheit waren solche Genomsequenzen nur kostenintensiv und sehr aufwändig herzustellen. So hat etwa das human genome project vor zwanzig Jahren mehrere Tausend Wissenschaftler über einen Zeitraum von über 10 Jahren beschäftigt und fast 3 Mrd. US-Dollar gekostet [52, 53]. Die Kosten einer Sequenzierung sind jedoch erheblich gesunken (z.B. auf ca. 1.500 € pro Humangenom), und verbesserte Computertechnologie sowie Algorithmik haben den zur Assemblierung benötigten Rechenaufwand deutlich reduziert. Heute ist daher die Erstellung eines Referenzgenoms einer Spezies auch für kleine Labore erschwinglich und der Arbeitsaufwand von einem Doktoranden in wenigen Wochen zu bewerkstelligen.
Als weiter kostensparende Alternative haben wir getestet, ob die AFS auch mit sehr schwach redundant sequenzierten und daher äußerst kostengünstig erstellten Genomsequenzen funktioniert.

Solche Genome, die typischerweise mit einer Redundanz von nur 5-10x produziert werden, erreichen selbstverständlich bei weitem nicht die hohe Qualität von gründlich erarbeiteten Referenzgenomen wie dem des Menschen. Die Basensequenz ist jedoch in der Regel nahezu vollständig vorhanden (> 90 %), liegt aber eben stark fragmentiert vor. Dennoch eignen sich diese günstig erstellten Genomsequenzen offenbar hervorragend für analytische Zwecke wie die AFS [unveröffentlichte Ergebnisse]. Zusätzlich planen große Genom-Konsortien die Sequenzierung von vielen Tausend Genomen [54, 55]. Daher ist davon auszugehen, dass in einigen Jahren Referenzgenome für die wichtigsten aller eukaryotischen Spezies vorhanden sein werden. Vermutlich werden damit alle lebensmittelrelevanten Spezies schon früher abgedeckt sein, sodass diese Limitierung der AFS voraussichtlich bald entfällt.

6.7 Schnellere Analysen durch algorithmische Weiterentwicklung

Während einer AFS-Analyse müssen bei der Kartierung der Reads alle Referenzgenome in den Arbeitsspeicher des Rechners geladen werden. Die Menge an benötigten Computerressourcen nimmt schon heute zum Teil problematische Dimensionen an und wird künftig mit mehr verfügbaren Referenzgenomen weiter steigen. Zum Teil werden diese Herausforderungen sicher durch Fortschritte in der Computertechnik kompensiert werden. Dennoch wird parallel eine Reduktion der Rechenlast erforderlich. Wir entwickeln daher die Software AFS-MetaCache, die durch geschickte Reduktion der zu durchsuchenden Genomgröße den Rechenaufwand stark reduziert [47, 56]. Die Vorgehensweise führt MinHash als algorithmisches Prinzip in DNA-Analyseverfahren ein: diese informatische data mining-Technik berechnet den Jaccard-Koeffizienten zweier Elemente und bestimmt deren Schnittmenge, sodass zügig Übereinstimmungen detektiert werden können. Die Methode stammt ursprünglich aus der Webentwicklung und wird u.a. von Internet-Suchmaschinen verwendet, um sich ähnelnde Websites zu filtern.

AFS-MetaCache unterteilt sich in zwei Arbeitsschritte: 1) einmalige Erstellung der Datenbank mit den Referenzgenomen und 2) Analyse einer Probe durch Abgleich der Reads mit der erstellten Datenbank. Bei der Konstruktion der Datenbank wird jedes zur Verfügung stehende Referenzgenom in Windows der Länge l unterteilt, die einer genomischen Region entsprechen (**Abbildung 4**). Innerhalb eines jeden Windows werden alle k Nukleotide langen Sequenzen, sog. k-mers, extrahiert und in eine Liste geschrieben. Auf jedes Element dieser Liste wird nun eine Hashfunktion angewandt, die jedem k-mer basierend auf seiner Sequenz einen eindeutigen Zahlenwert zuteilt. Nun folgt der Vorteil, der zur enormen Reduktion an Hardware-Ressourcen führt.

Anstatt wie rechenintensive Mapping-Algorithmen die gesamte verfügbare Information als Referenz zu nutzen, reduziert AFS-MetaCache die Sequenz der Referenzgenome nach einem reproduzierbaren Prinzip: aus der Liste an Hash-Werten jedes Windows werden nicht alle, sondern nur die s kleinsten k-mere für die folgende Analyse in die Datenbank geschrieben. Dadurch reduziert sich die Größe jedes zu durchsuchenden Referenzgenoms, während jedoch alle genomischen Regionen abgebildet bleiben.
Beim zweiten Schritt, der Klassifikation der sequenzierten Reads aus der Lebensmittelprobe, wird nun sehr ähnlich vorgegangen: Jeder sequenzierte Read wird in seine k-mere zerlegt und diese in einer Liste gespeichert, jeweils eine Hashfunktion angewandt und zuletzt nur die s kleinsten k-mere betrachtet (Abbildung 4). Anstatt die ganze Sequenz eines Reads zu vergleichen, müssen somit nur s k-mere ausgewertet werden, um die genomische Region des Reads und damit seine Herkunft und Spezies-Zuordnung zu identifizieren. Diese Reduktion des Suchaufwands hat eine über 400-fache Steigerung der Geschwindigkeit bei gleichbleibender Sensitivität und Spezifität zur Folge, sodass die bioinformatische Auswertung einer Probe mit 10 Mio. Reads auf deutlich unter 1 Minute reduziert werden kann [47, 56].

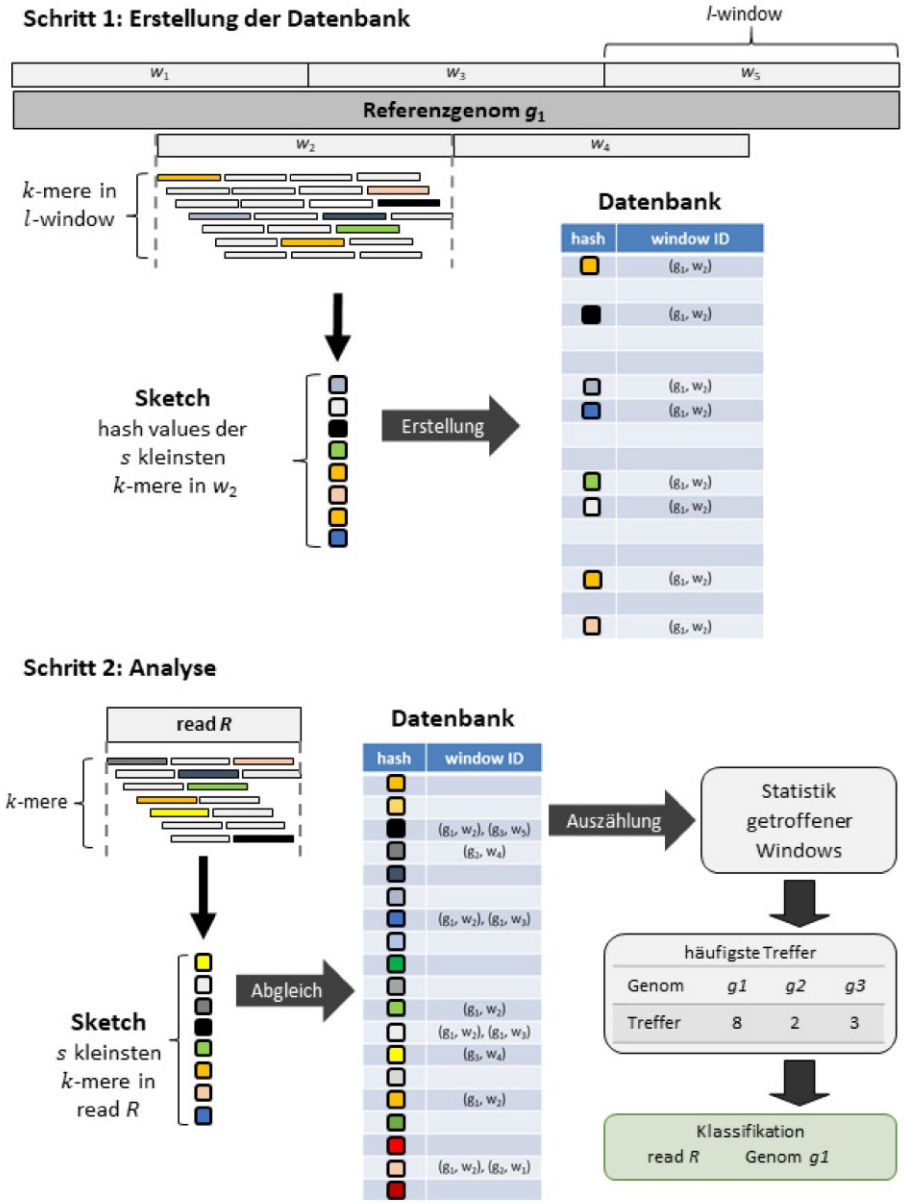

Abbildung 4: Konzeptioneller Ablauf von AFS-MetaCache. Die Methode läuft in zwei Schritten ab: Zunächst wird einmalig eine Datenbank aus allen zur Verfügung stehenden Referenzgenomen erstellt. Dabei wird jedes Genom in windows unterteilt. Anschließend wird für alle in einem window enthaltenen k-mere ein Hash-Wert berechnet. Nur die k-mere mit den s kleinsten Hash-Werten werden in die Datenbank übertragen. Der zweite AFS-MetaCache Schritt ist die Analyse: Jeder zu analysierende Read wird in seine k-mere zerlegt und deren Hash-Werte berechnet. Die s kleinsten Hash-Werte jeden Reads werden mit der Datenbank abgeglichen. Basierend auf den getroffenen k-meren in der Datenbank wird ermittelt, an welches window und Genom der jeweilige Read kartiert werden kann.

6.8 Kosten der AFS

Die Materialkosten einer AFS-Analyse reflektieren größtenteils die Reagenzien zur Erstellung der Illumina NGS-Library. Die Reagenzien für die Sequenzierung selbst von etwa 1 Mio. Reads sind hingegen sehr günstig. Kommerzielle Anbieter offerieren beispielsweise die Analyse von 40 Proben (parallel) mit je 2x 1 Mio. Reads („paired-end"-Sequenzierung) und 75 bp Leselänge für ca. 120€ pro Probe [57].

6.9 Zusammenfassung und Ausblick

Die All-Food-Seq (AFS) ist eine Screening-Methode zur Analyse von Lebensmittelkomponenten basierend auf Next-Generation Sequencing der Gesamt-DNA. Verglichen mit etablierten Methoden wie der qPCR oder ddPCR liefert die AFS eine vergleichbare oder sogar bessere Quantifizierungsaussage. Bedingung für eine Analyse mit der AFS ist das Vorhandensein von Referenzgenomen der zu untersuchenden Spezies. Durch kontinuierlich sinkende Kosten für Genomdaten ist mit deren Verfügbarkeit für die meisten Lebensmittel-relevanten Spezies in wenigen Jahren zu rechnen. Der Preis pro Analyse wird vermutlich ebenfalls weiter sinken.

Prinzip-bedingt ermöglicht AFS das Screening eines Lebensmittels ohne a priori-Annahmen, z.B. durch den Einsatz von PCR-Systemen. Außerdem stellt die Methode einen All-in-One Ansatz dar, da neben tierischen und pflanzlichen Komponenten auch die mikrobielle Belastung festgestellt werden kann. Dabei ist es unerheblich, ob es sich um Parasiten, Pilze, Bakterien und gar Viren handelt: Vertreter aller Domänen des Lebens können in nur einem Experiment parallel detektiert werden. Aktuell lassen sich Mikroorganismen zwar identifizieren, aber nur bedingt quantifizieren. Wir arbeiten daher daran, den Zusammenhang zwischen identifizierten bakteriellen Reads und koloniebildenden Einheiten herzustellen. Hierdurch wird es potenziell möglich, per AFS eine Aussage z.B. über die Anzahl von Fäulnisbakterien und somit den Frischegrad des Lebensmittels zu treffen. Allergene können in der AFS derzeit häufig nur qualitativ identifiziert werden; für viele relevante Pflanzenspezies fehlen aktuell noch die Referenzgenome.

Fortschritte in der Sequenziertechnologie werden sich vermutlich sehr positiv auf die AFS in der Praxis auswirken: aktuell testen wir die Eignung des Nanoporen-Sequenzierprinzips für die AFS. Diese ultraportable Sequenziertechnologie ermöglicht Analysen am Ort der Probennahme, also z.B. auf dem Markt oder direkt beim Produzenten, und damit in Echtzeit. Es entfällt die Notwendigkeit des Probentransports in ein Labor, was vor allem bei der Kontrolle von schnell verderblichen Lebensmitteln von Vorteil ist. Besonders die Kombination mit schnelleren Analyse-Algorithmen wie AFS-MetaCache ist dabei erfolgsversprechend. Somit kann die aktuell gängige Praxis der Beanstandung einer Probe im Nachhinein aufgrund der großen Zeitersparnis bei der Analyse schon bald der Vergangenheit angehören.

6.10 Danksagung

TH und SLH danken dem Bundesamt für Landwirtschaft und Ernährung (BLE, Förderkennzeichen 2816503814), dem Center for Computational Sciences der JGU Mainz (CSM) und dem Bundesministerium der Justiz und für Verbraucherschutz Rheinland-Pfalz für finanzielle Unterstützung.

6.11 Quellen

1. **Wong EHK, Hanner RH** (2008) *DNA barcoding detects market substitution in North American seafood*. Food Res Int 41:828–837

2. **Luber F., Demmel A., Hosken A., *et al.*** (2012) *Apricot DNA as an indicator for persipan: Detection and quantitation in marzipan using ligation-dependent probe amplification*. J Agric Food Chem 60:5853–5858

3. **Iwobi A., Sebah D., Kraemer I., *et al.*** (2015) *A multiplex real-time PCR method for the quantification of beef and pork fractions in minced meat*. Food Chem 169:305–313

4. **Cohen NJ, Deeds JR, Wong ES, *et al.*** (2009) *Public health response to puffer fish (Tetrodotoxin) poisoning from mislabeled product*. J Food Prot 72:810–817

5. **Doosti A, Ghasemi Dehkordi P, Rahimi E** (2014) *Molecular assay to fraud identification of meat products*. J Food Sci Technol 51:148–152

6. **Bundesinstitut für Risikobewertung**: *Problematik der Lebensmittelinfektion.* https://www.bfr.bund.de/de/problematik_der_lebensmittelinfektion-11100.html. Accessed 24 Feb 2020

7. **Verwaltungsgericht Kassel** (2019) *Betriebsschließung Firma Wilke*. https://verwaltungsgerichtsbarkeit.hessen.de/pressemitteilungen/betriebsschließung-firma-wilke-0. Accessed 24 Feb 2020

8. **Köppel R., Ruf J., Zimmerli F., Breitenmoser A.** (2008) *Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, chicken and turkey*. Eur Food Res Technol 227:1199–1203

9. **Eugster A., Ruf J., Rentsch J., Köppel R.** (2009) *Quantification of beef, pork, chicken and turkey proportions in sausages: use of matrix-adapted standards and comparison of single versus multiplex PCR in an interlaboratory trial*. Eur Food Res Technol 230:55–61

10. **Köppel R., Ruf J., Rentsch J.** (2011) *Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, horse and sheep*. Eur Food Res Technol 232:151–155

11. **Ulca P., Balta H., Çağın İ., Senyuva HZ** (2013) *Meat species identification and Halal authentication using PCR analysis of raw and cooked traditional Turkish foods*. Meat Sci 94:280–284

12. **Floren C., Wiedemann I., Brenig B., *et al.*** (2015) *Species identification and quantification in meat and meat products using droplet digital PCR (ddPCR)*. Food Chem 173:1054–1058

13. **Liu S., Xu K., Wu Z., *et al.*** (2016) *Identification of five highly priced tuna species by quantitative real-time polymerase chain reaction*. Mitochondrial DNA 27:3270–3279

14. **Song K.-Y., Hwang HJ, Kim JH** (2017) *Ultra-fast DNA-based multiplex convection PCR method for meat species identification with possible on-site applications*. Food Chem 229:341–346

15. **Spielmann G, Gerdes L, Miller A, *et al.*** (2018) *Molecular biological species identification of animal samples from Asian buffets*. J fur Verbraucherschutz und Leb 13:271–278

16. **Spielmann G., Diedrich J., Haszprunar G., *et al.*** (2019) *Comparison of three DNA marker regions for identification of food relevant crustaceans of the order Decapoda*. Eur Food Res Technol 245:987–995

17. **Köppel R., Ganeshan A., Weber S., *et al.*** (2019) *Duplex digital PCR for the determination of meat proportions of sausages containing meat from chicken, turkey, horse, cow, pig and sheep*. Eur Food Res Technol 245:853–862

18. **Roslin T., Majaneva S.** (2016) *The use of DNA barcodes in food web construction—terrestrial and aquatic ecologists unite!* Genome 59:603–628

19. **Staats M., Arulandhu AJ, Gravendeel B., *et al.*** (2016) *Advances in DNA metabarcoding for food and wildlife forensic species identification*. Anal. Bioanal. Chem. 408:4615–4630

20. **Böhme K., Calo-Mata P., Barros-Velázquez J., Ortea I.** (2019) *Review of Recent DNA-Based Methods for Main Food-Authentication Topics*. J Agric Food Chem 67:3854–3864

21. **Ratnasingham S., Hebert PDN** (2007) *BOLD: The Barcode of Life Data System: Barcoding*. Mol Ecol Notes 7:355–364

22. **Tillmar AO, Dell'Amico B., Welander J., Holmlund G.** (2013) *A universal method for species identification of mammals utilizing next generation sequencing for the analysis of DNA mixtures*. PLoS One 8:e83761

23. **Zhou X., Li Y., Liu S., *et al.*** (2013) *Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification*. Gigascience

24. **Newmaster SG, Grguric M., Shanmughanandhan D., *et al.*** (2013) *DNA barcoding detects contamination and substitution in North American herbal products*. BMC Med 11:1–13

25. **Deagle BE, Thomas AC, Shaffer AK, *et al.*** (2013) *Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: Which counts count?* Mol Ecol Resour 13:620–633

26. **Markoulatos P., Siafakas N., Moncany M.** (2002) *Multiplex polymerase chain reaction: A practical approach*. J Clin Lab Anal 16:47–51

27. **Berry D., Mahfoudh K. Ben, Wagner M., Loy A.** (2011) *Barcoded Primers Used in Multiplex Amplicon Pyrosequencing Bias Amplification*. Appl Environ Microbiol 77:7846–7849

28. **Tedersoo L., Anslan S., Bahram M., *et al.*** (2015) *Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi*. MycoKeys 10:1–43.

29. **Sze MA, Schloss PD** (2019) *The Impact of DNA Polymerase and Number of Rounds of Amplification in PCR on 16S rRNA Gene Sequence Data*. mSphere 4:e00163-19.

30. **Dunham I., Kundaje A., Aldred SF, *et al.*** (2012) *An integrated encyclopedia of DNA elements in the human genome*. Nature 489:57–74

31. **Kerstens HHD, Crooijmans RPMA, Veenendaal A, *et al.*** (2009) *Large scale single nucleotide polymorphism discovery in unsequenced genomes using second generation high throughput sequencing technology: Applied to Turkey*. BMC Genomics 10:479

32. **Kijas JW, Townley D., Dalrymple BP, *et al.*** (2009) *A Genome Wide Survey of SNP Variation Reveals the Genetic Structure of Sheep Breeds*. PLoS One 4:e4668

33. **Wade CM, Giulotto E., Sigurdsson S., *et al.*** (2009) *Genome sequence, comparative analysis, and population genetics of the domestic horse*. Science (80-) 326:865–867

34. **Wiedmann RT, Smith TPL, Nonneman DJ** (2008) *SNP discovery in swine by reduced representation and high throughput pyrosequencing*. BMC Genet 9:81

35. **Ripp F., Krombholz C., Liu Y., *et al.*** (2014) *All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing*. BMC Genomics 15:639

36. **Liu Y., Ripp F., Koeppel R., *et al.*** (2017) *AFS: identification and quantification of species composition by metagenomic sequencing*. Bioinformatics 33:btw822

37. **Li H., Durbin R.** (2009) *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics 25:1754–1760

38. **Altschul SF, Gish W., Miller W., *et al.*** (1990) *Basic local alignment search tool.* J Mol Biol 215:403–410

39. **Camacho C., Coulouris G., Avagyan V., *et al.*** (2009) BLAST+: *Architecture and applications*. BMC Bioinformatics 10

40. **Hellmann SL, Ripp F., Bikar SE, *et al.*** (2020) *Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq)*. Eur Food Res Technol 246:193–200

41. **Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit** (2020) *Kenntlichmachung von „Döner Kebab“ und „ähnlichen“ Erzeugnissen bei loser Abgabe*. https://www.lgl.bayern.de/downloads/lebensmittel/doc/merkblatt_doener_kebab.pdf. Accessed 24 Feb 2020

42. **Norddeutscher Rundfunk** (2016) *Häufig Fleischbrät und Zusatzstoffe im Döner.* https://www.ndr.de/ratgeber/verbraucher/Haeufig-Fleischbraet-und-Zusatzstoffe-im-Doener,doener164.html. Accessed 24 Feb 2020

43. **LACROS** (2009) *The composition and labelling of doner kebabs*. Lacors, Local Authorities Coord Regul Serv

44. **Liuzzo G., Rossi R., Giacometti F., *et al.*** (2016) *Mislabelling of döner kebab sold in Italy*. Ital J Food Saf 5

45. **Europäisches Parlament** (2004) *Verordnung (EG) Nr. 853/2004*. https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2004:139:0055:0205:DE:PDF. Accessed 24 Feb 2020

46. **Deutsche See**: *Nematoden*. https://www.deutschesee.de/wissen/fisch-ernaehrung/nematoden/. Accessed 24 Feb 2020

47. **Kobus R., Abuín JM, Müller A., *et al.*** (2020) *A big data approach to metagenomics for All-Food-Sequencing*. BMC Bioinformatics 21:102

48. **Borch E., Kant-Muermans ML, Blixt Y.** (1996) *Bacterial spoilage of meat and cured meat products*. Int J Food Microbiol 33:103–120

49. **Gennari M., Parini M., Volpon D., Serio M.** (1992) *Isolation and characterization by conventional methods and genetic transformation of Psychrobacter and Acinetobacter from fresh and spoiled meat, milk and cheese*. Int J Food Microbiol 15:61–75

50. **Greer GG, Dilts BD** (2002) *Control of Brochothrix thermosphacta spoilage of pork adipose tissue using bacteriophages*. J Food Prot 65:861–863

51. **Moye ZD, Woolston J., Sulakvelidze A** (2018) *Bacteriophage applications for food production and processing*. Viruses 10

52. **National Human Genome Research Institute** (2001) *The Human Genome Project*. https://www.genome.gov/human-genome-project. Accessed 24 Feb 2020

53. **Lander ES, Linton LM, Birren B, *et al.*** (2001) *Initial sequencing and analysis of the human genome*. Nature 409:860–921

54. **G10K Consortium**: *Vertebrate Genomes Project*. https://vertebrategenomesproject.org/. Accessed 24 Feb 2020

55. **G10K Consortium**: *Earth BioGenome Project*. https://genome10k.soe.ucsc.edu/earth-biogenome/. Accessed 24 Feb 2020

56. **Müller A., Hundt C., Hildebrandt A., *et al.*** (2017) *MetaCache: Context-aware classification of metagenomic reads using minhashing*. Bioinformatics 33:3740–3748

57. **StarSEQ**: *NGS Sequenzierung*. https://www.starseq.com/. Accessed 24 Feb 2020

2.3 A Big Data Approach to Metagenomics for All-Food-Sequencing

Kobus R, Abuín JM, Müller A, Hellmann SL, Pichel JC, Pena TF, Hildebrandt A, Hankeln T, Schmidt B

Published in: BMC Bioinformatics

DOI: https://doi.org/10.1186/s12859-020-3429-6

Own contributions to this publication:

- Comparison of AFS-MetaCache to prior AFS: Benchmarking against the alignment-based AFS for quantification accuracy and computational efficiency on AFS10 database (with Dr. R. Kobus)
- Reference database curation: Selecting and integrating 31 food-relevant eukaryote genomes plus RefSeq Release 90 bacteria/viruses/archaea (with Dr. R. Kobus and A. Müller)
- Writing: Original draft and review

Kobus *et al. BMC Bioinformatics* (2020) 21:102
https://doi.org/10.1186/s12859-020-3429-6

BMC Bioinformatics

SOFTWARE Open Access

A big data approach to metagenomics for all-food-sequencing

Robin Kobus[1†], José M. Abuín[2,3†], André Müller[1], Sören Lukas Hellmann[4], Juan C. Pichel[3], Tomás F. Pena[3], Andreas Hildebrandt[1], Thomas Hankeln[4] and Bertil Schmidt[1*]

Abstract

Background: All-Food-Sequencing (AFS) is an untargeted metagenomic sequencing method that allows for the detection and quantification of food ingredients including animals, plants, and microbiota. While this approach avoids some of the shortcomings of targeted PCR-based methods, it requires the comparison of sequence reads to large collections of reference genomes. The steadily increasing amount of available reference genomes establishes the need for efficient big data approaches.

Results: We introduce an alignment-free *k*-mer based method for detection and quantification of species composition in food and other complex biological matters. It is orders-of-magnitude faster than our previous alignment-based AFS pipeline. In comparison to the established tools CLARK, Kraken2, and Kraken2+Bracken it is superior in terms of false-positive rate and quantification accuracy. Furthermore, the usage of an efficient database partitioning scheme allows for the processing of massive collections of reference genomes with reduced memory requirements on a workstation (AFS-MetaCache) or on a Spark-based compute cluster (MetaCacheSpark).

Conclusions: We present a fast yet accurate screening method for whole genome shotgun sequencing-based biosurveillance applications such as food testing. By relying on a big data approach it can scale efficiently towards large-scale collections of complex eukaryotic and bacterial reference genomes. AFS-MetaCache and MetaCacheSpark are suitable tools for broad-scale metagenomic screening applications. They are available at https://muellan.github.io/metacache/afs.html (C++ version for a workstation) and https://github.com/jmabuin/MetaCacheSpark (Spark version for big data clusters).

Keywords: Next-generation sequencing, Metagenomics, Species identification, Eukaryotic genomes, Locality sensitive hashing, Big data

Background

Monitoring of food ingredients is becoming an increasingly important task. Relevant issues include correct labeling, fraud detection, and assessment of health risks [1]. This motivates the need for analytical methods that allow for accurate determination and quantification of food ingredients ideally spanning all kingdoms of life including animals, plants, bacteria, fungi, and possibly even viruses.

Quantitative real-time polymerase chain reaction (qPCR) [2] and droplet digital PCR (ddPCR) [3] are DNA-based technologies for food control that are widely used in practice. Unfortunately, these methods are limited by the number of target species within a single assay and thus are not suitable for broad-scale species screening. Similar restrictions apply to approaches based on sequencing of species-specific DNA bar codes [4].

High-throughput sequencing of total metagenomic DNA from biological samples provides the possibility to screen for a wide range of species as it does not require any prior definition of possible target species. However, subsequent bioinformatic analysis of large amounts

*Correspondence: bertil.schmidt@uni-mainz.de
†Robin Kobus and José M. Abuín contributed equally to this work.
[1]Department of Computer Science, Johannes Gutenberg University, 55099 Mainz, Germany
Full list of author information is available at the end of the article

of sequence-reads is required to identify and quantify actual food components. Our All-Food-Seq (AFS) pipeline [5, 6] maps each sequenced read to a number of reference genomes and then determines species composition and relative quantities based on a read counting procedure. Evaluation based on simulated as well as real data has demonstrated that AFS can detect anticipated species in food products and achieve quantification accuracy comparable to qPCR.

However, the AFS pipeline relies on applying a read alignment tool (such as BWA [7–9], Bowtie2 [10], or CUSHAW [11]) for each considered reference genome. Thus, runtime scales linearly with the number of considered genomes. For example, the quantification of a typical short read dataset consisting of a few million reads using ten mammalian and avian reference genomes with the BWA-based AFS pipeline already requires several hours on a standard workstation (not including the time for index construction). For broader scale screening of many species a much larger amount of reference genomes would be required, making this approach unfeasible.

More recently, a number of innovative techniques for fast taxonomic labeling in the field of bacterial metagenomics have been proposed. Wood and Salzberg [12] demonstrated that a *k*-mer-based exact matching approach can achieve high read classification accuracy while being around three orders-of-magnitude faster than the alignment tool MegaBLAST. It relies on building a database of all substrings of length *k* of each considered (bacterial) reference genome. A read is classified by querying the database using each of its *k*-mers as query. If a query returns a match a counter for the corresponding reference genome(s) is incremented. Finally, a read is taxonomically labeled based on high-scoring counters. Recent benchmark studies [13, 14] demonstrated that *k*-mer based tools such as Kraken [12], Kraken2+Bracken [15], CLARK [16], and MetaCache [17] can produce superior read assignment accuracy compared to several other tools including MetaPhlAn [18], mOTU [19], QIIME [20], and Kaiju [21] for selected bacterial metagenomic datasets. While being accurate, the major drawback of the *k*-mer based approach is high main memory consumption and long database construction times. For typical bacterial reference genome sets the databases used by Kraken and CLARK already consume several hundreds of gigabytes in size. The significantly higher complexities of eukaryotic reference genomes relevant for monitoring food ingredients therefore make an extension of this method to food-monitoring challenging.

Here, we present a novel computational method for broad-scale detection and quantification of species composition in food and other complex biological matters. It is based on our recently introduced MetaCache [17] bacterial metagenomic read classification algorithm. We employ a big data technique called minhashing to subsample *k*-mers in an intelligent way, thereby reducing the amount of stored *k*-mers by an order-of-magnitude. In this paper we show how this method can be extended from the taxonomic labeling of bacterial reads to the detection and quantification of ingredients in food samples that can span various kingdoms of life. MetaCache is augmented with the ability to estimate the abundance of organisms at a selectable taxonomic level as well as the possibility to filter out target references based on sequence coverage. Furthermore, we combine the minhashing algorithm used by MetaCache with efficient partitioning schemes. This allows us to employ databases that index large collections of reference genomes efficiently in terms of both construction times and memory consumption. We present two partitioning schemes and provide corresponding implementations for standard workstations based on C++ (AFS-MetaCache) and for big data clusters based on Apache Spark (MetaCacheSpark). Both version can be used as substitutes for the alignment tools previously employed in the AFS pipeline.

Our experimental results using a number of sequenced calibrator sausages of known species composition show that AFS-MetaCache runs orders-of-magnitude faster than the alignment-based AFS pipeline while yielding similar results. Furthermore, AFS-MetaCache and MetaCacheSpark yield lower false-positive rates and higher quantification accuracy compared to Kraken2, Kraken2+Bracken, and CLARK. They also provide faster database construction times and competitive query speeds. Our database partitioning scheme allows the reduction of peak main memory consumption on a single workstation or a cluster node significantly and therefore enables scalability to growing genome collections.

Implementation

Approach

Many tools in metagenomics struggle to keep pace with the increasing amount of available reference genomes. We address this issue by aiming at species identification and quantification at a large scale by using a combination of two big data techniques.

Minhashing: We adopt *minhashing* – a locality sensitive hashing (LSH) based data subsampling technique. It has been successfully applied by search engines to detect near duplicate web pages [22] but has recently gained popularity in bioinformatics with example applications including genome assembly [23], sequence clustering [24], and privacy-preserving read mapping [25]. Mash Screen [26] also employs minhashing for metagenomic analysis. While it allows to identify genomes contained in a sample, Mash Screen is not able to classify individual reads or quantify abundances by itself.

Partitioning: Because the RAM of a single workstation or a cluster node can become insufficient to hold a complete reference database, we employ a partitioning scheme to divide reference sequences into multiple chunks. The partitions can be queried successively on a single workstation or among multiple worker nodes of a distributed compute cluster. In order to support these two types of compute resources we have developed (*i*) AFS-MetaCache: a C++ version for individual workstations, and (*ii*) MetaCacheSpark: a distributed version based on the big data analytics engine Apache Spark [27] for compute clusters.

Database construction

Consider a collection G of m genomic sequences (reference genomes). Each reference genome is divided into windows of size l which overlap by $k-1$ base-pairs. Typically, l is of similar size to the anticipated read length (e.g. $l = 128$ for Illumina data as default). For each window a *sketch* is calculated using minhashing. A sketch consists of the s smallest k-mers (in strand-neutral canonical representation) contained in the window with respect to an applied hash function h_1. Thus, the sketching procedure selects only a subset of k-mers to be inserted into the database used for similarity computation. Assuming unique k-mers, the subsampling factor can be determined as $S = \frac{l-k+1}{s}$; i.e. for typical values such as $s = 8$, $k = 16$, and $l = 128$ this corresponds to a data reduction by over an order-of-magnitude ($S = 14.125$). Besides providing data reduction, minhashing also exhibits a desirable mathematical property when comparing two sketches: The relative intersection ratio between two sketched windows approximates the true Jaccard index evaluated on the whole k-mer space [22].

The hash table (database) for a given collection of reference genomes is constructed using open addressing. The entries of the hash table consist of key-target-list pairs. An associated hash function h_2 maps k-mers to slots in the hash table. If an identified slot is empty or occupied with the same k-mer, the corresponding k-mer is inserted as key and the corresponding location (genome ID, window ID) is appended to the target-list. If the slot is occupied by a different k-mer quadratic probing is used to iterate

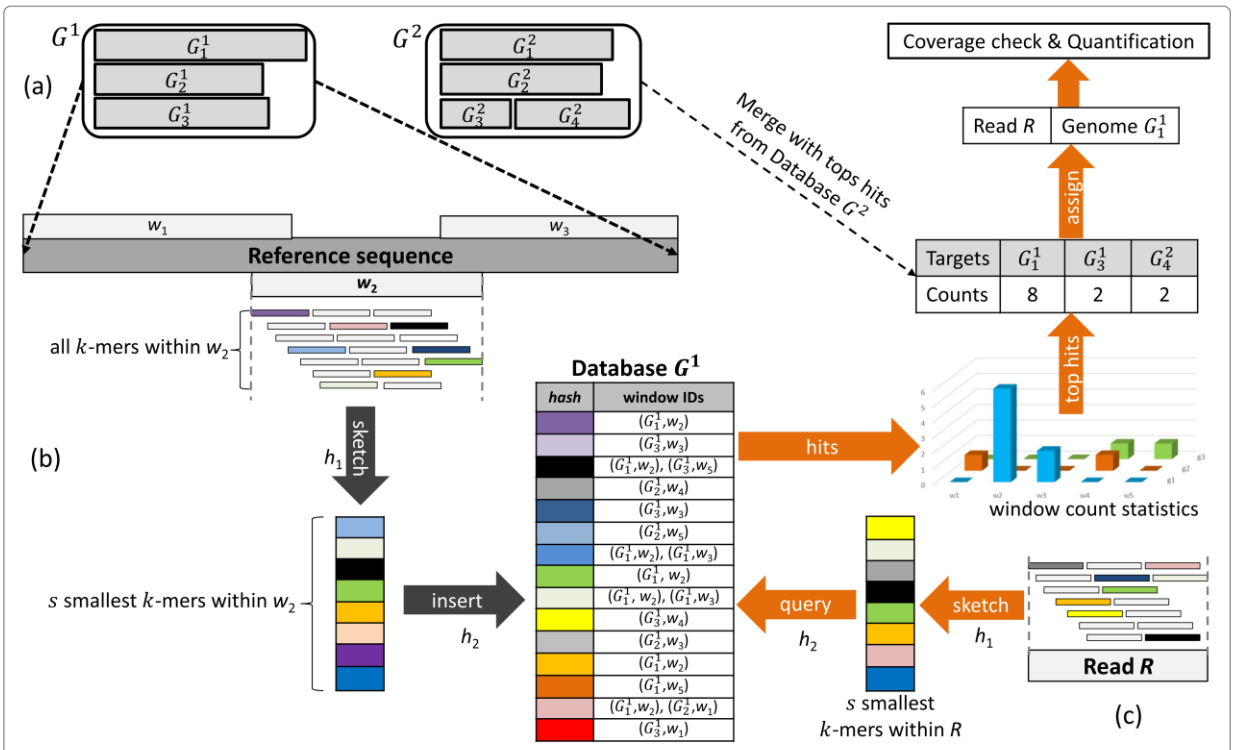

Fig. 1 Workflow: (**a**) Partitioning: reference sequences are divided into the sets G^1 and G^2. Each reference is further partitioned into slightly overlapping windows w_i. (**b**) Database construction: the s smallest k-mers of each window are computed and inserted into the database. (**c**) Classification: a database is queried with the s smallest k-mers of a read. The returned hits are used to count the number of hits within each window. Target reference genomes are identified by high scores in the window count statistics. In case of several partitions, the top hits from querying each database need to be merged in order to assign a read to a reference genome. After all reads have been processed, coverage check and quantification are performed

to the next slot. Target lists have a pre-defined maximum length. If the maximum length is reached, the corresponding k-mer is considered uninformative and deleted from the hash table at the end of the construction.

In the big data scenario we need to consider cases where the database is too large to fit into the RAM of a single workstation or a cluster node. Hence, it needs to be split into multiple parts which can be queried successively or distributed among multiple worker nodes of a cluster. Partitioning divides the collection of reference genomes G of total base-pair length M into disjoint buckets $G = \bigcup_{i=1}^{n} G^i$ of roughly equal size; i.e. $G^i = \{G_1^i, \ldots, G_{n_i}^i\}$ where $N_i = \sum_{j=1}^{n_i} \left|G_j^i\right| \approx M/n$. The partition size N_i can be chosen depending on the available main memory resources and the subsampling factor S. For each partition G^i a separate hash table (database) is constructed by the aforementioned method. Our partitioning scheme is illustrated in Fig. 1(a) and database construction in Fig. 1(b).

Single workstation

AFS-MetaCache constructs a separate database for each partition of reference sequences G^i and stores it as a database file on disk. We also allow to add sequences to previously constructed databases. This makes it easy to modify the set of reference genomes by either swapping out database partitions or including more sequences.

Spark

Apache Spark is a distributed memory computing engine [27]. It is able to process a large quantity of input data in parallel thanks to the combination of the Hadoop Distributed File System (HDFS) and Resilient Distributed Datasets (RDDs). These two features are used by MetaCacheSpark. Our algorithm consists of four phases that are illustrated in Fig. 2.

1. Reference genome sequences are loaded from HDFS and distributed proportionally among the Spark executors. In this way, each executor will contain a different subset of sequences to work with.
2. With these sequences loaded into memory, the Spark executors perform the described minhashing algorithm. Results are stored in a executor-local C++ hash table, similar to the one used by AFS-MetaCache.
3. We apply a map-reduce operation where the map operator receives the number of items belonging to the same key in each executor, and the reduction phase sums up the number of items calculating a global count. If the global item count per key exceeds a given threshold (by default 254), the corresponding items are deleted from all the executor-local hash tables.
4. Each hash table is written to a database file stored in HDFS.

At the end of the process, each executor will contain one, and only one, hash table. Note that a key can be present in several hash tables. However, items belonging to the same target ID (i.e., to the same reference sequence) will be present only in one hash table (this is important for the subsequent read assignment phase).

Furthermore, both versions have a pre-processing phase prior to database construction that builds a taxonomic tree of the considered reference genomes.

Individual read assignment

In order to assign reads to reference genome(s) minhashing is applied to any given read R in the same way as to a reference genome window using the hash function h_1. The produced sketch is used to query a loaded hash table using the hash function h_2. Each query returns a

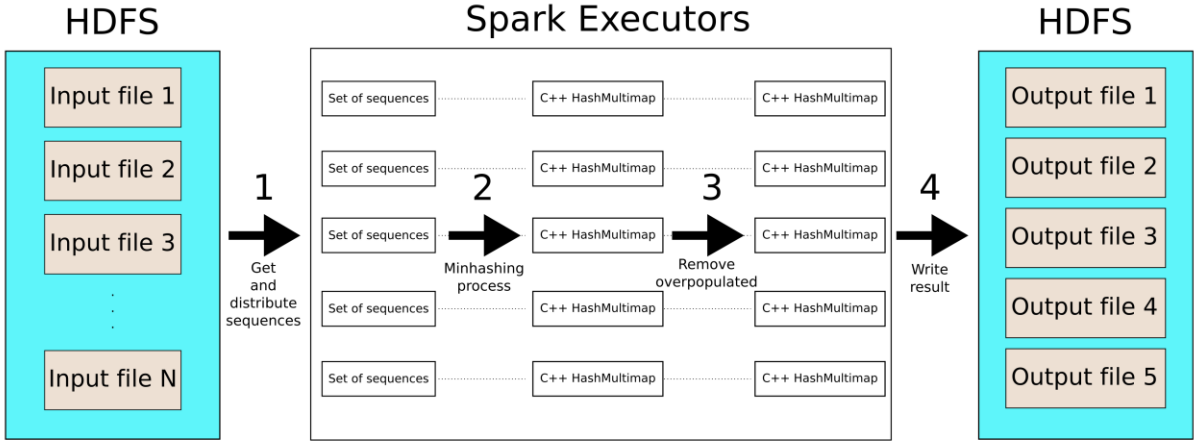

Fig. 2 Database construction algorithm used by MetaCacheSpark

(possibly empty) target list. The target lists are merged into a sparse two-dimensional data structure (called *window count statistic*) by accumulating identical (genome ID, window ID) pairs. High values in the window count statistic indicate a match of the read in the corresponding genome. The counts are sorted in descending order and the targets with the highest counts are considered in order to classify a read. This process is illustrated in Fig. 1(c).

However, a match of a paired-end (or even a single-end) read typically corresponds to a region in the genome that overlaps the borders of two or more windows in this genome. Thus, we accumulate the counters spanning a contiguous range of several neighboring windows to find the ranges with maximum hit counts. The considered read is assigned to the genome containing the best final count if it is significantly higher than the second best. If the count difference is small, the read is assigned to the lowest common ancestor (LCA) of multiple candidate genomes which are in a similar count range using the provided taxonomic tree.

Single workstation

AFS-MetaCache reads the database partitions from disk and queries them with the set of reads in succession. Subsequently, the individual results are merged to determine the final classification for each read. We further support multi-threading by processing chunks of reads independently in order to exploit multiple CPU cores.

Spark

Two inputs are needed: the database files created in the build phase and the input reads to be processed. The MetaCacheSpark algorithm consists of four steps (see Fig. 3):

1. Each hash table is loaded into the main memory of one executor. Furthermore, the taxonomy is loaded only in the Spark driver.
2. All executors read a block of N input reads to be processed from HDFS. Note that every executor needs to read all of them since the hash table is distributed. While reading the input sequences, each executor queries its local hash table to compute the (local) classification candidates with their corresponding hits. This process returns a set of key-value pairs, where the key is the ID of the read being processed, and the value is a list of possible candidates with their corresponding hit counts.
3. The next step is a reduction phase. Here, partial results from each executor are grouped using read IDs as keys. The driver then collects the N results and performs the assignment of reads to reference genomes (classification). This step uses the Spark function *reduceByKey()*, and it requires a *shuffle*.
4. Classification results from the previous step are written to the output file in HDFS. The algorithm goes back to Step 2 to process the next chunk of reads.

It is also important to note that:

- There is a guarantee that items belonging to the same reference sequence during the build phase are present in the same local hash table. Otherwise, calculating the hits in Step 2 would involve a distributed operation (such as *groupByKey()*) that would cause severe performance degradation.
- To gain speed, we further support multi-threading. Each thread processes a different set of input reads by means of a map-reduce job that corresponds to Steps 3 and 4.

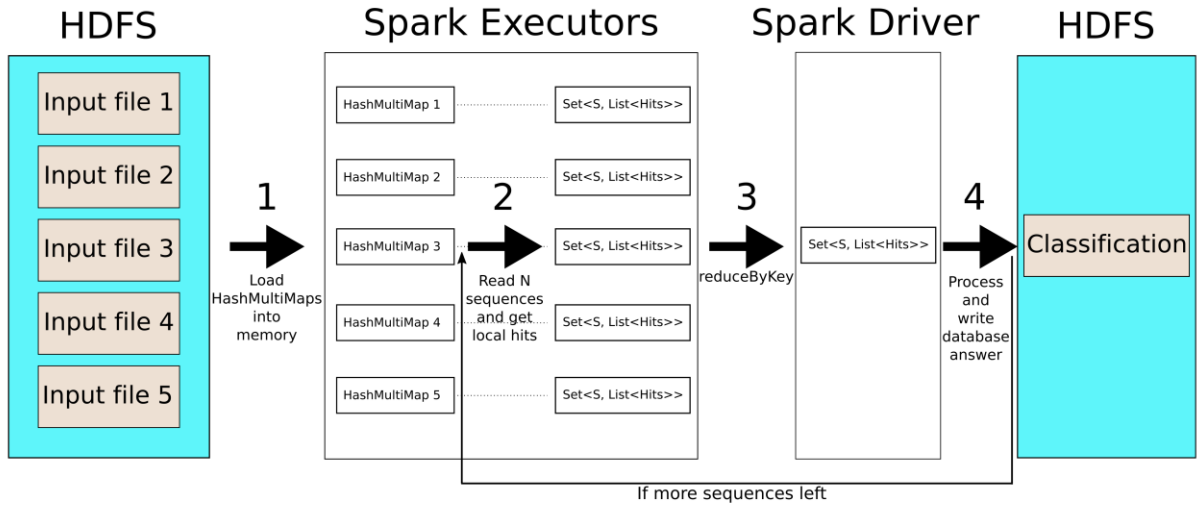

Fig. 3 Individual read classification algorithm used by the MetaCacheSpark

Table 1 Food-related reference genomes used for database construction

Item	Name	ID	Size on disk
1	Sus scrofa (pig)	GCF_000003025.6	2.4GB
2	Equus caballus (horse)	GCF_002863925.1	2.4GB
3	Meleagris gallopavo (turkey)	GCF_000146605.2	1.2GB
4	Mus musculus (house mouse)	GCF_000001635.26	2.7GB
5	Gallus gallus (chicken)	GCF_000002315.5	1.1GB
6	Ovis aries (sheep)	GCF_000298735.2	2.5GB
7	Rattus norvegicus (Norway rat)	GCF_000001895.5	2.8GB
8	Bos taurus (cattle)	GCF_002263795.1	2.6GB
9	Bubalus bubalis (water buffalo)	GCF_003121395.1	2.6GB
10	Cervus elaphus hippelaphus (red deer)	GCA_002197005.1	3.3GB
11	Capreolus capreolus (Western roe deer)	GCA_000751575.1	3.0GB
12	Struthio camelus australis (African ostrich)	GCA_000698965.1	1.2GB
13	Anas platyrhynchos (mallard)	GCF_003850225.1	1.1GB
14	Capra hircus (goat)	GCF_001704415.1	2.8GB
15	Oryctolagus cuniculus (rabbit)	GCF_000003625.3	2.6GB
16	Cavia aperea (Brazilian guinea pig)	GCA_000688575.1	2.6GB
17	Camelus ferus (Wild Bactrian camel)	GCF_000311805.1	1.9GB
18	Canis lupus familiaris (dog)	GCF_000002285.3	2.3GB
19	Felis catus (domestic cat)	GCF_000181335.3	2.4GB
20	Homo sapiens (human)	GCF_000001405.38	3.1GB
21	Equus asinus (ass)	GCA_001305755.1	2.3GB
22	Rangifer tarandus (reindeer)	GCA_004026565.1	2.9GB
23	Phasianus colchicus (Ring-necked pheasant)	GCA_004143745.1	987MB
24	Glycine max (soybean)	GCF_000004515.5	946MB
25	Zea mays (maize)	GCF_000005005.2	2.1GB
26	Triticum aestivum (bread wheat)	GCA_900519105.1	14.0GB
27	Secale cereale (rye)	GCA_900079665.1	1.8GB
28	Hordeum vulgare (barley)	GCA_004114815.1	3.8GB
29	Oryza sativa Japonica Group (Japanese rice)	GCF_001433935.1	362MB
30	Arachis hypogaea (peanut)	GCF_003086295.1	2.4GB
31	Saccharomyces cerevisiae S288C (baker's yeast)	GCA_000146045.2	12MB
Total			**74GB**

- The reduction generates a lot of traffic over the network and requires an expensive shuffle operation. In order to reduce the associated communication overhead, we have introduced an optional parameter (H) that is used to discard all candidates in Step 2 and Step 3 with less than H hits. However, if this parameter is used, results can be slightly different compared to the single workstation version.

Coverage filter

False positive read assignments can be caused by shared regions of DNA among multiple reference genomes [28]. We use coverage information to detect some of these cases as follows.

Before assigning reads to classification targets we can filter the list of candidate genomes identified during the read assignment phase by checking the coverage per genome as follows. We analyze which windows of a target genome are covered by reads from the dataset. If the percentage of covered windows of a genome is much lower compared to other genomes, it is likely to be a false positive and will be deleted from the list of possible target genomes. In fact we delete a quantile (e.g. 10%) of the target genomes with the lowest coverage. The

reads are then classified with respect to the remaining genomes.

Note that this strategy is only applicable if the number of reads is large enough to cover significant parts of the genomes. In our experience it proofed especially efficient in case of bacterial genomes which are orders of magnitudes smaller than animal or plant genomes.

Quantification

In addition to the per-read classification we are able to estimate the abundances of organisms contained in a dataset at a specific taxonomical rank. For each taxon which occurs in the dataset we count the number of reads assigned to it. We then build a taxonomic tree containing all found taxa.

Taxa on lower levels than the requested taxonomic rank are pruned and their read counts are added to their respective parents, while reads from taxa on higher levels are distributed among their children in proportion to the weights of the sub-trees rooted at each child. After the redistribution the estimated number of reads and abundance percentages are returned as outputs.

Results

Datasets

In order to measure performance and accuracy of our approach in comparison to other metagenomic tools, we have created databases of varying size containing different organisms. Food-related genomes (selection of main ingredients) used for database construction are listed in Table 1 while the considered bacteria, viruses, and archaea from NCBI RefSeq (Release 90) are summarized in Table 2. The created databases with their included reference genomes are described in Table 3.

We use ten short read datasets sequenced from calibrator sausage samples containing admixtures of a set of food relevant ingredients (chicken, turkey, pork, beef, horse, sheep) on an Illumina HiSeq machine (downloaded from ENA project ID PRJNA271645 (Kal_D and KAL_D) and PRJEB34001 (all other data)). Table 4 shows the read datasets together with the corresponding percentage of meat components used during preparation. The samples comprise meat proportions ranging from 0.5% to 80% and can be subdivided into two categories:

Table 2 Reference genomes from NCBI RefSeq (Release 90) used for database construction

Organism	Number of references	Size on disk
Bacteria	10838	41.0GB
Viral	7857	269MB
Archaea	269	656MB
Total	**18964**	**41.9GB**

Table 3 Data sets used for database construction

Name	Number of species	Size on disk
AFS10	Animal genomes from 1 to 10	22.3GB
AFS20	Animal genomes from 1 to 20	45.8GB
AFS20RS90	Animal genomes from 1 to 20 plus NCBI RefSeq (Release 90)	87.5GB
AFS31	Animal genomes from 1 to 31	76.8GB
AFS31RS90	Animal genomes from 1 to 31 plus NCBI RefSeq (Release 90)	118.5GB

Kal A-E consist only of mammalian meat, while KLyo A-D represent Lyoner-like sausages containing poultry in addition to mammals [29, 30]. The dataset KAL_D is identical to Kal_D but sequenced with higher coverage.

Quantification accuracy

Tables 5 and 6 show the quantification results returned by the tested tools (AFS-MetaCache (v.0.5.3), MetaCacheSpark, CLARK (v.1.2.6), Kraken2 (v.2.0.7-beta), and Kraken2 with subsequent abundance estimation by Bracken v.2.0.0 – all executed with default parameters) using AFS20 as reference database. Besides showing the quantification for each included meat component, we also show the (false positive) results for water buffalo (closely related to cattle) and goat (closely related to sheep). In addition, we provide the sum of all false positive (Σ FP) read classifications over all of the detected reference genomes that were not included in the sample. In addition, the sum of the deviations of the measured proportions

Table 4 Calibrator sausage datasets and their meat composition

Name	#Reads (paired-end)	Cattle	Sheep	Pig	Horse	Chicken	Turkey
KLyo_A	401K	14.0%	0.0%	80.0%	0.0%	0.5%	5.5%
KLyo_B	302K	36.0%	0.0%	58.0%	0.0%	2.0%	4.0%
KLyo_C	507K	58.0%	0.0%	36.0%	0.0%	4.0%	2.0%
KLyo_D	417K	80.0%	0.0%	14.0%	0.0%	5.5%	0.5%
Kal_A	830K	1.0%	9.0%	35.0%	55.0%	0.0%	0.0%
Kal_B	977K	9.0%	1.0%	55.0%	35.0%	0.0%	0.0%
Kal_C	404K	25.0%	25.0%	25.0%	25.0%	0.0%	0.0%
Kal_D	403K	35.0%	55.0%	9.0%	1.0%	0.0%	0.0%
Kal_E	289K	55.0%	35.0%	1.0%	9.0%	0.0%	0.0%
KAL_D	26,114K	35.0%	55.0%	9.0%	1.0%	0.0%	0.0%

Table 5 Quantification results for the Klyo samples using the reference dataset AFS20 and the average result for AFS31RS90

Dataset	Classifier	Cattle	Pig	W.Buf.	Goat	Chicken	Turkey	Σ FP	Σ Dev
KLyo_A	Expected	14.0%	80.0%	0.00%	0.00%	0.50%	5.50%		
	AFS-MC	16.6%	71.5%	0.04%	0.02%	0.60%	4.64%	**0.28**%	**12.39**%
	MCSpark	16.9%	71.2%	0.04%	0.02%	0.60%	4.64%	0.32%	12.99%
	CLARK	16.4%	70.4%	0.20%	0.09%	0.62%	4.61%	0.51%	13.55%
	Kraken2	15.9%	70.0%	0.27%	0.11%	0.65%	4.59%	0.87%	13.82%
	K2+Brack	17.6%	70.3%	0.30%	0.14%	0.66%	4.63%	0.97%	15.33%
KLyo_B	Expected	36.0%	58.0%	0.00%	0.00%	2.00%	4.00%		
	AFS-MC	37.6%	51.0%	0.12%	0.04%	2.05%	2.99%	**0.50**%	10.16%
	MCSpark	37.9%	50.5%	0.12%	0.04%	2.06%	3.02%	0.60%	11.11%
	CLARK	35.9%	50.4%	0.47%	0.19%	2.10%	3.01%	1.03%	**9.84**%
	Kraken2	34.5%	49.9%	0.68%	0.24%	2.12%	2.99%	1.57%	12.11%
	K2+Brack	39.1%	50.2%	0.32%	0.78%	2.15%	3.02%	1.84%	13.93%
KLyo_C	Expected	58.0%	36.0%	0.00%	0.00%	4.00%	2.00%		
	AFS-MC	57.7%	27.1%	0.16%	0.06%	3.56%	1.16%	**0.95**%	**11.47**%
	MCSpark	57.7%	26.9%	0.16%	0.06%	3.63%	1.18%	0.95%	11.48%
	CLARK	54.1%	25.9%	0.69%	0.29%	3.58%	1.16%	1.88%	17.11%
	Kraken2	52.2%	25.7%	0.95%	0.36%	3.57%	1.17%	2.58%	19.94%
	K2+Brack	58.6%	25.8%	1.07%	0.46%	3.60%	1.18%	2.89%	14.90%
KLyo_D	Expected	80.0%	14.0%	0.00%	0.00%	5.50%	0.50%		
	AFS-MC	74.7%	10.9%	0.23%	0.08%	4.66%	0.33%	**0.93**%	10.27%
	MCSpark	74.7%	10.8%	0.23%	0.08%	4.69%	0.33%	1.09%	10.58%
	CLARK	70.8%	10.8%	0.94%	0.39%	4.73%	0.35%	1.94%	15.27%
	Kraken2	68.0%	10.7%	1.26%	0.48%	4.70%	0.36%	2.42%	18.62%
	K2+Brack	77.6%	10.8%	1.45%	0.62%	4.76%	0.36%	2.87%	**9.35**%
Average	AFS-MC			**0.14**%	**0.05**%			**0.67**%	**11.07**%
	MCSpark			**0.14**%	**0.05**%			0.74%	11.54%
	CLARK			0.58%	0.24%			1.34%	13.94%
	Kraken2			0.79%	0.30%			1.86%	16.12%
	K2+Brack			0.71%	0.50%			2.14%	13.38%
AFS31RS90 Average	AFS-MC							**0.58**%	**13.97**%
	MCSpark							0.59%	14.08%

AFS-MC: AFS-MetaCache, MC-Spark: MetaCacheSpark, K2+Brack: Kraken2 with subsequent Bracken, W.Buf: Water Buffalo, Σ FP: Sum of all false positive read classifications, Σ Dev: Sum of absolute deviations to the given meat composition (best results for each dataset in bold)

to the real sausage composition (Σ Dev) as well as the averages over all tested datasets are shown.

In terms of sensitivity, all methods are able to detect the included meat components. In addition, several tools detect false positive signals; e.g., Kraken2+Bracken detects over 1% of water buffalo in KLyo_C and KLyo_D and over 3% of goat in Kal_C, Kal_D, and Kal_E. False positive quantities in these cases correlate with the amount of beef and the amount of sheep present in the respective sample. Overall, AFS-MetaCache achieves the lowest FP-rates for each tested dataset with an average FP-sum per sample of only 0.67% for the Klyo samples and 1.12% for the Kal samples. This is much lower compared to CLARK (1.34% for Klyo, 3.59% for Kal), Kraken2 (1.86% for Klyo, 3.87% for Kal), and Kraken2+Bracken (2.14% for Klyo, 4.41% for Kal). The relative differences become even more significant when looking at some of the individual FP signals. In the Klyo samples (Table 5) AFS-MetaCache only detects negligible amounts of goat (0.05% on average) and water buffalo (0.14%), while the amounts detected by CLARK, Kraken2, and Kraken2+Bracken are higher by factors of 4.2 and 4.8, 5.6 and 6.0, and 5.1 and 10.0,

Table 6 Quantification results for the Kal samples using the reference dataset AFS20 and the average result for AFS31RS90

Dataset	Classifier	Cattle	Sheep	Pig	Horse	W.Buf.	Goat	Σ FP	Σ Dev
	Expected	1.00%	9.0%	35.0%	55.0%	0.00%	0.00%		
	AFS-MC	1.25%	11.0%	30.5%	54.1%	0.01%	0.29%	**0.42**%	8.13%
Kal_A	MCSpark	1.27%	11.1%	30.3%	54.1%	0.01%	0.29%	0.45%	8.42%
	CLARK	1.29%	9.1%	31.1%	54.0%	0.09%	0.89%	1.15%	**6.43**%
	Kraken2	1.23%	8.7%	30.9%	53.9%	0.08%	0.96%	1.31%	6.99%
	K2+Brack	1.43%	10.3%	31.0%	54.0%	0.10%	1.12%	1.53%	8.24%
	Expected	9.0%	1.00%	55.0%	35.0%	0.00%	0.00%		
	AFS-MC	10.5%	1.42%	49.3%	35.6%	0.03%	0.06%	**0.27**%	8.43%
Kal_B	MCSpark	10.6%	1.42%	49.1%	35.7%	0.03%	0.06%	0.30%	8.92%
	CLARK	10.3%	1.26%	50.0%	35.8%	0.17%	0.18%	0.56%	**7.85**%
	Kraken2	10.0%	1.21%	49.6%	35.7%	0.20%	0.20%	1.03%	8.40%
	K2+Brack	11.0%	1.40%	35.8%	49.7%	0.22%	0.23%	1.09%	9.60%
	Expected	25.0%	25.0%	25.0%	25.0%	0.00%	0.00%		
	AFS-MC	23.3%	29.6%	19.2%	23.0%	0.06%	0.73%	**1.08**%	15.28%
Kal_C	MCSpark	23.5%	29.6%	19.0%	22.9%	0.06%	0.73%	1.18%	15.32%
	CLARK	23.4%	25.6%	19.4%	23.2%	0.45%	2.56%	3.38%	**12.98**%
	Kraken2	22.7%	24.7%	19.4%	23.1%	0.49%	2.69%	3.48%	13.65%
	K2+Brack	24.8%	27.8%	19.4%	23.2%	0.54%	3.02%	3.89%	14.35%
	Expected	35.0%	55.0%	9.00%	1.00%	0.00%	0.00%		
	AFS-MC	32.9%	51.5%	7.14%	1.14%	0.09%	1.50%	**2.07**%	**9.62**%
Kal_D	MCSpark	33.2%	51.2%	7.03%	1.13%	0.09%	1.49%	2.23%	9.91%
	CLARK	32.8%	43.1%	7.31%	1.16%	0.72%	4.40%	5.69%	21.61%
	Kraken2	31.6%	41.3%	7.26%	1.16%	0.79%	4.62%	5.77%	24.75%
	K2+Brack	35.8%	48.4%	7.28%	1.16%	0.89%	5.40%	6.70%	15.96%
	Expected	55.0%	35.0%	1.00%	9.00%	0.00%	0.00%		
	AFS-MC	50.4%	33.7%	0.99%	7.80%	0.12%	0.96%	**1.52**%	**8.55**%
Kal_E	MCSpark	50.7%	33.4%	0.97%	7.73%	0.12%	0.95%	1.66%	8.82%
	CLARK	50.7%	28.7%	1.02%	7.81%	0.84%	3.07%	4.43%	16.26%
	Kraken2	49.2%	27.6%	1.00%	7.80%	0.99%	3.28%	4.58%	18.96%
	K2+Brack	54.1%	31.4%	1.00%	7.81%	1.10%	3.71%	5.15%	10.86%
	Expected	35.0%	55.0%	9.00%	1.00%	0.00%	0.00%		
	AFS-MC	30.3%	49.6%	7.27%	1.16%	0.08%	1.25%	1.38%	**13.36**%
KAL_D	MCSpark	30.4%	49.5%	7.25%	1.16%	0.08%	1.26%	**1.36**%	**13.36**%
	CLARK	30.8%	43.3%	7.51%	1.20%	0.86%	4.57%	6.30%	23.85%
	Kraken2	29.6%	41.3%	7.47%	1.19%	0.95%	4.98%	7.03%	27.86%
	K2+Brack	33.5%	48.7%	7.58%	1.19%	1.08%	5.84%	8.07%	17.44%
	AFS-MC					**0.07**%	**0.80**%	**1.12**%	**10.56**%
	MCSpark					**0.07**%	**0.80**%	1.20%	10.79%
Average	CLARK					0.51%	2.61%	3.59%	14.83%
	Kraken2					0.58%	2.79%	3.87%	16.77%
	K2+Brack					0.66%	3.22%	4.41%	12.74%
AFS31RS90 Average	AFS-MC							**1.84**%	**13.38**%
	MCSpark							**1.84**%	13.63%

AFS-MC: AFS-MetaCache, MC-Spark: MetaCacheSpark, K2+Brack: Kraken2 with subsequent Bracken, W.Buf: Water Buffalo, Σ FP: Sum of all false positive read classifications, Σ Dev: Sum of absolute deviations to the given meat composition (best results for each dataset in bold)

Table 7 Runtimes and peak memory consumption for non-partitioned database construction (build) and querying for different data sets on a workstation with 512 GB RAM

Data set		AFS-MetaCache	CLARK	Kraken2	Kraken2+Bracken
AFS20	Build time	**1h 11m**	15h 37m	1h 27m	5h 32m
	Build memory	**64 GB**	428 GB	69 GB	147 GB
	Query time	136 s	93 s	**37 s**	111 s
	Query speed	11.5 MR/m	16.9 MR/m	**43.2 MR/m**	14.2 MR/m
	Query memory	**50 GB**	152 GB	54 GB	54 GB
AFS31	Build time	**1h 47m**	-	3h 19min	11h 41min
	Build memory	**91 GB**	-	107 GB	296 GB
	Query time	175 s	-	**44 s**	58 s
	Query speed	8.9 MR/m	-	**35.9 MR/m**	27.0 MR/m
	Query memory	78 GB	-	**72 GB**	**72 GB**
AFS20RS90	Build time	**1h 42m**	-	2h 58m	8h 53m
	Build memory	110 GB	-	**94 GB**	168 GB
	Query time	180 s	-	**43 s**	117 s
	Query speed	8.7 MR/m	-	**37.0 MR/m**	13.5 MR/m
	Query memory	94 GB	-	**79 GB**	**79 GB**
AFS31RS90	Build time	**3h 10m**	-	5h 55min	17h 44min
	Build memory	135 GB	-	**134 GB**	329 GB
	Query time	217 s	-	**49 s**	61 s
	Query speed	7.2 MR/m	-	**32.1 MR/m**	25.7 MR/m
	Query memory	117 GB	-	**97 GB**	**97 GB**

Query speeds are measured for the KAL_D dataset in terms of million reads per minute (MR/m). For the cases with "-" the corresponding program exceeds the main memory capacity of 512 GB. Fastest runtimes and lowest memory consumption for each dataset are indicated in bold

respectively. Similar results can be observed for the Kal samples (Table 6): AFS-MetaCache only detects 0.07% of water buffalo meat on average and 0.80% of goat meat on average, while the amounts detected by CLARK, Kraken2, and Kraken2+Bracken are higher by factors of 7.3 and 3.3, 8.3 and 3.5, and 9.4 and 4.0, respectively.

In terms of deviation from the expected foodstuff ingredients, AFS-MetaCache shows the lowest average of the sums of absolute differences for both Klyo (11.07%) samples and Kal samples (10.56%). Kraken2+Bracken (13.38% and 12.74%) has smaller deviations on average than Kraken2 alone (16.12% and 16.77%), showing that quantification after read assignment is beneficial.

As can be seen in Tables 5 and 6 there are small differences between the results of AFS-MetaCache and MetaCacheSpark. They are caused by the constraint list of target genomes with highest scores (tophits) of MetaCacheSpark and by the different ordering of targets with the same score. The differences could be reduced by increasing the tophits list size, but we decided for a smaller list in favor of faster querying speeds.

Table 8 Partitioned build time and query speed for AFS31RS90 database

Tool	Build time	Max. Memory	Query Speed	Max. Memory
AFS-MetaCache (1 part.)	3h 10min	135 GB	7.2 MR/m	117 GB
AFS-MetaCache (2 part.)	3h 04min	82 GB	3.1 MR/m	70 GB
AFS-MetaCache (4 part.)	3h 45min	52 GB	2.5 MR/m	39 GB
MetaCacheSpark (8Ex-32Th)	2h 57min	175 GB	4.3 MR/m	76 GB
MetaCacheSpark (16Ex-16Th)	1h 57min	100 GB	3.4 MR/m	48 GB
MetaCacheSpark (32Ex-8Th)	1h 25min	69 GB	2.2 MR/m	37 GB
MetaCacheSpark (64Ex-4Th)	1h 03min	45 GB	1.4 MR/m	29 GB

Query speed measured for dataset KAL_D in million reads per minute (MR/m). For MetaCacheSpark, the number of executors and threads per executor are indicated

When scanning the calibrator sausage read datasets with AFS-MetaCache using the bigger AFS31 and AFS31RS90 databases, we can make the following observations: (1) More *k*-mers are removed from the hash table due to overflowing target lists. Therefore, the number of classified reads is reduced and total deviation increases slightly. (2) Additional false positive targets are introduced, but the total number of false positives is reduced for the Klyo datasets (excluding bacteria).

A benefit of screening for microbiota and eukaryotic foodstuff species at the same time is a lower false positive rate. Usually reads of a dataset are queried against either one or the other and only the remaining unclassified reads are investigated further. This can lead to false assumptions about the data. In our experiments some reads are falsely classified as Triticum aestivum (bread wheat) when using the AFS31 database. With the AFS31RS90 database, however, those reads are identified as bacterial or unspecific (classified as the lowest common ancestor of bread wheat and bacteria).

Runtime and memory consumption for non-Partitioned databases

Runtime and memory consumption where the whole database can fit into the available main memory are measured on a system with a dual Xeon E5-2630v4 (2.2 GHz, 2 × 10 cores) CPU with 512 GB of DDR4 RAM. We have compared the speed and the peak memory consumption during database construction and classification of the default versions of AFS-MetaCache (v.0.5.3), CLARK (v.1.2.6), Kraken2 (v2.0.7-beta), and Kraken2 with subsequent abundance estimation by Bracken v.2.0.0 (Kraken2+Bracken) using 40 threads. Table 7 shows the results for the reference genome datasets listed in Table 3 and the KAL_D read dataset (26 million paired-end reads of length 101 bp) for classification. Note, that the time to load the databases is excluded when measuring query speed for all programs to make the results independent of dataset size.

AFS-MetaCache is fastest for database construction for all tested data sets. Furthermore, it requires least memory for constructing the database for AFS20 and AFS31, but requires slightly more memory than Kraken2 for AFS20RS90 and AFS31RS90.

Kraken2 is fastest in terms of query (classification) speed. If Kraken2 is executed with subsequent quantification by Bracken, corresponding runtimes increase. Even though query speeds of MetaCache-AFS are slowest, corresponding execution times are still competitive (only around three minutes for the largest data set (KAL_D)).

For common data set sizes in food control applications runtimes for database construction (a few hours) are typically much higher than for the classification stage (a few minutes). Since the amount of relevant reference genomes is increasing rapidly corresponding databases have to be constructed or extended frequently. Thus, fast built times are of high importance. Besides having the fastest database construction time, AFS-MetaCache is also the only tool that supports the functionality of extending an existing database.

Table 9 Runtimes and peak memory consumption for database construction (build) and querying for AFS10

Data set		AFS-MetaCache	AFS-previous
AFS10	Build time	**47m**	7h 0m
	Build memory	35 GB	**5GB**
	Query speed	**17.1 MR/m**	0.04 MR/m
	Query memory	30 GB	**6GB**

Query speeds are measured for the KAL_D dataset in terms of million reads per minute (MR/m)

Runtime and memory consumption for partitioned databases

In this subsection we evaluate the ability of AFS-MetaCache and MetaCacheSpark to reduce the consumed main memory by partitioning the database into smaller chunks. AFS-MetaCache is again evaluated on a workstation with a dual Xeon E5-2630v4 CPU and 512 GB of DDR4 RAM. MetaCacheSpark has been tested on a big data cluster composed of 12 Dell EMC PowerEdge R730 servers, each one with a dual Xeon E5-2630v4 (2.2GHz 10 cores) CPU with 384 GB RAM and 32 TB HDDs running Java version Openjdk 1.8.0_201, gcc 7.3.1, Spark 2.2.0, and Hadoop 2.7.3.

Table 8 shows the speed and memory consumption of AFS-MetaCache and MetaCacheSpark for partitioned database construction and querying using the AFS31RS90 reference genome dataset and the KAL_D dataset. Using four partitions, AFS-MetaCache can reduce the main memory consumption from 135 GB to only 52 GB while the construction time only slightly increases from 3h 10m to 3h 45m. In addition, memory consumption for classification is reduced from 117 GB to 39 GB. However, the corresponding query speed decreases from 7.2 MR/m to 2.5 MR/m since the partitions have to be queried by all

Table 10 Average quantification results for the Klyo and Kal samples using the reference dataset AFS10

Dataset	Classifier	Σ FP	Σ Dev
KLyo Average	AFS-MC	**0.37**%	**10.71**%
	AFS-prev	**0.37**%	10.80%
Kal & KAL_D Average	AFS-MC	**0.19**%	8.43%
	AFS-prev	0.33%	**6.65**%

AFS-MC: AFS-MetaCache, AFS-prev: previous AFS pipeline, Σ FP: Sum of all false positive read classifications, Σ Dev: Sum of absolute deviations to the given meat composition

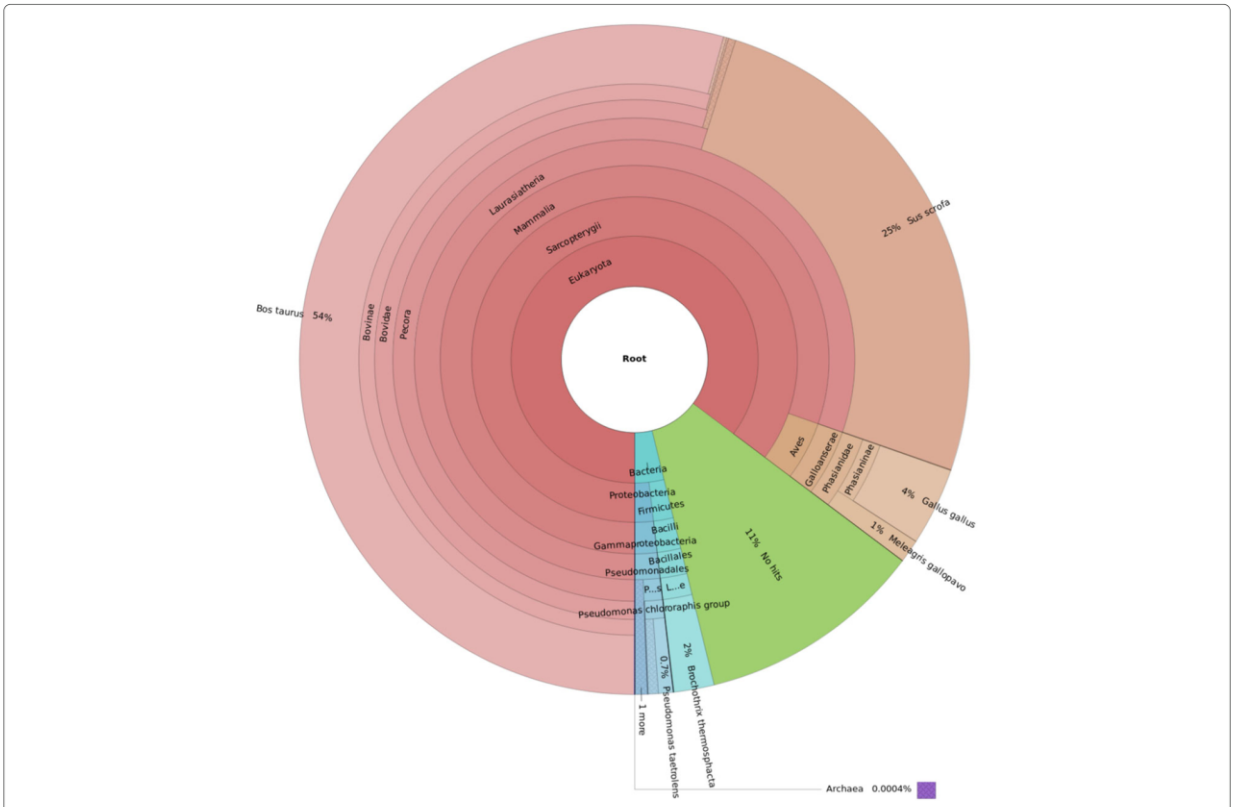

Fig. 4 Visualization of the AFS-MetaCache results using Krona [31] for the dataset KLyo_C using the AFS31RS90 reference data set

reads in succession and the individual results need to be merged.

The results show that memory requirements per node and build time for MetaCacheSpark both decrease when increasing the number of executors. As the number of executors increases, the benefits of using the Spark version are revealed. For 64 executors the AFS31RS90 database can be built in around one hour using 45 GB of memory per node. This is 3×, 3.6×, 5.6×, 16.9× faster than AFS-MetaCache, 4-partitioned AFS-MetaCache, Kraken2 and Kraken2+Bracken, respectively. Important reductions in the memory consumed per node can also be observed.

MetaCacheSpark consumes less memory in the classification phase than in the build phase. Some additional memory is required to store query hits. However, this memory can be re-used with each batch of sequences being classified. As a trade-off to fast build time and low memory consumption per node, the query speeds of MetaCacheSpark are lower compared to non-partitioned AFS-MetaCache. This can be explained by the necessity to perform a costly shuffle operation for the reduce-by-key function. Its cost increases with the number of executors as can be seen in Table 8: query speed reduces from 4.2 MR/m with 8 executors to 1.4 MR/m with 64. Nevertheless, runtimes are still acceptable in application scenarios where relevant read datasets are small compared to the utilized databases.

Table 11 Detected bacteria in dataset KLyo_C using reference dataset AFS31RS90

Genus	AFS-MetaCache	Kraken2	Kraken2+Bracken
Brochothrix	1.94%	1.94%	1.98%
Pseudomonas	1.23%	1.73%	1.92%
Psychrobacter	0.59%	1.43%	1.45%

Genera with less 500 than hits (< 0.1% of the dataset) are omitted

Comparison to previous aFS pipeline

To compare AFS-MetaCache to our previous alignment-based AFS pipeline the same dual-socket workstation as before is used. Runtimes and memory consumption of both approaches are shown in Table 9. For the small genome dataset AFS10 the previous AFS pipeline already takes several hours to construct the index. Querying of the KAL_D dataset takes even more than 10 hours. For bigger numbers of reference genomes this approach becomes unfeasible because the runtime scales linearly

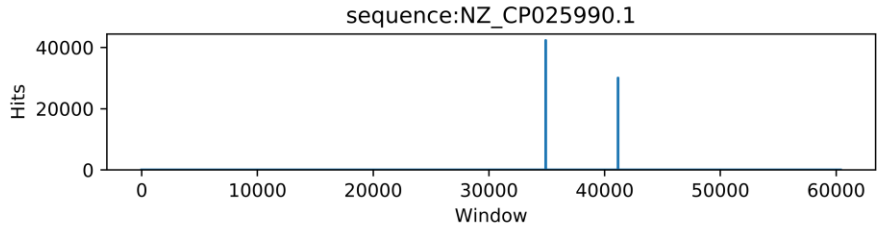

Fig. 5 Genome coverage of Actinoalloteichus for the dataset KLyo_C. The sparse coverage is an indicator for false positives

with the number reference genomes. On the other hand, AFS-MetaCache takes less than an hour for database construction of AFS10 while the query speed improves by more than two orders of magnitude. As shown before even larger databases like AFS31 can be built by AFS-MetaCache in just a few hours and query speed drops by less than a factor of two.

The average quantification results for the Klyo and Kal samples produced by AFS-MetaCache and the previous AFS pipeline are shown in Table 10. The *k*-mer based AFS-MetaCache is able to match quantification accuracy of the previous alignment-based pipeline for the KLyo datasets. The average deviation to the meat components is even lower for AFS-MetaCache. For the Kal datasets AFS-MetaCache reduces the false positive rate while the average deviation increases slightly. However, it is still possible to identify the correct components with the benefit of less false positives.

Detection of microbiota

A major strength of next generation sequencing when applied to foodstuffs, is its theoretically infinite range of species that can be detected. We therefore analyzed the microbiota detected by AFS-MetaCache and MetaCacheSpark in more detail. A visualization of the AFS-MetaCache results using Krona [31] for the dataset KLyo_C using the AFS31RS90 reference data set is shown in Fig. 4. The results of Kraken2 and Bracken agree on the most prominent bacteria as shown in Table 11. The detected bacterial genera Brochothrix, Pseudomonas, and Psychrobacter are well known representatives in foodstuffs. In some sausages a very high amount of the species Brochothrix thermosphacta and even the corresponding Brochothrix phage BL3 could be found, possibly indicating meat spoilage. Furthermore, in several cases a significant amount of Actinoalloteichus was initially detected which has no known relation to foodstuff. However, after application of the coverage filter these matches could be detected as false positives and were removed.

Figures 5 and 6 show the corresponding genome coverage diagrams for Actinoalloteichus and Brochothrix thermosphacta for the KLyo_C read dataset. The highly uneven genome coverage of Actinoalloteichus is taken as an indicator by AFS-MetaCache for a false-positive species identification. The Brochothrix genome is evenly covered by reads and is thus classified as a true positive.

Discussion

The determination and quantification of food ingredients is an important issue in official food control [1]. Furthermore, microbiological contamination or the presence of non-declared allergenic food components establishes the need for a broad-scale screening method that allows for precise determination and quantification of ingredients ideally spanning all kingdoms of life including plants, animals, fungi, and bacteria. DNA-based methods like quantitative real-time PCR are established technologies for analyzing foodstuff. However, they have the drawback of being limited to a set of target species within a single assay that need to be defined beforehand. The usage of next-generation sequencing of total genomic DNA from biological samples followed by bioinformatics analyses based on comparisons to available reference genomes can overcome this limitation. Our previous alignment-based AFS-pipeline was found suitable to screen for species in processed food samples [5, 6]. However, the utilized

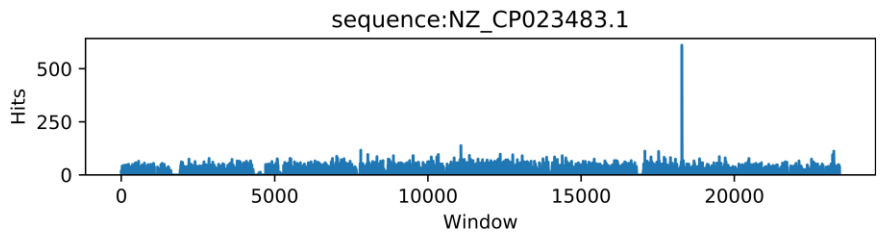

Fig. 6 Genome coverage of Brochothrix thermosphacta for the dataset KLyo_C. The even coverage is an indicator for true positives

algorithms put limitations on the number species to be screened and on the computational throughput.

Here, we have presented AFS-MetaCache and MetaCacheSpark as new computational methods for the efficient detection and quantification of species composition in food samples from sequencing reads. Being based on an alignment-free exact *k*-mer matching approach, we gain significant speed compared to our previous alignment-based AFS method at the expense of a higher memory consumption for constructing and querying reference genome databases. We apply an intelligent subsampling technique based on minhashing within local windows to reduce the database size. Further reductions of peak memory consumption can be achieved by the introduced partitioning schemes either for single workstations (AFS-MetaCache) or for big data clusters (MetaCacheSpark) at the expense of query speed. Applications of our previous alignment-based AFS pipeline have been limited to around ten complex genomes. With AFS-MetaCache we are able to significantly extend this limit, which is of high importance since the amount of available reference genomes continues to grow rapidly [32, 33]. Thus, our results are particularly encouraging since AFS-MetaCache and MetaCacheSpark are fastest in terms of database construction times. Corresponding peak memory consumption is competitive and can be even further reduced by the partitioned version of AFS-MetaCache on a single workstation or by using MetaCacheSpark on a big data cluster.

While AFS-Metacache can achieve higher query speed than MetaCacheSpark, it takes some manual setup for the partitioned version. MetaCacheSpark on the other hand allows for faster database creation and can easily be deployed on existing Spark infrastructure, while being faster than the partitioned version of AFS-Metacache. Spark, while being fault tolerant, also enables to use a cluster of lower powered computers than we used for our benchmarks.

Within this study we have applied our approach on a broad set of reference samples, containing admixtures of a set of food relevant ingredients (chicken, turkey, pork, beef, horse, sheep). The results demonstrate that our approach is able to reliably detect the components even at the 0.5% level. The comparison to the established metagenomics tools Kraken2, CLARK, and Kraken2+Bracken shows that AFS-MetaCache and MetaCacheSpark are superior in terms of false positive (FP) rates. In particular for pairs of closely related genomes AFS-MetaCache can achieve almost an order-of-magnitude lower FP-rates. These results demonstrate that our classification approach based on counting *k*-mer matches within small windows is effective compared to simply counting *k*-mer matches over an entire genome (as used by CLARK and Kraken) and to an alignment-based approach (as used by the our previous AFS pipeline). Our results also show that AFS-MetaCache achieves the lowest sum of absolute deviations to the included food ingredients. As different types of tissue can contain different concentrations of DNA (matrix effect), deviations could possibly be further reduced by a subsequent normalization procedure that takes tissue ratios into account.

Applications of AFS-MetaCache and MetaCacheSpark are not limited to the study of foodstuff but can be used to analyze high throughput sequencing datasets of metagenomic DNA from other complex biological samples as well, including diverse environmental materials, in-vitro cell cultures, and biopharmaca.

Conclusion

We have presented a fast screening and quantification method together with two corresponding publicly available implementations (AFS-MetaCache and MetaCacheSpark) for whole genome shotgun sequencing-based biosurveillance applications such as food testing. By relying on a big data approach our approach can scale efficiently towards large-scale collections of complex eukaryotic and bacterial reference genomes making both tools suitable for broad-scale metagenomic screening applications.

Availability and requirements

Project name: AFS-MetaCache
Project home page: https://muellan.github.io/metacache/afs.html
Operating system(s): Linux
Programming language: C++
Other requirements: gcc
License: GPL-3
Any restrictions to use by non-academics: according to license

Project name: MetaCacheSpark
Project home page: https://github.com/jmabuin/MetaCacheSpark
Operating system(s): Linux
Programming language: Java and C++
Other requirements: Openjdk, gcc, Spark, Hadoop
License: GPL-3
Any restrictions to use by non-academics: according to license

Abbreviations

AFS: All-Food-Sequencing; ddPCR: Droplet digital real-time polymerase chain reaction; HDFS: Hadoop distributed file system; LCA: Lowest common ancestor; LSH: Locality sensitive hashing; qPCR: Quantitative real-time polymerase chain reaction; RDD: Resilient distributed datasets; Σ FP: Sum of all false positive read classifications; Σ Dev: Sum of absolute deviations

Acknowledgements

Not applicable.

Authors' contributions
RK, JA, and AM implemented and tested the software. RK, JA, and SH performed the experiments. BS, TH, and AH wrote the draft of the manuscript. BS proposed the project. BS, TH, AH, JP, and TP supervised the project. RK, JA, AM, SH, JP, TP, TH, and BS edited the manuscript and analyzed the results. The author(s) read and approved the final manuscript.

Funding
This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG), Project HySim, the MINECO under award RTI2018-093336-B-C21, Xunta de Galicia under awards ED481B 2018/013 and ED431C 2018/19, the European Regional Development Fund, and by the Federal Office for Agriculture and Food. The funders had no role in study design, data collection, interpretation of data, decision to publish, or preparation of the manuscript.

Availability of data and materials
AFS-MetaCache and MetaCacheSpark are available at https://muellan.github.io/metacache/afs.html and https://github.com/jmabuin/MetaCacheSpark. The utilized Illumina sequence read datasets can be downloaded at ENA projects PRJNA271645 (Kal_D and KAL_D) and PRJEB34001 (all others). PRJEB34001 will be made publicly available upon acceptance of the paper.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Computer Science, Johannes Gutenberg University, 55099 Mainz, Germany. [2]IPCA, Polytechnic Institute of Cávado and Ave, 4750-810 Barcelos, Portugal. [3]CiTIUS, Universidade de Santiago de Compostela, 15782 Santiago de Compostela, Spain. [4]Molecular Genetics and Genome Analysis, Institute of Organismal and Molecular Evolution, Johannes Gutenberg University, 55099 Mainz, Germany.

Received: 19 August 2019 Accepted: 24 February 2020
Published online: 12 March 2020

References

1. Esteki M, Regueiro J, Simal-Gándara J. Tackling fraudsters with global strategies to expose fraud in the food chain. Compr Rev Food Sci Food Saf. 2019;18(2):425–40.
2. Köppel R, Ruf J, Rentsch J. Multiplex real-time pcr for the detection and quantification of dna from beef, pork, horse and sheep. Eur Food Res Technol. 2011;232(1):151–5.
3. Köppel R, Ganeshan A, van Velsen F, Weber S, Schmid J, Graf C, Hochegger R. Digital duplex versus real-time pcr for the determination of meat proportions from sausages containing pork and beef. Eur Food Res Technol. 2019;245(1):151–7.
4. Tillmar AO, Dell'Amico B, Welander J, Holmlund G. A universal method for species identification of mammals utilizing next generation sequencing for the analysis of dna mixtures. PloS ONE. 2013;8(12):83761.
5. Ripp F, Krombholz CF, Liu Y, et al. All-food-seq (afs): a quantifiable screen for species in biological samples by deep dna sequencing. BMC Genomics. 2014;15:639.
6. Liu Y, Ripp F, Koeppel R, et al. Afs: identification and quantification of species composition by metagenomic sequencing. Bioinformatics. 2017822. https://doi.org/10.1093/bioinformatics/btw822.
7. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
8. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.
9. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2. 2013.
10. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357.
11. Liu Y, Schmidt B, Maskell DL. Cushaw: a cuda compatible short read aligner to large genomes based on the burrows–wheeler transform. Bioinformatics. 2012;28(14):1830–7.
12. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.
13. Lindgreen S, Adair KL, Gardner P. An evaluation of the accuracy and speed of metagenome analysis tools. Sci Rep. 2016;6(19233):. https://doi.org/10.1038/srep19233.
14. Seppey M, Manni M, Zdobnov EM. Lemmi: A live evaluation of computational methods for metagenome investigation. bioRxiv. 2019. https://doi.org/10.1101/507731. https://www.biorxiv.org/content/early/2019/04/16/507731.full.pdf.
15. Lu J, Breitwieser FP, Thielen P, Salzberg SL. Bracken: estimating species abundance in metagenomics data. PeerJ Comput Sci. 2017;3:104.
16. Ounit R, Wanamaker S, Close TJ, et al. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. BMC Genomics. 2015;16(1):1–13. https://doi.org/10.1186/s12864-015-1419-2.
17. Müller A, Hundt C, Hildebrandt A, Hankeln T, Schmidt B. Metacache: context-aware classification of metagenomic reads using minhashing. Bioinformatics. 2017;33(23):3740–8.
18. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12(10):902–3. https://doi.org/10.1038/nmeth.3589.
19. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013;10(12):1196.
20. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, et al. Qiime allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7(5):335–6.
21. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with kaiju. Nat Commun. 2016;7:11257.
22. Broder AZ. Identifying and Filtering Near-Duplicate Documents. In: Proc. 11th Annual Symposium on Combinatorial Pattern Matching, COM '00; 2000. p. 1–10. http://dl.acm.org/citation.cfm?id=647819.736184.
23. Berlin K, Koren S, Chin C-S, et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotech. 2015;33:623–30. https://doi.org/10.1038/nbt.3238.
24. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Phillippy AM. Mash: fast genome and metagenome distance estimation using minhash. Genome Biol. 2016;17(1):132. https://doi.org/10.1186/s13059-016-0997-x.
25. Popic V, Batzoglou S. A hybrid cloud read aligner based on minhash and kmer voting that preserves privacy. Nat Commun. 2017;8:15311.
26. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM. Mash screen: high-throughput sequence containment estimation for genome discovery. Genome Biol. 2019;20(1):232. https://doi.org/10.1186/s13059-019-1841-x.
27. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ, et al. Apache spark: a unified engine for big data processing. Commun ACM. 2016;59(11):56–65.
28. Dadi TH, Renard BY, Wieler LH, Semmler T, Reinert K. Slimm: species level identification of microorganisms from metagenomes. PeerJ. 2017;5:3138.
29. Köppel R, Ruf J, Rentsch J. Multiplex real-time pcr for the detection and quantification of dna from beef, pork, horse and sheep. Eur Food Res Technol. 2011;232(1):151–5.
30. Eugster A, Ruf J, Rentsch J, Köppel R. Quantification of beef, pork, chicken and turkey proportions in sausages: use of matrix-adapted standards and comparison of single versus multiplex pcr in an interlaboratory trial. Eur Food Res Technol. 2009;230(1):55.
31. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a web browser. BMC Bioinformatics. 2011;12(1):385.
32. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big data: astronomical or genomical?. PLoS Biol. 2015;13(7):1002195.
33. Schmidt B, Hildebrandt A. Next-generation sequencing: big data meets high performance computing. Drug Discov Today. 2017;22(4):712–7.

Publisher's Note

2.4 High-Throughput Seafood Surveillance by All-Food-Sequencing reveals Mislabelling and Hidden Allergens

Hellmann SL, Huntemann T, Peter M, Müller A, Schmidt B, Hankeln T

Own contributions to this publication:

- Sample acquisition: Preparation of calibration samples and collection of seafood samples
- Data analysis: Bioinformatic AFS analysis of calibration samples and seafood samples (with T. Huntemann)
- *De novo* assembly: Generation of authentic genomic references for AFS analysis (with M. Peter)
- Quantification normalization: Implementation of genome-size scaling for read counts across species
- Matrix calibration: Calculation of calibration to correct matrix-driven bias
- Microbiome profiling: Identification of microbial species in unclassified reads
- Visualization: design and refinement of all schematics

Experimental design, data analysis, data interpretation, and drafting of the manuscript were conducted in collaboration with Prof. Dr. T. Hankeln. The project was managed by Prof. Dr. T. Hankeln.

High-Throughput Seafood Surveillance by All-Food-Sequencing reveals Mislabelling and Hidden Allergens

S. Lukas Hellmann[1,2,*], Theresa Huntemann[1], Matthias Peter[1,3], André Müller[4], Bertil Schmidt[4], Thomas Hankeln[1]

[1] Institute of Organismic and Molecular Evolution, Molecular Genetics & Genome Analysis, Johannes Gutenberg University Mainz, Mainz, Germany

[2] Nucleic Acids Core Facility, Johannes Gutenberg University Mainz, Mainz, Germany

[3] TRON-Translational Oncology, Medical Center of Johannes Gutenberg University Mainz gGmbH, Mainz, Germany

[4] Department of Computer Science, Johannes Gutenberg University Mainz, Mainz, Germany

[*] Corresponding author: lukas.hellmann@uni-mainz.de

Abstract

Accurate seafood authentication and quantification protect consumers, deter fraud, and support regulatory enforcement. All-Food-Seq (AFS) is an untargeted whole-genome shotgun sequencing workflow that can profile a broad range of taxa across domains of life, but its quantitative performance in fish-dominated processed foods has not been systematically assessed. Here, we applied AFS-MetaCache to defined "FishCal" calibration sausages containing five economically important fish species and to 11 commercial seafood products (fish cake, surimi, paella, and shrimp pâté). Major ingredients were consistently identified. Quantification improved markedly after genome-size scaling and further improved after applying calibration functions derived from FishCal to mitigate matrix-specific biases: the mean absolute deviation between expected and measured proportions in calibration samples decreased from 44.5 % to 19.9 % (genome-size scaled) and to 14.2 % after calibration. In retail products, AFS

confirmed declared ingredients and additionally detected undeclared taxa (e.g., squid, celery, and Japanese threadfin bream). Species-level discrimination within closely related taxa remained challenging in some cases. Microbial community profiles varied substantially across products and were more consistent with processing and storage than with recipe. Overall, AFS combined with genome-size scaling enables high-throughput, comprehensive surveillance of eukaryotic constituents and accompanying microbiota in processed seafood.

Keywords

Food authentication – label verification – WGS – NGS – quantitative metagenomics – taxonomic profiling – species quantification – calibration materials – seafood – fishcake – surimi

Introduction

Accurate species identification in seafood is essential for consumer protection, regulatory enforcement, and fraud prevention, particularly in product categories prone to substitution and mislabelling (Chang et al., 2016; Galimberti et al., 2013; Liou et al., 2020). Beyond economic deception, undeclared species can violate dietary preferences and religious requirements and may pose health risks, for example through allergen exposure (Everstine et al., 2013; Huck et al., 2016).

DNA barcoding and, more recently, metabarcoding are widely used for seafood authentication by amplifying defined marker loci (e.g., COI for animals) and comparing sequences against reference databases such as BOLD or GenBank (Galimberti et al., 2013; Macher et al., 2023; Sayers et al., 2022). While these PCR-based approaches are sensitive and cost-effective, they rely on prior marker selection and primer binding, which constrains taxonomic scope and introduces amplification bias that can compromise quantitative interpretation, especially in complex or highly processed matrices (Kappel et al., 2023; Macher et al., 2023; Staats et al., 2016).

Untargeted whole-genome shotgun sequencing of food DNA (All-Food-Seq, AFS) aims to overcome these limitations by profiling all DNA present without defining target taxa a priori (Hellmann, Ripp, et al., 2020; Kobus et al., 2020; Ripp et al., 2014). In AFS,

sequencing reads are assigned to reference genomes, and read counts are used to estimate relative species contributions. In practice, however, quantitative accuracy can be affected by matrix-dependent DNA recovery, genome size, and incomplete or ambiguous reference representation - factors that are particularly relevant for seafood products, where closely related taxa and intensive processing are common.

Here, we evaluate AFS-MetaCache for taxonomic authentication and quantification in fish-dominated matrices using defined "FishCal" calibration sausages containing economically important fish species and a set of commercial seafood products. We assess qualitative identification performance and quantify the impact of genome-size scaling and calibration functions on agreement between expected and measured species proportions, providing a framework for improved quantitative AFS-based surveillance of processed seafood.

Materials & Methods

Sample collection and preparation

Lophius piscatorius, Melanogrammus aeglefinus, Salmo salar, Sebastes norvegicus and *Thunnus albacares* for the preparation of calibration sausages (FishCal) were purchased at a local fish monger (Fisch Jackob GmbH, Mainz, Germany). DNA extraction was outsourced to StarSEQ GmbH (Mainz, Germany) and performed following the provider's standard protocol for animal tissue. To verify the correctness of species identity, DNA barcoding was performed by StarSEQ GmbH (Mainz, Germany) by PCR amplification of a mitochondrial barcode marker COI, followed by Sanger sequencing and sequence comparison against BOLD database. Six calibration samples with various amounts of each species were prepared by weighing fish filet in proportions of 1, 4, 9, 15, 20 and 71 %, respectively. Processed food samples (SeaFood and Surimi) were bought at various supermarkets in Rhinehessen (Germany). For both calibration and processed food samples, DNA extraction and Illumina library preparation was performed by StarSEQ GmbH (Mainz, Germany). Sequencing was carried out according to the provider's protocols on an Illumina MiSeq instrument.

Bioinformatic Sequence Read Preparation

To remove sequence adapters and low-quality sequences, all samples were pre-processed using BBDuk from the BBTools software package v37.93 (https://sourceforge.net/projects/bbmap/). Sequencing adapters were trimmed if a match with k=23 (k=11, if the matched kmer is detected at the very end of the each read) against a reference sequence allowing one mismatch was detected. Subsequently 3' bases with Phred-quality <20 were discarded. If the average Phred-quality or the length of a read was below 20, the entire read was discarded. Success of quality control was inspected visually using FastQC and MultiQC (Andrew, 2010; Ewels et al., 2016).

All-Food-Seq

All food samples were analysed using AFS-MetaCache v2.5.0 (Kobus et al., 2020) in multiple steps: The initial database was constructed for each sample using the reference genomes of all ingredients listed on the product; or in case of calibration samples, all species utilised for their creation. A first round of AFS-MC was performed using the before mentioned database with parameters "-maxcand 4" and "-hitmin 4". All reads that could not be classified during the first iteration were extracted and matched against a "Barcoding Database" consisting of commonly used DNA-barcode markers from BOLD and GenBank. For all additional species identified this way, a reference genome was added to the initial AFS-MC database and a second round of AFS-MC was performed using a more defined database.

Subsequent calibration of the main components identified with AFS-MC was performed to account for the variability in genome sizes across different species, a crucial step necessitated by the principle that larger genomes yield higher concentrations of DNA during extraction and thereby leading to a higher number of reads in sequencing. To achieve a standardized comparison across all samples, normalization procedures were applied based on the respective genome sizes. The size data for each genome were primarily sourced from two comprehensive repositories, the Animal Genome Size Database (Gregory, 2025) and the Plant DNA C-values Database (Leitch IJ et al., 2019). For species where no Genome Size data was present in the databases, an estimation was performed (see next section).

Estimation of genome sizes

Addressing the challenge of missing species genome sizes, we incorporated an estimation technique: K-mer abundance in NGS read data resemble the underlying genome's composition, enabling the inference of genome size through the analysis of their distribution and abundance patterns. Sequencing reads were processed to extract data for k=23 and create kmer-histograms with the number of unique 23-mers for each sequencing depth using bbnorm.sh from the BBTools software package v37.93. These histograms were then used for GenomeScope to calculate genome size estimations for each species no represented in c-value databases (Vurture et al., 2017).

Identification of unclassified reads

All reads that remained unclassified after the second round of AFS-MC were queried against the Archaea, Bacteria and Viral sequences from the NCBI RefSeq database, release 229, accessed March 22nd 2025, (Goldfarb et al., 2025) with "-maxcand 2" and "-hitmin 0". Reads that remained unclassified even after this analysis, BLASTN v2.12.0+ was performed to map the reads to entries in the NCBI non-redundant nucleotide collection (nr/nt) database with output of Reference Taxonomic IDs (staxids). The results were filtered to only include a single best hit based on bit score for each read. The species of origin was identified for each read by matching of Taxonomic IDs with the NCBI taxonomy database.

Results

AFS-MC quantification of fish ingredients in calibration samples

As shown in previous studies, AFS-MC cal. is capable of identifying and quantifying the species proportions of meat samples (Hellmann, Ripp, et al., 2020; Kobus et al., 2020; Ripp et al., 2014). Here, we tested whether the same concept transfers to fish-dominated matrices by analysing six defined FishCal mixtures containing commonly consumed fish species. Quantification was evaluated for the default AFS-MC output, after genome-size scaling (AFS-MC norm.), and after an additional calibration step (AFS-MC cal.; Fig. 1).

With default AFS-MC, salmon was consistently overestimated across the FishCal series (Fig. 1, blue). Read-based proportions primarily reflect DNA contribution, not the weighed-in biomass fraction. Salmon is therefore expected to contribute disproportionately more DNA per gram tissue for two reasons: (i) aquaculture production frequently involves triploid salmon, and triploidy increases nuclear DNA content per cell (and thus per tissue mass) relative to diploids (Piferrer et al., 2009); and (ii) salmonids have a substantially larger genome than the other fish species in this set, a consequence of a salmonid-specific whole-genome duplication event (Macqueen & Johnston, 2014; Near et al., 2012). Together, these factors provide a coherent explanation for overestimation of salmon when proportions are inferred from raw read counts. This effect is most dominant in FishCal A, where the salmon signal exceeds the target by >8-fold.

To correct this systematic bias, we normalized species-assigned read counts by genome size. This single step markedly improved quantitative agreement across the calibration series, reducing the overall deviation from the expected composition from 44.5 % to 19.9 % after genome-size scaling (Fig. 1; AFS-MC norm., orange). Remaining discrepancies were further reduced by applying calibration functions (linear regression) to compensate residual, composition-dependent matrix effects (Fig. 1; AFS-MC cal., purple), lowering the overall deviation to 14.2 %. For salmon, the improvement is illustrated by FishCal A, where the strong overcall is reduced to ~2.3-fold after both corrections. Importantly, residual salmon deviations are not strictly one-directional after calibration (e.g., FishCal B trends slightly below the target), indicating that genome size explains a major, but not the only, driver of quantitative bias in these fish matrices.

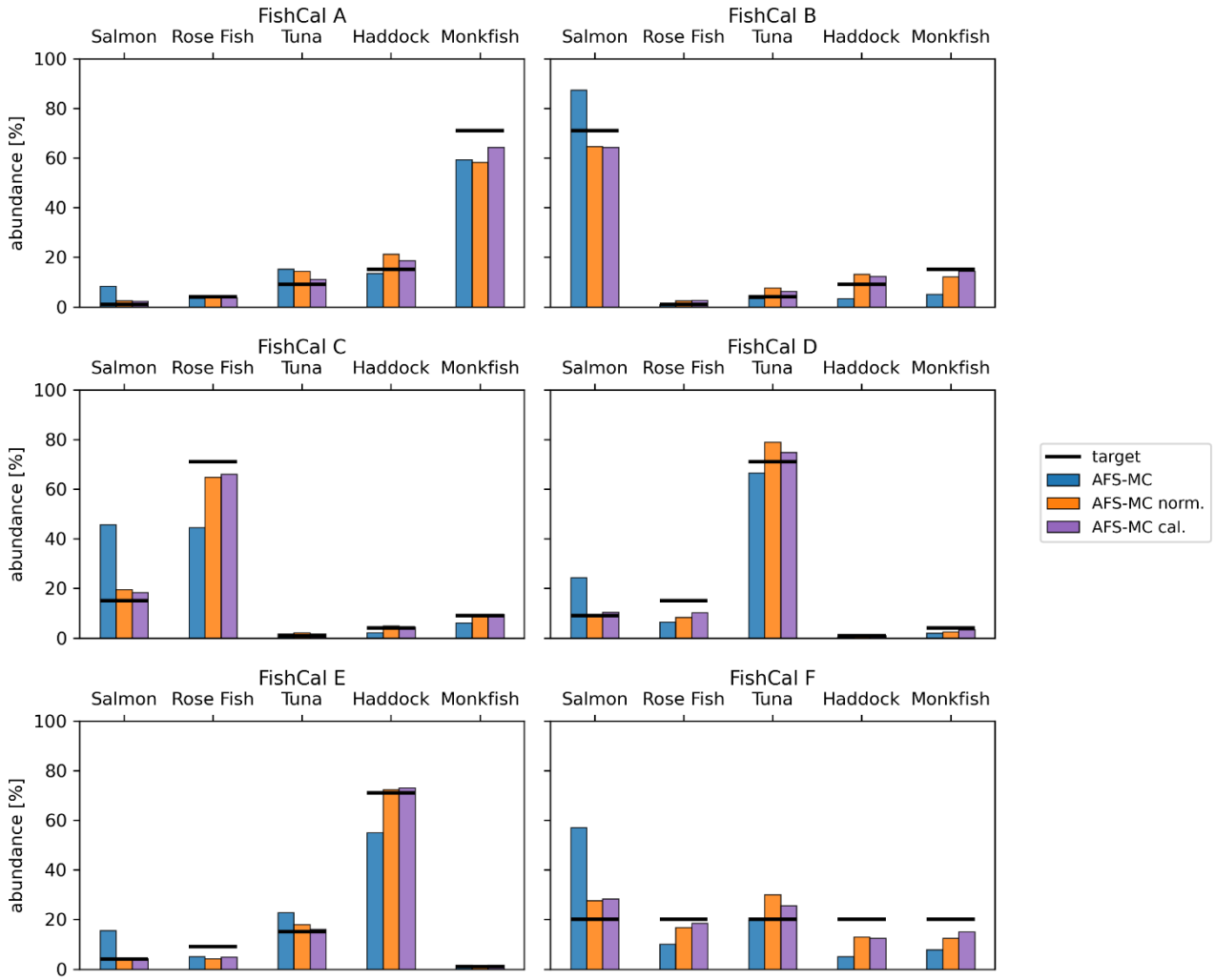

Figure 1: *Quantification of fish ingredients in calibration samples (FishCal A–F). For each sample, the expected (weighed-in) proportions are shown as black horizontal segments ("target"). Bars show the relative read abundance (%) obtained by AFS-MetaCache (AFS-MC; blue), AFS-MC after genome-size normalization (AFS-MC norm.; orange), and AFS-MC after genome-size normalization plus calibration (AFS-MC cal.; purple).*

Real fish product samples

In addition to the calibration series, we analysed 11 commercial products to assess AFS performance in real marine food matrices: six processed "SeaFood" products and five "Surimi"-style products. All samples were subjected to WGS and analysed with AFS, followed by genome-size scaling to improve quantitative comparability. Because these retail foods represent highly heterogeneous matrices and lack mixture-matched reference material, the additional calibration step used for FishCal was not applied here.

Across the SeaFood set, three samples (SeaFood1, SeaFood2, SeaFood4) were fishcake-type products (Fig. 2). SeaFood1 and SeaFood4 were declared to be primarily Alaska pollock; however, both profiles were dominated by a broader *Gadus* signal split across Alaska pollock (*G. chalcogrammus*) and other cod species (notably Atlantic cod *G. morhua* and Pacific cod *G. macrocephalus*). SeaFood4 additionally contained a substantial fraction assigned to *Merluccius* spp., dominated by *M. productus* (~31 %) alongside smaller *Merluccius* contributions and a minor pollock signal. In both fishcake products, low-level plant ingredients (e.g., onion; cereals such as rice or wheat) were detected, consistent with the expectation of binding agents in processed foods.

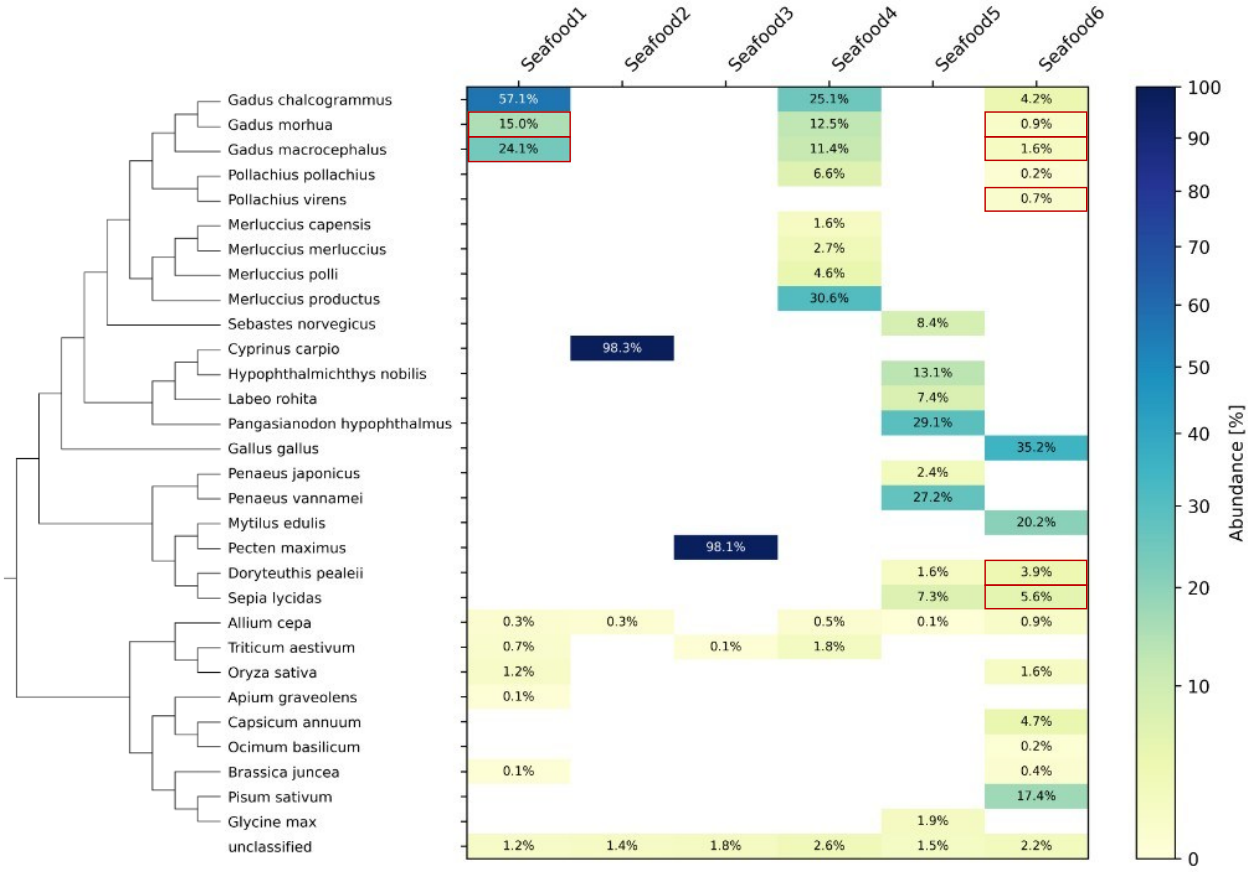

Figure 2: *Species composition of commercial SeaFood products assessed by AFS. Heatmap shows relative abundances (%) of detected taxa normalized for genome size. Cell color encodes abundance (0-100 %) and numbers indicate the corresponding percentages. The "unclassified" row denotes reads remaining unassigned after the eukaryotic ingredient analysis. The tree on the left shows phylogenetic relationship. Unlabelled ingredients are highlighted in red.*

SeaFood2 ("carp grilled sausage") was essentially a single-species product (*Cyprinus carpio*, ~98 %), with only minor plant components. Similarly, SeaFood3 ("scallops with sauce") consists of almost entirely *Pecten maximus* accounting for ~98 % of the

classified fraction (Fig. 2). By contrast, SeaFood5 (shrimp pâté) showed a more complex composition: two shrimp taxa (*Penaeus vannamei* ~27 % and *P. japonicus* ~2 %) co-occurred with several fish species (dominated by pangasius *Pangasianodon hypophthalmus* ~29 %, plus carp, rohu and rose fish) and cephalopods (notably cuttlefish *Sepia lycidas* ~7 %; Fig. 2). SeaFood6 (paella-style with chicken) was treated by sequencing only the non-rice fraction to increase effective depth for the mixed ingredients; nevertheless, a small rice signal remained (~2 %), while chicken was the major component (~35 %) together with mussel (~20 %), peas (~17 %), and cephalopods (squid/cuttlefish both in the single-digit percent range). Minor signals from cod/pollock-related taxa were also present.

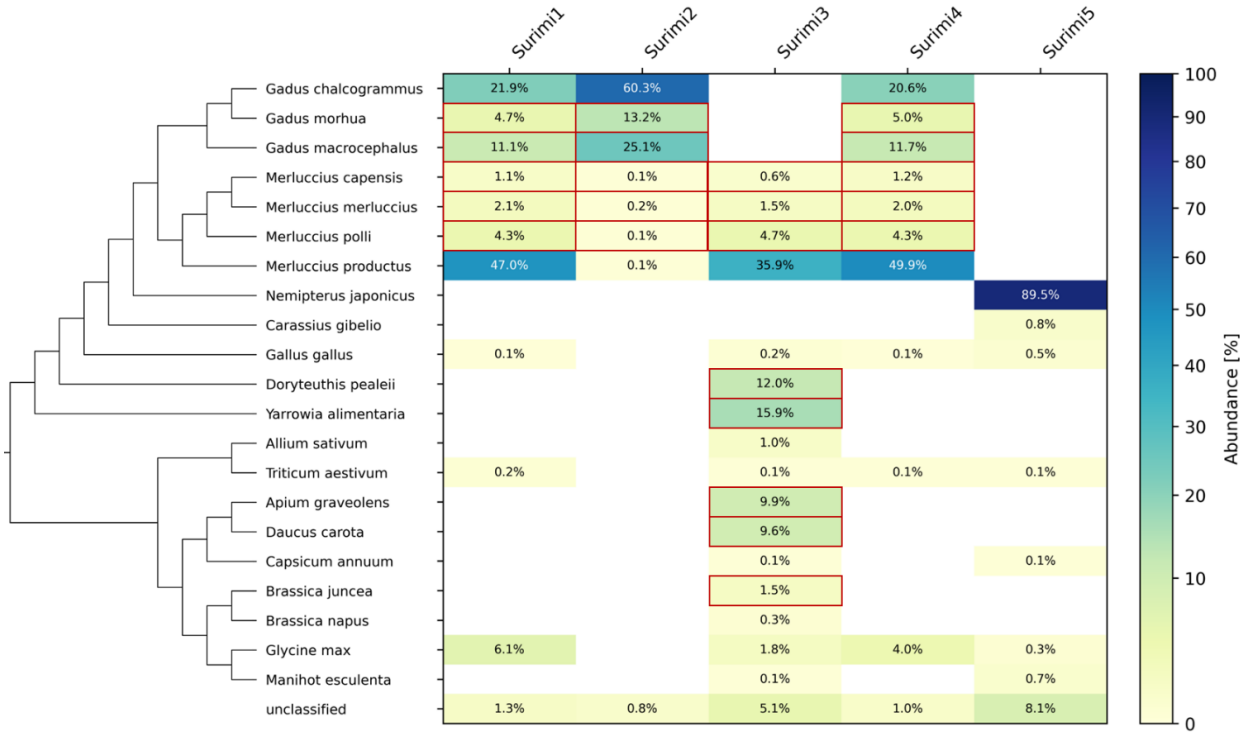

Figure 3: *Species composition of commercial Surimi products assessed by AFS. Heatmap shows relative abundances (%) of detected taxa normalization for genome size. Cell color encodes abundance (0-100 %) and numbers indicate the corresponding percentages. The "unclassified" row denotes reads remaining unassigned after the eukaryotic ingredient analysis. The tree on the left shows phylogenetic relationship. Unlabelled ingredients are highlighted in red.*

The Surimi products (Surimi1-Surimi5) showed both brand-to-brand variability and recurring patterns typical for surimi matrices. Surimi1 was dominated by North Pacific hake (*Merluccius productus*, ~47 %) and Alaska pollock (*G. chalcogrammus*, ~22 %), with additional contributions from other *Gadus* assignments and a notable soy fraction (~6 %; Fig. 3). This pattern supports the declared "pollock and/or hake"-type

formulation, while also illustrating that species-level assignments within closely related groups (e.g., *Gadus*/*Merluccius*) can distribute across multiple near neighbours in the reference set. Surimi2 was almost exclusively *Gadus* (~99 %), dominated by *G. chalcogrammus* (~60 %) with additional *Gadus* contributions, consistent with a Alaska pollock-based surimi.

Surimi3 stood out as the most compositionally complex sample: alongside a hake-dominated fish fraction (*M. productus* ~36 %), it contained substantial signals from squid (~12 %), vegetables (celery ~10 % and carrot ~10 %), and an unexpectedly large fraction assigned to the yeast *Yarrowia alimentaria* (~16 %; Fig. 3). Surimi4 showed a mixed fish composition with *M. productus* (~50 %) and *Gadus* (~37 %) plus soy (~4 %). Surimi5 differed markedly from the others, being dominated by Japanese threadfin bream (*Nemipterus japonicus*, ~90 %) and showing the highest "unclassified" fraction (~8 %), indicating incomplete reference representation in the database.

Microbial communities identified

In addition, we characterised the microbial communities for each sample at genus level. Across all products, Pseudomonas was detected and reached high relative abundances in several cases, including SeaFood2, where it represented the dominant genus (Fig. 4). However, most samples were dominated by other taxa in a strongly product-specific manner: SeaFood1 was dominated by Mycobacterium (with a substantial Lactobacillus fraction), SeaFood3 by Cutibacterium, SeaFood4 by Aeromonas, SeaFood5 by Shewanella, and SeaFood6 showed a mixed dominance of Rouxiella and Vibrio. The surimi products also differed markedly from each other, with dominance patterns shifting between Corynebacterium (Surimi1; with notable Bacillus), Xanthomonas (Surimi2), Periweissella (Surimi3; alongside elevated Acinetobacter), and a lactic-acid-bacteria-dominated profile in Surimi4 (Streptococcus with high Lactobacillus and Limosilactobacillus). Surimi5 was dominated by Citrobacter (approaching roughly half of the microbial community) and additionally showed pronounced signals of Streptomyces, Acinetobacter and Psychrobacter. Overall, the microbial profiles were highly sample-specific and did not correlate to the declared main ingredients: even products within the same category (e.g., the surimi set or the fishcake-type samples) displayed clearly distinct community patterns.

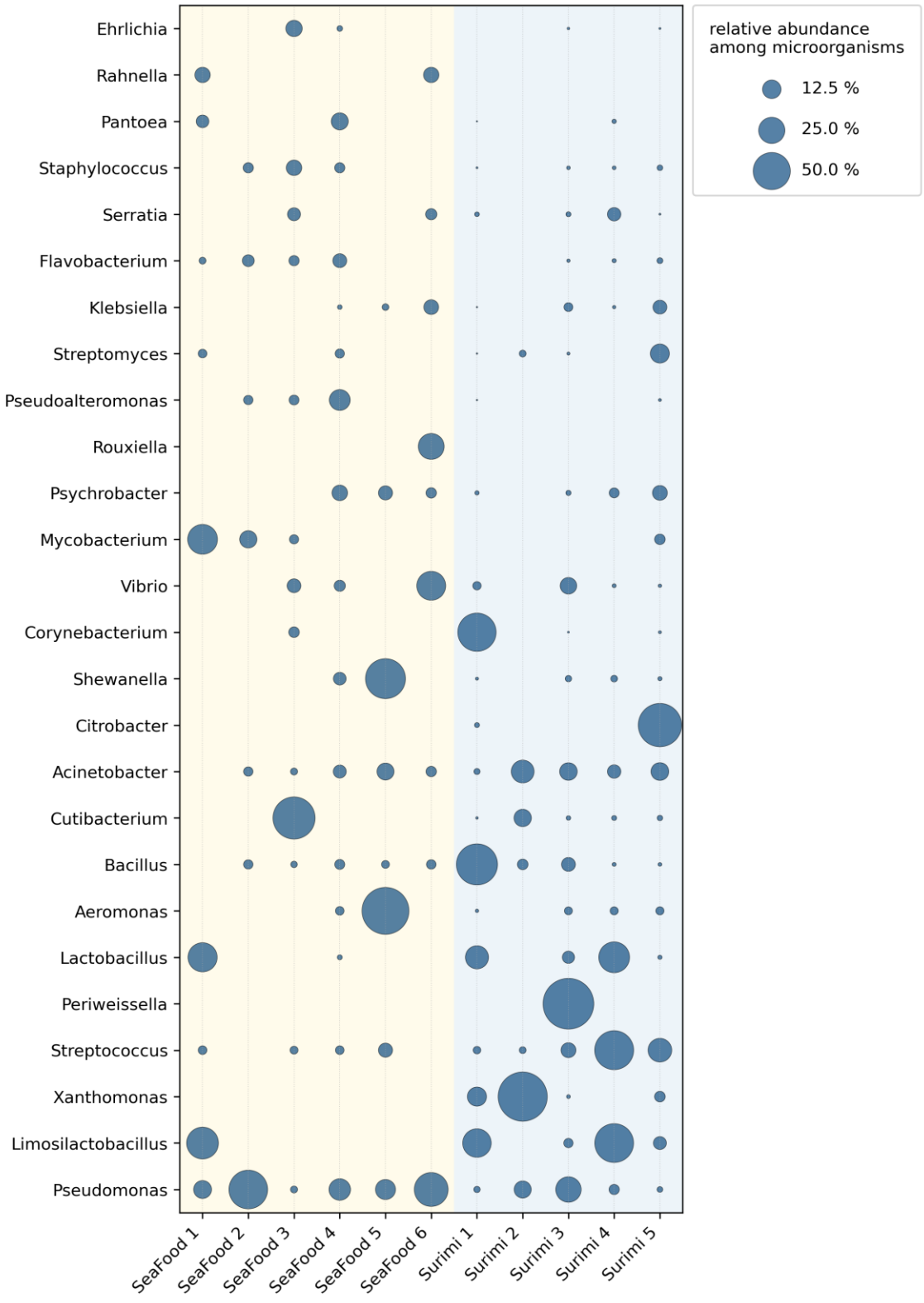

Figure 4: *Genus-level microbial profiles of SeaFood and Surimi samples. Bubbles show the relative abundances of the most common bacterial genera per sample; bubble size corresponds to the genus' relative abundance within the microbial fraction of the respective sample. Background shading separates SeaFood (yellow) and Surimi (blue) products.*

Qualitative identification

For most samples, the fraction of reads remaining unclassified after AFS analysis was low, consistent with comprehensive assignment of the major ingredients. To investigate the origin of the residual unclassified fraction and identify species missed during the main analysis, we subjected these reads to large-scale BLAST searches. Across all samples, a consistent pattern emerged. First, a subset of reads matched species that were already detected by AFS, indicating that part of the unclassified fraction reflects reads from known ingredients that were not assigned by the classifier. Second, BLAST revealed additional microbial sequences that were not captured in the initial AFS profiling, demonstrating that the unclassified fraction also contains low-abundance microbiota. Third, BLAST recovered traces of declared plant ingredients that fell below the AFS reporting threshold of 0.01 % (e.g., parsley in Surimi3) and were therefore omitted from the main results. Finally, some BLAST hits pointed to taxa that cannot be represented in the primary AFS workflow because suitable reference genomes are missing (e.g., *Trachurus trachurus* detected in Surimi5).

Discussion

Quantification Accuracy of AFS for Fish Ingredients

AFS has shown promising capabilities in the accurate quantification of species proportions in mixed meat samples as reported in prior studies. Our current investigation expands these findings by exploring the potential of AFS in quantifying fish ingredients within complex food matrices. For an initial test of the method, we created calibration sausages from commonly consumed fish species, before undergoing AFS analysis.

Across FishCal A–F, default AFS-MC systematically overestimated salmon, reaching >8-fold in the most extreme case. This bias is mechanistically plausible because read proportions primarily reflect the relative amount of recovered DNA rather than biomass. Salmonids carry an unusually large genome due to a salmonid-specific whole-genome duplication, and aquaculture salmon can additionally be triploid in some production contexts; both factors increase nuclear DNA content per unit tissue and therefore inflate read-based abundance estimates if uncorrected (Macqueen & Johnston, 2014; Near et al., 2012; Piferrer et al., 2009). Genome-size scaling directly targets this source of bias and substantially improved quantitative agreement in

FishCal, reducing the average per-sample deviation from the expected composition from 44.54 % to 19.90 %. After scaling, salmon estimates moved much closer to the targets, while a smaller residual overestimation remained for tuna (generally <2-fold), and the remaining species showed mixture-dependent over- and underestimation.

However, genome-size scaling is not a universal "always improves" correction; its usefulness depends on whether genome size is the dominant driver of bias in the specific matrix. In our earliest AFS work, a similar c-value-based normalization did not improve overall quantification in mixed meat sausages (Ripp et al., 2014), likely because tissue composition introduced strong, opposing biases (e.g., low-DNA lard vs. higher-DNA skin), which can outweigh genome-size effects (Cai et al., 2014; Köppel et al., 2011). In contrast, FishCal was prepared only from fish fillet, reducing tissue-type heterogeneity and making genome-size differences a major, correctable determinant. Nevertheless, even within fillets, DNA yield per gram can vary with biological and processing factors (e.g., muscle structure, fat content, age, and handling), which can propagate into extraction efficiency and thereby affect any DNA-based quantification approach (Kirpičnikov, 1987; Rehbein & Oehlenschläger, 2009).

To address residual, composition-dependent effects beyond genome size, we additionally applied calibration functions to correct matrix effects, such as species-specific differences in DNA recovery during extraction (Cankar et al., 2006; Josefsen et al., 2015). This further reduced the remaining deviation (to 14.20 % on average) and strongly dampened the salmon bias (e.g., the average overcall decreased from ~3.7-fold to ~1.3-fold). However, this step requires either prior knowledge of the mixture or suitable calibration materials, which are typically unavailable for routine retail surveillance. For that reason, calibration is best viewed as a performance upper bound and a tool for method characterization, whereas genome-size scaling provides a broadly applicable correction that already captures a large fraction of the systematic bias observed in fish matrices.

Level of Discrimination of closely related species

The extension of the AFS-MC method from controlled calibration samples to real-world processed marine food products provides a critical validation of its practical utility. We therefore analysed eleven products from local supermarkets to unravel their ingredient composition.

Across all products that contained Gadus spp. (SeaFood1, SeaFood4, SeaFood6, Surimi1, Surimi2 and Surimi4), reads were consistently split between Alaska pollock (*Gadus chalcogrammus*), Pacific cod (*Gadus macrocephalus*), and Atlantic cod (*Gadus morhua*) (Fig. 2&3). While the absolute proportions varied between recipes, a strikingly recurrent ~4:2:1 pattern emerged for *G. chalcogrammus* : *G. macrocephalus* : *G. morhua*. This stable ratio across otherwise different products strongly suggests that the signal is not simply reflecting three independent ingredients, but rather a systematic read-allocation ambiguity within a very closely related species group. The three taxa diverged only around ~5 MYA, so extensive sequence conservation is expected (Hughes et al., 2018; Kumar et al., 2017; Y. Ma et al., 2022), and a classifier that assigns reads to the "best matching" reference can repeatedly distribute reads across near-neighbour genomes in a characteristic way. In other words, even if a sample contains predominantly Alaska pollock, conserved genomic regions can still be assigned reproducibly to Pacific cod and (spuriously) to Atlantic cod, yielding the observed 4:2:1 split. The fact that *G. chalcogrammus* is consistently the largest fraction is compatible with pollock being the true main ingredient, but the presence of Pacific cod cannot be excluded on sequence evidence alone.

Catch-area metadata provide an additional plausibility check. The products are stated to originate from the Northeast Pacific (FAO 67), where Alaska pollock and Pacific cod share habitat, whereas Atlantic cod does not occur (Thünen Institute of Fisheries Ecology). We therefore interpret the recurring *G. morhua* signal as the most likely false-positive species assignment driven by sequence similarity, while *G. chalcogrammus* is most plausibly driving the dominant component of the pattern; *G. macrocephalus* could represent either spillover or a true co-ingredient, and cannot be resolved confidently from these data. Consequently, for these products the most defensible taxonomic statement from AFS alone is Gadus spp., rather than a fully reliable species-level confirmation.

A similar limitation was observed for hake in Surimi1, Surimi3, Surimi4, and SeaFood4. In each case, multiple Merluccius spp. were reported, but assignments were strongly

skewed toward Pacific hake (*Merluccius productus*), which exceeded the other *Merluccius* signals by approximately 7- to 70-fold (Fig. 2&3). Several hake species in the dataset are closely related (e.g., divergence times around ~4 MYA), whereas *M. productus* is reported as more distantly related (~17 MYA; Kumar et al., 2017; Rabosky et al., 2018). Despite this increased divergence, many genomic regions remain conserved within the genus, which can still lead to ambiguous read assignment and low-level false-positive detections of congeners. Importantly, when we analysed a control sample consisting of 100 % Pacific hake, we observed a similar low-level redistribution of reads to other *Merluccius* species (Supplement), supporting the interpretation that these minor signals can arise from classifier ambiguity rather than true additional ingredients. As the labelling lists Pacific hake as the ingredient and no other hakes are expected for the stated Northeast Pacific origin (Thünen Institute of Fisheries Ecology), *M. productus* is the most plausible true contributor. Nevertheless, as for cod, the conservative conclusion remains Merluccius spp. unless additional orthogonal evidence (e.g., targeted markers or catch documentation) is available.

Overall, the recurring structured splits (e.g., the 4:2:1 pattern within *Gadus*) highlight a practical boundary of untargeted WGS-based ingredient profiling: for very recently diverged taxa with high sequence similarity, AFS read assignment can be reproducibly biased across near-neighbour references, and genus-level reporting is the robust interpretation even when one species dominates the assignments.

Identification of Species substitutions

Surimi5 illustrates a different, regulatory-relevant ambiguity: it was marketed using the term "whitefish", which is a culinary label rather than an unambiguous taxonomic group. In Germany, the Federal Office for Agriculture and Food lists a defined set of species that may be marketed under this designation (German Federal Office for Agriculture and Food, 2025). Our AFS profile, however, was dominated by Japanese threadfin bream (*Nemipterus japonicus*; ~90 %). This species is commonly used as a "whitefish" in parts of Asian cuisine (Guo et al., 2020), but it is not included in the German list for marketing as such and would therefore likely be challenged in official control. Given that the product was imported from Asia and that "whitefish" is used differently across culinary contexts, this pattern is most consistent with mislabelling driven by differing naming conventions, rather than deliberate substitution for economic gain.

Beyond taxonomic ambiguity, several retail samples raise clear labelling and consumer-protection issues. Surimi3 contained a substantial celery signal (~10 %) without declaration. In contrast to the *Gadus*/*Merluccius* cases, celery is taxonomically distant from the main ingredients, making a classification artefact due to close relatives unlikely; while the exact proportion may still be imperfect, the presence of celery is plausible and relevant. This is particularly problematic because celery is a regulated allergen that must be declared to protect consumers (European Parliament & Council, 2011; Köppel et al., 2014; Worm et al., 2014). Similarly, the "paella-style with chicken" product (SeaFood6) contained additional seafood components, including mussels and cephalopods, and also showed a cod signal. Even if these ingredients are common in paella-style dishes, they need to be declared clearly to avoid consumer deception and to mitigate potential health risks.

Trace identification

Many of the low-abundance plant detections can be plausibly explained by culinary additives and spices, including basil, garlic, parsley, and mustard, which were typically present at <0.1 %. Such ingredients are generally used in small quantities because of their strong flavour. Across all surimi products, paprika was consistently detected at low levels (~0.01-0.09 %). This is consistent with its common use as a colouring agent to impart the characteristic red appearance associated with crab-like surimi products.

SeaFood3 provides a useful example of how processing can decouple DNA-based estimates from ingredient mass. This product consisted almost exclusively of great scallop, but we also detected traces of cattle DNA. SeaFood3 is marketed as "Deluxe scallops filled with creamy white wine sauce" and contains dairy ingredients (e.g., clarified butter, skimmed milk powder, crème fraîche), which plausibly explain the cattle signal. Because these dairy components are highly processed and often fat-rich, they contain comparatively little recoverable DNA per unit weight. Consequently, their contribution can be substantially underestimated by a DNA-based quantification approach, and the true mass fraction of milk-derived ingredients in the filling is likely higher than the ~0.04 % indicated by read-based estimates.

Wheat was detected in several products, mostly at low levels, and its presence is consistent with typical processing aids. In the fishcake-type products SeaFood1 and SeaFood4, wheat-associated signals (~0.7-1.8 %) likely reflect added bread-like

components or binders used in the recipe. In the surimi products, wheat occurred only in trace amounts (~0.07-0.18 %), consistent with the use of wheat-derived starch. Starches are widely applied in processed foods due to their functional properties as thickeners and stabilizers, and may again be underestimated by DNA-based quantification because they contribute relatively little DNA.

Databases as limitations for quantification

Unclassified reads varied between samples. To better understand their origin, we analysed the unclassified read fraction by BLAST. Across all products, a substantial share of these reads matched taxa that were already identified in the main AFS analysis. This indicates that "unclassified" does not necessarily imply "unknown ingredient", but often reflects technical and reference-related limitations of read assignment. Three factors likely contribute.

First, most reference genomes represent only one (or a few) individuals and therefore do not capture the full spectrum of intraspecific variation; reads originating from divergent haplotypes or structural variants may fail to match the reference sufficiently and remain unassigned (Gui et al., 2022; Sherman et al., 2018). Second, reference quality and completeness vary widely. Fragmented assemblies, misassemblies, or missing sequence can reduce mappability and thus classification success, particularly for short reads (Park et al., 2023). Third, some genomic regions are intrinsically difficult for short-read–based assignment, including repetitive or duplicated ("camouflaged") regions that lead to ambiguous mapping and conservative filtering (Ebbert et al., 2019; Ryan & Corvin, 2023). These limitations are expected to diminish as pangenome resources expand and more high-quality reference assemblies become available, but they currently contribute measurably to the residual unclassified fraction.

Long-read sequencing offers a plausible route to mitigate parts of this problem. Longer reads span repetitive elements and provide more unique context, improving mapping robustness and potentially supporting better discrimination of closely related taxa in complex mixtures (Albertsen, 2023; Kovaka et al., 2023; Nurk et al., 2022). While long reads do not eliminate reference bias, they can reduce ambiguity in regions that are problematic for short-read workflows.

In addition to imperfect assignment to known taxa, BLAST also indicated the presence of organisms for which suitable reference genomes are not available in the AFS

reference set. For example, the unclassified fraction was comparatively high (~8 %) in Surimi5, and BLAST results suggested *Trachurus trachurus* as a potential additional fish component. Without an appropriate reference genome, such taxa can at best be detected qualitatively by similarity search, but they cannot be quantified within the standard AFS read-count framework. This limitation is inherent to reference-based metagenomic quantification and will gradually become less restrictive as genome coverage in public databases increases.

Detection of potentially lacking hygiene standards

Pseudomonas spp. was detected in every sample (Fig. 4) and reached high relative abundances in several products. This is consistent with its ecology as a ubiquitous environmental genus and with its well-established role as a psychrotolerant spoilage organism in fish that can grow at refrigeration temperatures (Andreani & Fasolato, 2017; Streeter & Katouli, 2016). In addition, several genera frequently associated with marine habitats occurred repeatedly across the dataset, including Vibrio, Shewanella, Acinetobacter and Flavobacterium (Egerton et al., 2018). Their presence supports the interpretation that the initial microbial load of raw seafood material contributes to the communities observed in the final products.

At the same time, microbial profiles were highly sample-specific and did not track the declared main ingredients: even products within the same category (e.g., the surimi set) showed distinct dominance patterns. This points to processing-related factors: handling, equipment hygiene, and storage conditions as major determinants shaping the microbial community structure beyond the ingredient list alone.

Two samples showed particularly pronounced dominance of taxa that can be informative from a hygiene perspective. Surimi4 was dominated by Streptococcus, a genus commonly found in the mucosal microbiota of humans and animals and frequently used as an indicator group in contamination contexts; this pattern is compatible with post-processing introduction via personnel or contact surfaces, although sequencing alone cannot prove the contamination route (Byappanahalli et al., 2012; Iwamoto et al., 2010). Likewise, Surimi5 was dominated by Citrobacter, which may reflect contamination and/or growth during processing and storage, and warrants attention as a potential hygiene-related signal (Andeta et al., 2018).

In addition to the bacterial profiles, Surimi3 contained a conspicuous yeast signal assigned to *Yarrowia alimentaria* (syn. *Candida alimentaria*), accounting for ~16 % in the Surimi3 profile. *Yarrowia* spp. are aerobic, salt-tolerant yeasts that are repeatedly reported from protein-rich, processed and/or fermented foods, where they can either contribute to desirable ripening reactions or act as spoilage organisms depending on the product and storage conditions (Nielsen et al., 2008). Their metabolism is well suited to such matrices because many *Yarrowia* strains can utilise lipids and proteins and can generate volatile compounds and other metabolites that influence odour and flavour, which is why they are frequently discussed in the context of quality changes in processed foods (Zinjarde, 2014). Although *Y. alimentaria* itself has, to our knowledge, not been reported for surimi products, its close relative *Y. lipolytica* has been linked to spoilage of high-fat fish products, supporting the plausibility that a *Yarrowia* signal in surimi could reflect unintended growth during processing or refrigerated storage (Nielsen et al., 2008).

At the same time, *Yarrowia* spp. are also relevant as industrial production organisms. They are known producers of enzymes and food-relevant metabolites (e.g., organic acids, aroma compounds, emulsifiers/surfactants), which makes them attractive in biotechnological and food-processing contexts (Zinjarde, 2014). The use of *Y. lipolytica* biomass as a dietary supplement has been described, and defatted yeast preparations can be protein- and fibre-rich and contain notable micronutrients (Gottardi et al., 2021; Turck et al., 2019). Recent work also discusses *Y. alimentaria* in the context of proteolytic activity and protein hydrolysis, highlighting potential applications where proteases are desirable (Liu et al., 2025). Therefore, two scenarios remain plausible for Surimi3: (i) unintended contamination and growth (spoilage-associated), or (ii) deliberate or incidental technological use (processing-related). Distinguishing these explanations would require targeted follow-up, such as culture-based confirmation, strain typing, and/or time-series testing during storage to determine whether *Yarrowia* increases post-production.

Acknowledgements

We thank Dr. Gesche Spielmann and Dr. Ingrid Huber at the Bavarian Health and Food Safety Authority (Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit, LGL) for providing the samples used in this study. SLH and TH gratefully acknowledge funding by the Federal Office for Agriculture and Food (project ID: 2816503814).

Andrew, S. (2010). FastQC.

BLE. (2024). List of trade names for fishery and aquaculture products. https://www.ble.de/SharedDocs/Downloads/DE/Fischerei/Fischwirtschaft/HandelsbezeichnungDLat.html

Cai, Y., Li, X., Lv, R., Yang, J., Li, J., He, Y., & Pan, L. (2014). Quantitative Analysis of Pork and Chicken Products by Droplet Digital PCR. BioMed Research International, 2014. https://doi.org/10.1155/2014/810209

Cankar, K., Stebih, D., Dreo, T., Žel, J., & Gruden, K. (2006). Critical points of DNA quantification by real-time PCR - Effects of DNA extraction method and sample matrix on quantification of genetically modified organisms. BMC Biotechnology, 6(1), 1–15. https://doi.org/10.1186/1472-6750-6-37/TABLES/4

Chang, C. H., Lin, H. Y., Ren, Q., Lin, Y. S., & Shao, K. T. (2016). DNA barcode identification of fish products in Taiwan: Government-commissioned authentication cases. Food Control, 66, 38–43. https://doi.org/10.1016/J.FOODCONT.2016.01.034

Deiner, K., Walser, J. C., Mächler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. Biological Conservation, 183, 53–63. https://doi.org/10.1016/J.BIOCON.2014.11.018

Ebbert, M. T. W., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., Kauwe, J. S. K., Belzil, V., Pregent, L., Carrasquillo, M. M., Keene, D., Larson, E., Crane, P., Asmann, Y. W., Ertekin-Taner, N., Younkin, S. G., Ross, O. A., Rademakers, R., Petrucelli, L., & Fryer, J. D. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. Genome Biology, 20(1). https://doi.org/10.1186/S13059-019-1707-2

European Parliament & Council. (2011). REGULATION (EU) No 1169/2011.

Everstine, K., Spink, J., & Kennedy, S. (2013). Economically Motivated Adulteration (EMA) of Food: Common Characteristics of EMA Incidents. Journal of Food Protection, 76(4), 723–735. https://doi.org/10.4315/0362-028X.JFP-12-399

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354

Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martellos, S., & Labra, M. (2013). DNA barcoding as a new tool for food traceability. Food Research International, 50(1), 55–63. https://doi.org/10.1016/J.FOODRES.2012.09.036

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D. H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., … Murphy, T. D. (2025). NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. Nucleic Acids Research, 53(D1), D243–D257. https://doi.org/10.1093/NAR/GKAE1038

Gottardi, D., Siroli, L., Vannini, L., Patrignani, F., & Lanciotti, R. (2021). Recovery and valorization of agri-food wastes and by-products using the non-conventional yeast Yarrowia lipolytica. Trends in Food Science & Technology, 115, 74–86. https://doi.org/10.1016/J.TIFS.2021.06.025

Gregory, T. R. (2025). Animal Genome Size Database.

Guo, N., Chen, H. X., Zhang, L. P., Zhang, J. Y., Yang, L. Y., & Li, L. (2020). Infection and molecular identification of ascaridoid nematodes from the important marine food fish Japanese threadfin bream Nemipterus japonicus (Bloch) (Perciformes: Nemipteridae) in China. Infection, Genetics and Evolution, 85, 104562. https://doi.org/10.1016/J.MEEGID.2020.104562

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. https://doi.org/10.1098/RSPB.2002.2218

Hellmann, S. L., Ripp, F., Bikar, S. E., Schmidt, B., Köppel, R., & Hankeln, T. (2020). Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq). European Food Research and Technology, 246(1), 193–200. https://doi.org/10.1007/s00217-019-03404-y

Huck, C. W., Pezzei, C. K., & Huck-Pezzei, V. A. (2016). An industry perspective of food fraud. Current Opinion in Food Science, 10, 32–37. https://doi.org/10.1016/J.COFS.2016.07.004

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur, R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhoui, Z., Huang, H., Chen, S., … Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proceedings of the National Academy of Sciences of the United States of America, 115(24), 6249–6254. https://doi.org/10.1073/PNAS.1719358115/-/DCSUPPLEMENTAL

Josefsen, M. H., Andersen, S. C., Christensen, J., & Hoorfar, J. (2015). Microbial food safety: Potential of DNA extraction methods for use in diagnostic metagenomics. Journal of Microbiological Methods, 114, 30–34. https://doi.org/10.1016/J.MIMET.2015.04.016

Kirpičnikov, V. S. (1987). Genetische Grundlagen der Fischzüchtung. Dt. Landwirtschaftsverl.

Kobus, R., Abuín, J. M., Müller, A., Hellmann, S. L., Pichel, J. C., Pena, T. F., Hildebrandt, A., Hankeln, T., & Schmidt, B. (2020). A big data approach to metagenomics for all-food-sequencing. BMC Bioinformatics, 21(1). https://doi.org/10.1186/s12859-020-3429-6

Köppel, R., Rentsch, J., Ruf, J., Eugster, A., Graf, C., Felderer, N., Pietsch, K., & Ilg, E. (2014). Results of an International Interlaboratory Trial to Determine Twelve Allergens Using Real-time PCR- and ELISA-based Assays. Chimia, 68(10), 721–725. https://doi.org/10.2533/CHIMIA.2014.721

Köppel, R., Ruf, J., & Rentsch, J. (2011). Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, horse and sheep. European Food Research and Technology, 232(1), 151–155. https://doi.org/10.1007/s00217-010-1371-y

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Molecular Biology and Evolution, 34(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Leitch IJ, Johnston E, Pellicer J, Hidalgo O, & Bennett MD. (2019). Plant DNA C-values Database (Release 7.1).

Liou, P., Banda, A., Isaacs, R. B., & Hellberg, R. S. (2020). Labeling compliance and species authentication of fish fillets sold at grocery stores in Southern California. Food Control, 112, 107137. https://doi.org/10.1016/J.FOODCONT.2020.107137

Liu, Y., Gao, X., Zang, M., Sun, B., Zhang, S., Xie, P., & Liu, X. (2025). Insights into Microbial Community and Its Enzymatic Profiles in Commercial Dry-Aged Beef. Foods, 14(3), 529. https://doi.org/10.3390/FOODS14030529/S1

Liu, Y., Ripp, F., Koeppel, R., Schmidt, H., Lukas Hellmann, S., Weber, M., Krombholz, C. F., Schmidt, B., & Hankeln, T. (2017). AFS: identification and quantification of species composition by metagenomic sequencing. Bioinformatics, January, btw822. https://doi.org/10.1093/bioinformatics/btw822

Ma, Y., Li, Y., Jiang, C., Zheng, L., Liu, S., & Zhao, L. (2022). High-quality chromosome-level genome assembly of Pacific cod, Gadus macrocephalus. Frontiers in Marine Science, 9. https://doi.org/10.3389/fmars.2022.1067526

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. Proceedings of the Royal Society B: Biological Sciences, 281(1778). https://doi.org/10.1098/RSPB.2013.2881

Müller, A., Hundt, C., Hildebrandt, A., Hankeln, T., & Schmidt, B. (2017). MetaCache: Context-aware classification of metagenomic reads using minhashing. Bioinformatics, 33(23), 3740–3748. https://doi.org/10.1093/bioinformatics/btx520

Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Wainwright, P. C., Friedman, M., & Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. Proceedings of the National Academy of Sciences of the United States of America, 109(34), 13698–13703. https://doi.org/10.1073/PNAS.1206625109/SUPPL_FILE/ST02.DOC

Nielsen, D. S., Jacobsen, T., Jespersen, L., Koch, A. G., & Arneborg, N. (2008). Occurrence and growth of yeasts in processed meat products – Implications for potential spoilage. Meat Science, 80(3), 919–926. https://doi.org/10.1016/J.MEATSCI.2008.04.011

Piferrer, F., Beaumont, A., Falguière, J. C., Flajšhans, M., Haffray, P., & Colombo, L. (2009). Polyploid fish and shellfish: Production, biology and applications to aquaculture for performance improvement and genetic containment. Aquaculture, 293(3–4), 125–156. https://doi.org/10.1016/J.AQUACULTURE.2009.04.036

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. Nature, 559(7714), 392–395. https://doi.org/10.1038/S41586-018-0273-1

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System: Barcoding. Molecular Ecology Notes, 7(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Rehbein, H., & Oehlenschläger, J. (2009). Fishery Products: Quality, Safety and Authenticity. Fishery Products: Quality, Safety and Authenticity, 1–477. https://doi.org/10.1002/9781444322668

Ripp, F., Krombholz, C., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., & Hankeln, T. (2014). All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. BMC Genomics, 15(1), 639. https://doi.org/10.1186/1471-2164-15-639

Ryan, N. M., & Corvin, A. (2023). Investigating the dark-side of the genome: a barrier to human disease variant discovery? Biological Research, 56(1), 42. https://doi.org/10.1186/S40659-023-00455-0

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. Nucleic Acids Research, 50(D1), D20–D26. https://doi.org/10.1093/NAR/GKAB1112

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 2016 408:17, 408(17), 4615–4630. https://doi.org/10.1007/S00216-016-9595-8

Thünen Institute of Fisheries Ecology. (n.d.). Nordostpazifik - Fischbestände (Northeast Pacific - Fish stocks). Retrieved May 29, 2024, from https://www.fischbestaende-online.de/fao-fanggebiete/nordostpazifik

Turck, D., Castenmiller, J., de Henauw, S., Hirsch-Ernst, K., Kearney, J., Maciuk, A., Mangelsdorf, I., McArdle, H. J., Naska, A., Pelaez, C., Pentieva, K., Siani, A., Thies, F., Tsabouri, S., Vinceti, M., Cubadda, F., Engel, K., Frenzel, T., Heinonen, M., … Knutsen, H. K. (2019). Safety of Yarrowia lipolytica yeast biomass as a novel food pursuant to Regulation (EU) 2015/2283. EFSA Journal, 17(2), e05594. https://doi.org/10.2903/J.EFSA.2019.5594

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics, 33(14), 2202–2204. https://doi.org/10.1093/BIOINFORMATICS/BTX153

Worm, M., Jappe, U., Kleine-Tebbe, J., Schäfer, C., Reese, I., Saloga, J., Treudler, R., Zuberbier, T., Waßmann, A., Fuchs, T., Dölle, S., Raithel, M., Ballmer-Weber, B., Niggemann, B., & Werfel, T. (2014). Food allergies resulting from immunological cross-reactivity with inhalant allergens: Guidelines from the German Society for Allergology and Clinical Immunology (DGAKI), the German Dermatology Society (DDG), the Association of German Allergologists (AeDA) and the Society for Pediatric Allergology and Environmental Medicine (GPA). Allergo Journal International, 23(1), 1. https://doi.org/10.1007/S40629-014-0004-6

Zinjarde, S. S. (2014). Food-related applications of Yarrowia lipolytica. Food Chemistry, 152, 1–10. https://doi.org/10.1016/J.FOODCHEM.2013.11.117

2.5 Improved Metagenomic Analysis for All-Food-Sequencing with AFS-MetaCache2: Illumina vs. Nanopore

Müller A, Wichmann A, Kallenborn F, Hellmann SL, Hankeln T, Schmidt B

Submitted to: bioRxiv; planned: BMC Bioinformatics

DOI: https://doi.org/10.64898/2025.12.18.694891

Own contributions to this publication:

- Data generation: Sequencing of calibration samples on PromethION
- Data analysis: Comparative evaluation of quantification accuracy and false-positive rates for Illumina short-read and Oxford Nanopore long-read (with A. Müller)
- Writing: Original draft

Experimental design, data analysis, data interpretation, and drafting of the manuscript were conducted in collaboration with A. Müller, Prof. Dr. T. Hankeln and Prof. Dr. B. Schmidt. The project was managed by Prof. Dr. T. Hankeln and Prof. Dr. B. Schmidt.

Improved Metagenomic Analysis for All-Food-Sequencing with AFS-MetaCache2: Illumina vs. Nanopore

André Müller[1*], Alexander Wichmann[1], Felix Kallenborn[1], S. Lukas Hellmann[2,3], Thomas Hankeln[2*], Bertil Schmidt[1*]

[1]Department of Computer Science, Johannes Gutenberg University, Mainz, 55099, Germany.
[2]Institute of Organismic and Molecular Evolution, Johannes Gutenberg University, Mainz, 55099, Germany.
[3]Nucleic Acids Core Facility, Johannes Gutenberg University, Mainz, 55099, Germany.

*Corresponding author(s). E-mail(s): muellan@uni-mainz.de; hankeln@uni-mainz.de; bertil.schmidt@uni-mainz.de;
Contributing authors: alwichma@uni-mainz.de; kallenborn@uni-mainz.de; lukas.hellmann@uni-mainz.de;

Abstract

Background: All-Food-Sequencing (AFS) is a method for untargeted metagenomic analysis that allows for the detection and quantification of food ingredients. While this approach avoids some of the shortcomings of targeted PCR-based methods, its performance depends on sequencing technologies, taxonomic classification tools, and genomic reference databases.

Results: AFS-MetaCache2 implements an improved reference database construction mechanism compared to prior approaches. To demonstrate the effectiveness to AFS, we sequenced sausages composed of mammalian and avian species using both short-read (Illumina) and long-read (Oxford Nanopore Technologies) platforms. While both approaches reliably detect the main components, our comparison shows that long-read sequencing is superior in terms of both quantification accuracy and false positive rates. The evaluation of representative metagenomic tools (Kraken2+Bracken, KrakenUniq, AFS-MetaCache1) demonstrates that AFS-MetaCache2 yields the best accuracy and fastest database build times, while reducing peak main memory consumption. It thus allows for efficient scaling to large reference genome sets.

Conclusion: Our study suggests that deep sequencing of total genomic DNA from samples with heterogeneous taxon composition, using 3rd generation sequencing technology followed by metagenomic analysis with AFS-MetaCache2, is a valuable approach for bio-surveillance of food ingredients. Our software is available at https://github.com/muellan/metacache.

Keywords: Next-generation sequencing, Third-generation sequencing, Long-read sequencing, Quantitative metagenomics, Species identification, Food authentication, Eukaryotic genomes, Big data

Background

Ensuring safety and quality standards is of high importance to the food industry. However, errors and fraud in the production and wrong labeling of food have garnered political and media attention in recent years. This motivates the need for analytical methods that allow for accurate monitoring of food species composition, ideally spanning various kingdoms of life including animals, plants, bacteria, fungi, and viruses. Quantitative real-time polymerase chain reaction (qPCR) [1–3], droplet digital PCR (ddPCR) [4–6], and sequencing of species-specific DNA bar codes [7–10] are technologies for food control that are widely used in practice. However, they are limited by the number and phylogenetic diversity of target species within a single assay and thus are not suitable for broad-scale screening.

As an alternative technology, high-throughput sequencing of metagenomic DNA from food samples has the potential to simultaneously screen for a wide range of species. Although prior definition of possible target species is not needed, subsequent bioinformatic analysis based on comparisons to genomic databases is required to identify and quantify actual food components. Our original All-Food-Seq (AFS) approach [11] mapped Illumina sequencing reads to a (small) number of reference genomes using BWA, and then determined species composition and relative quantities based on a read counting procedure. This detected anticipated species in food products with quantification accuracy comparable to ddPCR and simultaneously identified unexpected species in an untargeted approach [12].

Application of this approach in practical settings faces two challenges;

1. storing large amounts of reference genomes in increasingly big databases, and
2. accurate bioinformatics solutions for mapping and quantification.

To address these issues, AFS-Metacache1 [13] proposed a k-mer-based exact matching approach to compare each read to a database of reference genomes. To gain efficiency, subsampling of k-mers based on minhashing is employed to reduce both memory consumption and database construction times. Nevertheless, when considering scaling to large reference genome collections, the associated databases may still consume hundreds of gigabytes or even terabytes of memory, which makes scaling challenging.

In addition, false positive read classifications might be caused by regions of DNA shared among multiple reference genomes due to sequence conservation. A number of

prior studies [14–16] in the field of microbial metagenomics or marker gene sequencing based on Illumina short-read sequencing have shown that performance depends both on the taxonomic classifier and utilized reference databases. Recent work further indicted that the accuracy of microbial taxonomic classification and profiling methods can be further improved by using long-read sequencing [17–20]. However, similar studies of metagenomics with complex eukaryotic genomes have not been conducted so far.

In this paper we therefore address both the **accuracy** and the **scalability** challenge of k-mer based AFS. Our contributions are two-fold:

Long read sequencing: We are the first to investigate the impact of longer sequencing reads to broad-scale identification and quantification of species composition in food. Our experimental results using a number of calibrator sausages of known species sequenced with both Illumina and Oxford Nanopore Technologies (ONT) platforms show that using long-read datasets yield lower false-positive rates and at the same time can provide higher quantification accuracy.
Scalable database construction: We also address the drawback of high main memory consumption and long database construction times of contemporary classification tools by proposing a novel reference genome database construction scheme. It produces a taxonomy-aware partitioning that automatically assigns similar reference genomes to different partitions by minimizing bucket overflows. The corresponding implementation is provided in *AFS-MetaCache2*. Our performance evaluation shows that AFS-MetaCache2 yields lower false-positive rates and higher quantification accuracy compared to Kraken2+Bracken [21, 22], KrakenUniq [23, 24], and AFS-MetaCache1 [13], while reducing both database construction times and peak main memory consumption by taking advantage of modern multi-core CPUs. Therefore, our new approach enables scalability to growing genome collections which is needed for practical broad-scale food screening.

Methods

DNA sequencing of sausage calibration samples

Thirteen calibration sausage samples containing admixtures of cattle, chicken, pig, sheep, horse and turkey at defined amounts were produced by a professional butchery and provided by the Official Food Control Authority of the Canton Zürich, Switzerland (see Table 1) [1]. The samples were prepared for calibration of foodstuff detection methods and reflect three different recipes of sausage production:

Kal A-E: all meat sausage,
KLyo A-D: Lyoner style sausage (matrix of meat, rind and lard), and
KGeflLyo A-D: poultry Lyoner (matrix of meat and skin).

Total DNA was extracted out of 200 mg homogenized sausage samples using the Wizard Plus system (Promega, 117 Madison, USA) according to the manufacturer's protocol.

Oxford Nanopore sequencing

DNA concentration was quantified using a DS-11 FX+ spectrophotometer/fluorometer (DeNovix). DNA integrity was assessed using a Fragment Analyzer 5200 with the DNF-474-1000 kit (Agilent Technologies).

Libraries were prepared using the Native Barcoding Kit 24 V14 (SQK-NBD114.24; Oxford Nanopore Technologies) according to the manufacturer's ligation sequencing gDNA native barcoding v14 protocol. Briefly, 400 ng per sample was subjected to DNA repair using Ultra™ II End Repair/dA-Tailing Module (New England Biolabs). Native barcodes were ligated using the NEB Blunt/TA Ligase Master Mix (New England Biolabs). Library concentration was measured using the Qubit dsDNA HS assay (Thermo Fisher Scientific). Barcoded samples were purified using 0.4 x AMPure XP Beads (Beckmann Coulter) and pooled equimolarly. Sequencing adapters were ligated to the pooled DNA using Quick T4 DNA Ligase (New England Biolabs) and the library was loaded onto a R10.4.1 PromethION flow cell (FLO-PRO114M). Sequencing was performed on a PromethION 2 Solo for 48 h.

For downstream analysis, super-accuracy (SUP) basecalling was performed using Guppy basecall server v7.1.4 with model *dna_r*10.4.1_*e*8.2_400*bps_sup*@*v*4.2.0 with demultiplexing and adapter trimming enabled and discarding low quality reads (PHRED <8). This resulted in 1.9 to 4.3 million reads per sample (median: 1.5 million) with median read length of 374 bp and maximum read length of 21 kbp.

Illumina sequencing

Sequencing library preparation and sequencing were performed by a commercial provider (StarSEQ, Mainz, Germany). The Nextera DNA Library Preparation Kit (Illumina, San Diego, USA) was applied following the manufacturer's instructions. Typically, 1 ng of total DNA was used. Sequencing was carried out on an Illumina MiSeq instrument using reagent kit v.2 in 150 bp paired-end mode. All datasets were quality checked, trimmed and filtered by using the FASTQC data evaluation software and the trimmomatic v0.33 trimming tool. This resulted in 0.195 to 1.659 million reads per sample (median: 0.8 million) with median read length of 126 bp and maximum read length of 136 bp.

Datasets for both Illumina and ONT sequencing have been submitted to the SRA database under the project name PRJEB34001.

AFS-MetaCache2 Pipeline

Our pipeline can be separated into two distinct phases: *build* and *query*. AFS-MetaCache2 extends the pipeline of AFS-MetaCache1 (as presented in [13]) by introducing a novel method for scalable database construction in the build phase. Figure 1 provides an overview of both phases which are further outlined in the following.

Build Phase

We consider a collection of N reference genomes as input. Each reference genome is divided into windows of size l which overlap by $k-1$ base-pairs. For each window a

sketch is calculated using minhashing [25]. A sketch consists of the s smallest k-mers (in strand-neutral canonical representation) contained in the window with respect to an applied hash function. Thus, the sketching procedure selects only a subset of k-mers to be inserted into the database used for similarity computation, typically reducing their amount by around one order-of-magnitude.

The database is stored as a hash table consisting of key-target-list pairs. A second hash function maps k-mers to slots in the hash table. If an identified slot is empty or occupied with the same k-mer, the corresponding k-mer is inserted as key and the corresponding location (genome ID, window ID) is appended to the target-list (see *Genome Hash Tables* in Figure 1).

Due to the ever increasing size of reference genome collections, the memory of a single workstation might not be sufficient to create a database for all considered reference sequences at once. Thus, we have developed a new approach to partition the reference genomes into separate databases (hash tables) which allows us to not only reduce memory consumption; but also limit the amount of overflowing buckets by storing similar genomes in different partitions. At the same time our approach reduces runtimes by employing multiple threads of common multi-core CPUs.

Our new database construction procedure for a total of N reference genomes $G_1, \ldots, G_N$ works as follows. If t CPU threads are available in total, $\min(t-1, N)$ individual hash tables T_i for each of the reference sequences G_i are constructed in parallel. They are then subsequently merged into larger hash tables using a separate, concurrently running thread. By default all genome-specific tables T_i are merged into one unified hash table which results in a database as produced by previous versions of AFS-MetaCache2. However, if the user either sets a maximum memory size per database partition or the number of individual partitions or both of these options, the genome-specific tables will be merged into separate hash tables each representing a database partition.

Since the distribution of k-mer occurrences in different genomes is usually highly skewed, a large fraction of k-mers occur only once while few occur thousands or even millions times. To limit memory consumption, the maximum number of locations stored per k-mer is limited to a predefined value (254 per default). In the case of highly similar genomes, the number of bucket collisions can become significant, since identical k-mers occur in several different genomes. This may lead to inaccurate assignments for reads stemming from repetitive genomic regions.

Thus, the merging of hash tables into multiple partitions P_j aims to preserve as much genomic information as possible by minimizing the number of overflowing hash table buckets. The procedure starts with a user-defined number of required partitions (by default 1) and successively inserts the contents of the previously generated per-genome hash tables T_i into these partitions. Each insertion round for any of the tables T_i starts by computing the number of resulting hash bucket overflows that would occur when inserting into any of the partitions P_k. This can be achieved by querying each bucket key (k-mer) in T_i against P_k and in case of a match checking if the sum of the bucket sizes exceeds the maximum bucket size. Finally, the genome table is inserted into the partition P_{opt} that incurs the least amount of overflowing buckets and has the smallest size. If inserting the current genome table would exceed the maximum

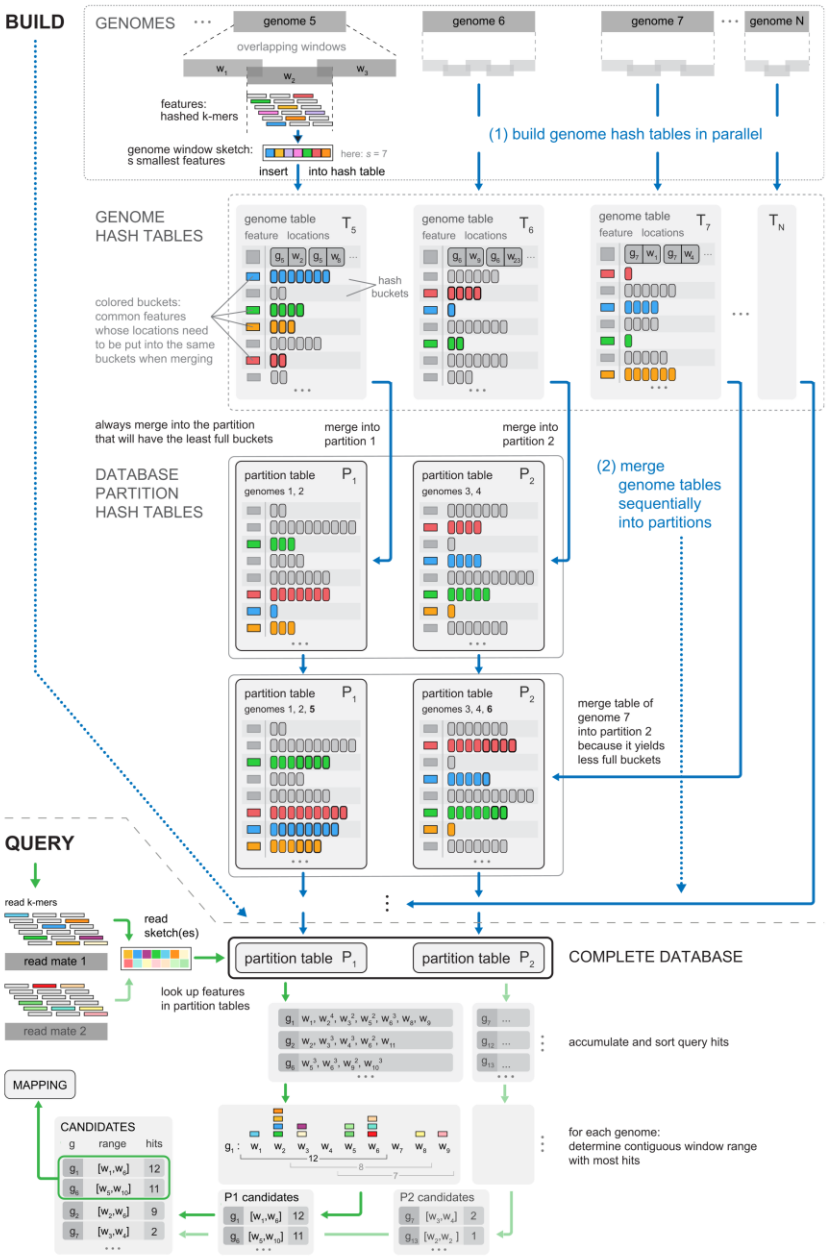

Fig. 1 Top (build phase): Database building scheme. First, hash tables T_i are built in parallel for each reference genome i. Then, they are sequentially merged into the final database partitions P_j while minimizing overflowing hash buckets. Bottom (query phase): Querying pipeline. Features of each read (pair) are queried against each database partition and the resulting top candidates are merged to obtain the classification result.

memory limit for partition P_{opt}, the table T_i is not inserted but used to start a new partition instead.

Bucket truncation that could lead to a loss of information due to many k-mer collisions can thus be reduced significantly in our new build approach if two or more partitions are used. Moreover, it is not necessary to manually partition reference genomes in order to obtain size-limited partitions. Note that, as the partitioning works on a reference sequence level and k-mer distributions might differ between input sequences the resulting partitions might vary in size.

The parallel construction of individual sequence tables in AFS-MetaCache2 also significantly speeds up database construction by around one order-of-magnitude compared to prior methods such as AFS-MetaCache1, Kraken2, and KrakenUniq which are limited to a single thread operating a single hash table during the build phase. Note that merging incurs a small additional memory overhead since not yet merged individual tables T_i have to be kept alongside the partition tables which itself have to be kept large enough to accommodate new insertions, while not exceeding the maximum hash table load factor.

Query Phase

To classify a read, its sequence is first split into windows of the same length as used in the database. From each window all canonical k-mers are generated and minhashing is applied to produce a sketch. All elements of the sketch are then queried against the hash table(s). The resulting location lists are merged and identical locations are accumulated. This yields a (sparse) histogram of hit counts per window in the reference genomes (window count statistic) which indicates the similarity of this region with the read. To account for single-end or paired-end Illumina reads spanning multiple windows, the window count statistic is scanned with a sliding window approach to find target regions with the highest aggregated hit counts in a contiguous window range. The top m counts (top hits) are then used to classify the read. In case of multiple partitions, each partition needs to be queried separately and the top-hit results need to be merged accordingly. If the difference of the highest and second highest count is above a threshold, the read is labeled as belonging to the taxon of the genome corresponding to the maximum count. Otherwise, all targets with counts close to the maximum are considered, the lowest common ancestor (LCA) of the corresponding taxa is calculated and used to label the read. Classifying ONT reads is identical to classifying (single-end) Illumina reads with the difference that ONT reads are typically longer and cover more windows.

Following the per-read classification, we perform *quantification* by estimating the abundances of organisms contained in a dataset at a specific taxonomical rank. For each taxon which occurs in the dataset we count the number of reads assigned to it. We then build a taxonomic tree containing all found taxa. Taxa on lower levels than the requested taxonomic rank are pruned and their read counts are added to their respective parents, while reads from taxa on higher levels are distributed among their children in proportion to the weights of the sub-trees rooted at each child. After the redistribution the estimated number of reads and abundance percentages are returned as outputs.

Table 1 Meat composition of our utilized calibrator sausages.

Name	Cattle	Sheep	Pig	Horse	Chicken	Turkey
KGefLyo_A	0.5%	0.0%	5.5%	0.0%	14.0%	80.0%
KGefLyo_B	2.0%	0.0%	4.0%	0.0%	36.0%	58.0%
KGefLyo_C	4.0%	0.0%	2.0%	0.0%	58.0%	36.0%
KGefLyo_D	5.5%	0.0%	0.5%	0.0%	80.0%	14.0%
KLyo_A	14.0%	0.0%	80.0%	0.0%	0.5%	5.5%
KLyo_B	36.0%	0.0%	58.0%	0.0%	2.0%	4.0%
KLyo_C	58.0%	0.0%	36.0%	0.0%	4.0%	2.0%
KLyo_D	80.0%	0.0%	14.0%	0.0%	5.5%	0.5%
Kal_A	1.0%	9.0%	35.0%	55.0%	0.0%	0.0%
Kal_B	9.0%	1.0%	55.0%	35.0%	0.0%	0.0%
Kal_C	25.0%	25.0%	25.0%	25.0%	0.0%	0.0%
Kal_D	35.0%	55.0%	9.0%	1.0%	0.0%	0.0%
Kal_E	55.0%	35.0%	1.0%	9.0%	0.0%	0.0%

Table 2 Description of utilized read datasets for each sequenced calibrator sausage for both Illumina and ONT.

Dataset	Illumina			ONT		
	Number of Seq.	Seq. Lengths Mean	Max	Number of Seq.	Seq. Lengths Mean	Max
KGefLyo A	675 648	123	136	3 730 696	483	13 550
KGefLyo B	772 140	127	136	3 541 003	553	18 939
KGefLyo C	703 550	128	136	3 606 972	509	14 817
KGefLyo D	1 613 574	125	136	2 713 627	436	16 044
KLyo A	801 976	128	136	1 878 926	580	21 061
KLyo B	603 544	122	136	2 110 760	704	12 684
KLyo C	1 014 268	125	136	2 617 000	674	21 048
KLyo D	833 068	120	136	2 873 895	516	14 454
Kal A	1 659 486	126	136	2 844 028	387	6 138
Kal B	195 360	132	136	4 341 385	412	9 142
Kal C	808 574	131	136	3 306 965	372	8 094
Kal D	805 542	126	136	4 154 567	473	11 428
Kal E	577 428	130	136	3 750 967	411	9 507

Results

Datasets

Sequencing read datasets have been obtained from thirteen calibrator sausage samples (Kal A-E, KLyo A-D, KGeflLyo A-D) with known meat composition (admixtures of chicken, turkey, pork, beef, horse, and sheep) as shown in Table 1. Table 2 provides a summary of the corresponding sequencing read datasets for both Illumina and ONT in terms of size and read length distribution. Due to the skewed read length distribution of ONT, we removed all reads shorter than 200-bp from the corresponding datasets.

Food-related genomes (selection of main ingredients) used for database construction are listed in Table 3. We created databases of various sizes using the following reference genomes sets in order to compare quantification accuracy, runtime, and memory consumption of various tools:

AFS8: Genomes from 1 to 8 in Table 3.
AFS20: Genomes from 1 to 20 in Table 3.
AFS31: Genomes from 1 to 31 in Table 3.
AFS42: Genomes from 1 to 42 in Table 3.

To test the performance of AFS-MetaCache2 with partitioned databases, we built several version of AFS42 using one (AFS42), two (AFS42-P2), and four partitions (AFS42-P4).

Quantification accuracy

Read classification and abundance estimation was performed for all sequencing read datasets listed in Table 2 using AFS-MetaCache2, Kraken2 v2.1.5 in combination with Bracken v2.5.3, and KrakenUniq v1.0.4 – all using default settings. AFS-MetaCache2 and Kraken2+Bracken were run with all eukaryotic databases of various sizes. KrakenUniq was only run with the AFS8 database, because it was not able to successfully build larger databases on our test system with 512 GB main memory.

In order to assess the overall performance of a mapping tool in combination with a given reference genome set and sequencing technology we use two metrics:

Absolute deviations: We sum up the absolute deviations from the known meat composition percentages for each dataset and compute the averages and standard deviations of these deviation sums over all datasets.
False positives: Read assignments to taxa other than the known main meat components are presumed to be false positive classifications[1].

The results for absolute deviations are presented in Figure 2 which compares the average sums and mean deviations per dataset for AFS-MetaCache2, Kraken2+Bracken, and KrakenUniq. AFS-MetaCache2 yields consistently lower ground truth deviations than Kraken2+Bracken for all genome databases and both sequencing technologies. While KrakenUniq's abundance results for AFS8 is comparable to AFS-MetaCache2 for the same database, it is not able to reliably scale to bigger genome collections. Deviation sums obtained with AFS-MetaCache2 range from $(11.7 \pm 3.1)\%$ to $(13.8 \pm 2.6)\%$ for Illumina reads and from $(11.1 \pm 3.4)\%$ to $(12.1 \pm 2.4)\%$ for ONT reads, while the corresponding values for Kraken2+Bracken range from $(15.7 \pm 4.9)\%$ to $(16.3 \pm 4.8)\%$ for Illumina reads and from $(14.8 \pm 3.4)\%$ to $(16.3 \pm 4.6)\%$ for ONT reads. The deviation sums for KrakenUniq with the AFS8 database are $(12.7 \pm 3.1)\%$ for Illumina and $(12.6 \pm 2.8)\%$ for ONT respectively. Each tested tool thus achieves better average abundance accuracy for ONT reads than for Illumina reads.

The sums of false positive abundances per dataset and averaged over all datasets are shown in Figure 3. AFS-MetaCache2 outperforms all other tested tools for

[1] Note that some of these reads could potentially represent unknown contamination or other ingredients that are in fact correctly mapped.

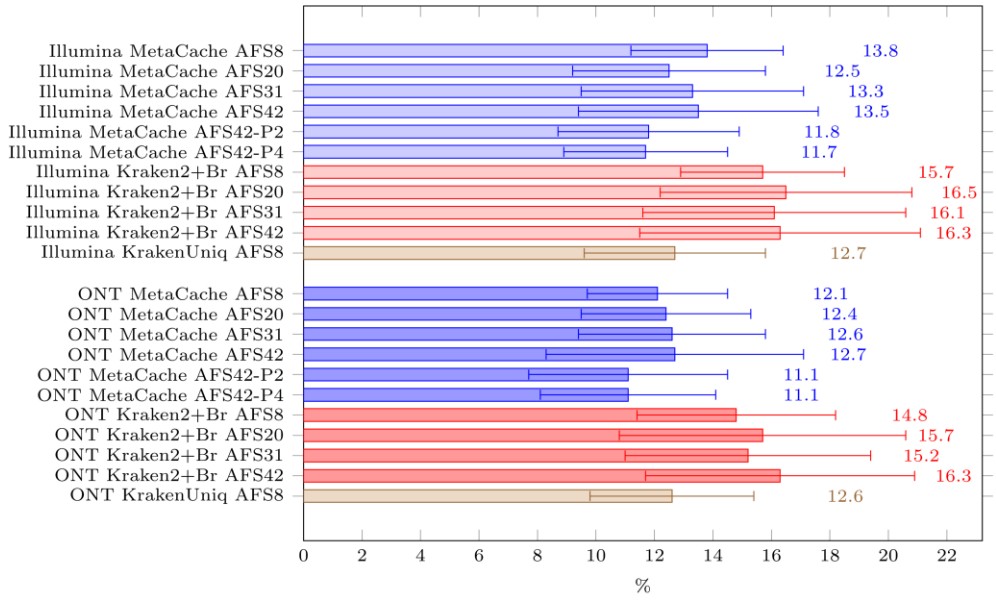

Fig. 2 Averages and standard deviations (over all 13 sausage datasets) of the per-dataset sum of the absolute abundance deviations from the ground truth.

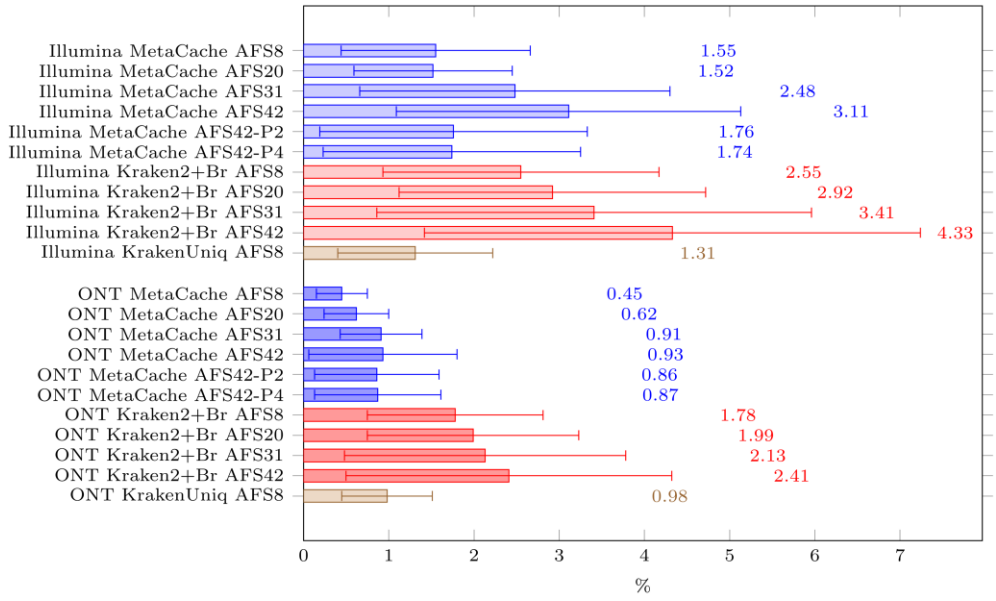

Fig. 3 Averages and standard deviations (over all 13 sausage datasets) of the per-dataset sum of presumed false positive abundances, i.e., all abundances of taxa other than the known main components.

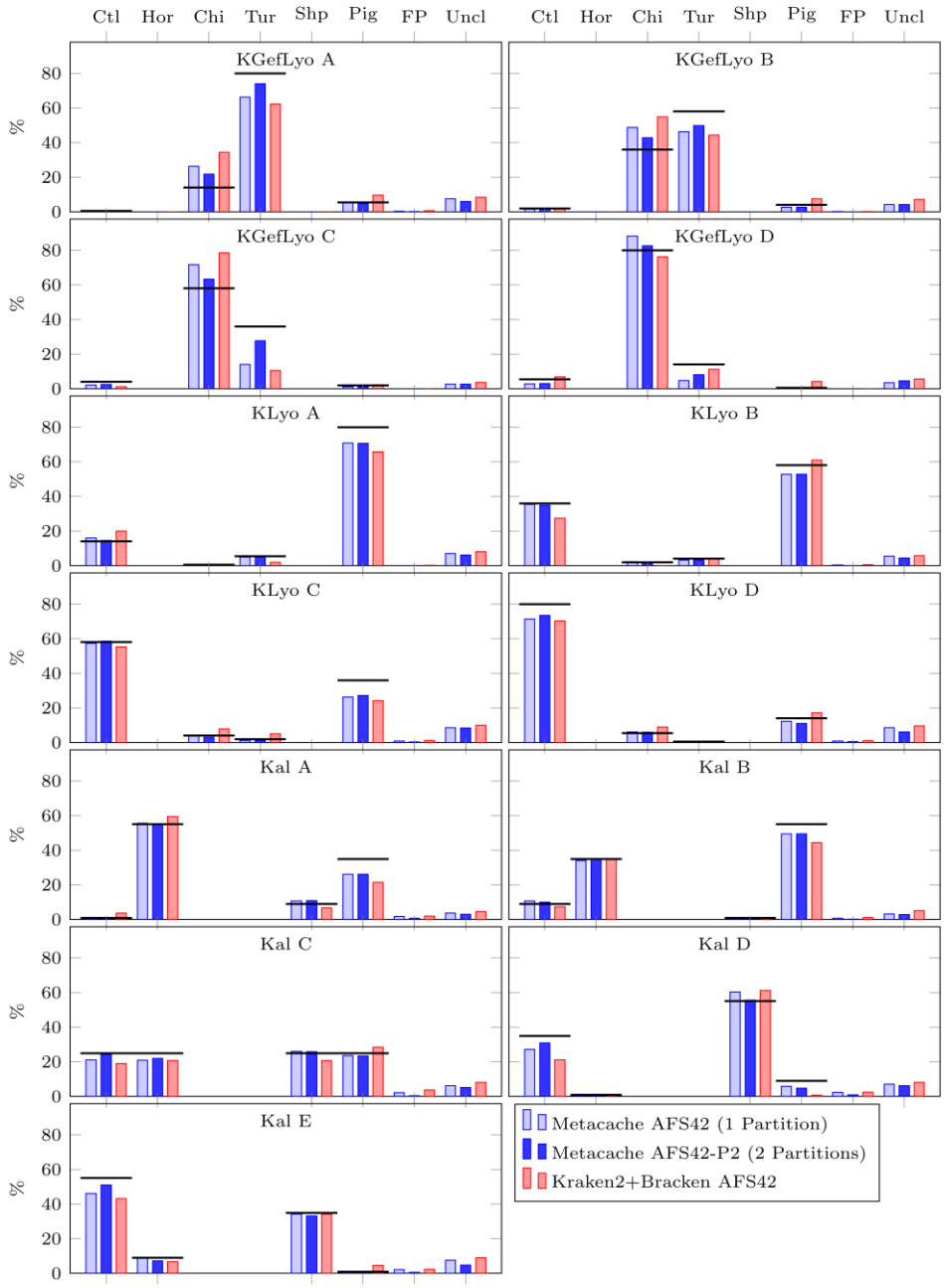

Fig. 4 Classified abundances for each ONT read dataset. Thick horizontal bars indicate ground truth abundances. Abbreviations: Ctl=Cattle, Hor=Horse, Chi=Chicken, Tur=Turkey, Shp=Sheep, Uncl=Unclassified, FP = presumed false positives (all taxa other than known main components).

Table 3 List of considered eukaryotic reference genomes

Item	Name	Accession ID	Size on disk
01	*Sus scrofa* (Pig)	GCF_000003025.6	2.4 GB
02	*Equus caballus* (Horse)	GCF_002863925.1	2.4 GB
03	*Meleagris gallopavo* (Turkey)	GCF_000146605.3	1.1 GB
04	*Mus musculus* (House Mouse)	GCF_000001635.26	2.7 GB
05	*Gallus gallus* (Chicken)	GCF_000002315.5	1.1 GB
06	*Ovis aries* (Sheep)	GCF_002742125.1	2.8 GB
07	*Rattus norvegicus* (Norway rat)	GCF_000001895.5	2.8 GB
08	*Bos taurus* (Cattle)	GCF_002263795.1	2.6 GB
09	*Bubalus bubalis* (Water buffalo)	GCF_003121395.1	2.6 GB
10	*Cervus elaphus hippelaphus* (Red deer)	GCA_002197005.1	3.3 GB
11	*Capreolus capreolus* (Western roe deer)	GCA_000751575.1	3.0 GB
12	*Struthio camelus australis* (African ostrich)	GCF_000698965.1	1.2 GB
13	*Anas platyrhynchos* (Mallard)	GCF_003850225.1	1.1 GB
14	*Capra hircus* (Goat)	GCF_001704415.1	2.8 GB
15	*Oryctolagus cuniculus* (Rabbit)	GCF_000003625.3	2.6 GB
16	*Cavia aperea* (Ginea pig)	GCA_000688575.1	2.6 GB
17	*Camelus ferus* (Wild batrian camel)	GCF_009834535.1	2.0 GB
18	*Canis lupus familiaris* (Dog)	GCF_000002285.3	2.3 GB
19	*Felis catus* (Domestic Cat)	GCF_000181335.3	2.4 GB
20	*Homo sapiens* (Human)	GCF_000001405.39	3.1 GB
21	*Equus asinus* (Donkey)	GCF_001305755.1	2.3 GB
22	*Rangifer tarandus* (Reindeer)	GCA_004026565.1	2.9 GB
23	*Phasanius colchicus* (Ring necked pheasant)	GCF_004143745.1	989 MB
24	*Glycine max* (Soybean)	GCF_000004515.5	946 MB
25	*Zea mays* (Maize)	GCF_902167145.1	2.1 GB
26	*Triticum aestivum* (Bread wheat)	GCA_002220415.3	15 GB
27	*Secale cereale* (Rye)	GCA_900079665.1	1.8 GB
28	*Hordeum vulgare* (Barley)	GCA_903813605.1	4.1 GB
29	*Oryza sativa japonica* (Rice)	GCF_000005425.2	370 MB
30	*Arachis hypogaea* (Peanut)	GCF_003086295.2	2.5 GB
31	*Saccharomyces cerevisiae* (Bakers yeast)	GCF_000146045.2	12 MB
32	*Notamacropus eugenii* (Tammar wallaby)	GCA_000004035.1	3.0 GB
33	*Brassica nigra* (Black mustard)	GCA_001682895.1	389 MB
34	*Brassica juncea* (Brown mustard)	GCA_001687265.1	924 MB
35	*Apium graveolens* (Celery)	GCA_009905375.1	3.2 GB
36	*Corylus avellana* (Hazelnut)	GCA_901000735.1	515 MB
37	*Sesamum indicum* (Sesame)	GCF_000512975.1	268 MB
38	*Juglans regia* (Walnut)	GCF_001411555.2	557 MB
39	*Pistacia vera* (Pistachio)	GCF_008641045.1	649 MB
40	*Prunus dulcis* (Almond)	GCF_902201215.1	220 MB
41	*Lupinus albus* (White lupine)	GCA_010261695.1	540 MB
42	*Lupinus angustifolius* (Blue lupine)	GCF_001865875.1	591 MB
Total			89 GB

ONT. For Illumina AFS-MetaCache2 achieves lower average false positive sums than Kraken2+Bracken for each genome database. For the small AFS8 database, KrakenUniq achieves a slightly lower rate than AFS-MetaCache2 for the same database. Overall, false positive sums are consistently and up to 3.4 times lower for ONT datasets compared to their corresponding Illumina counterparts over all tool/database combinations.

The results in figures 2 and 3 also show that the new database partitioning introduced with AFS-MetaCache2 is effective: For all tests the accuracy using AFS42-P2 (AFS42-P4) clearly improve in comparison to AFS42: Average absolute abundance deviations are reduced from 13.5% to 11.8% (11.7%) for Illumina and from 12.7% to 11.1% (11.1%) for ONT, while average false positives are reduced from 3.11% to 1.76% (1.74%) for Illumina and from 0.93% to 0.86% (0.87%) for ONT.

The relatively high variance observed in dataset-average abundance deviations can be better understood by looking at the breakdown of abundance results of the main meat components in relation to their ground truth values for each individual ONT dataset as shown in Figure 4. A large part of the variance is driven by the abundance deviations for chicken and turkey in the KGefLyo datasets where the abundance of chicken is in most cases overestimated, while that of turkey is often underestimated by Kraken2+Bracken and also (to a lesser extent) by AFS-MetaCache2. Also, both tools underestimate the percentage of pork more when the overall fraction of pork in a sample is large.

Table 6 shows a more detailed comparison of the relative abundances obtained with AFS-MetaCache2 using reference database AFS42-P2 for both Illumina and ONT reads for all taxa whose abundance was at least 0.1% of all classified reads in a dataset. Illumina reads assigned to other animal genomes than the main components account for 0.23% to 8.47% per dataset. They were mainly identified as goat, donkey, deer and to a lesser extent ($\leq 0.2\%$) as water buffalo or guinea pig. The ONT mappings show significantly lower abundances of animals other than the main components with percentages ranging from 0.25% to 2.3% per dataset while the animal species were mostly the same except that less than 0.01% were assigned to guinea pig and additionally 0.12% were identified as pheasant.

For both sequencing technologies, the percentage of reads mapped to any of the plant genomes in the database ranged from 0.1% to 0.7% per dataset. Plants identified with abundances greater than 0.1% included wheat, barley, celery, mustard and sesame. The percentage of unclassified reads ranged from 2.65% to 9.33% per dataset for Illumina samples and 2.63% to 8.68% for ONT samples.

Runtime and memory consumption

All database builds and experiments were run on a workstation with an AMD EPYC 7713P 64-core processor running Linux, 512 GB of RAM, and a PCI NVMe SSD for storing the read files and genome databases. Build times and memory consumption for the database construction used for the accuracy tests for all classification tools are shown in Table 4. AFS-MetaCache2 consumes less than half the amount of memory than Kraken2+Bracken for building databases. In terms of runtime it takes only a few minutes to build each database and is around one order-of-magnitude faster than Kraken2+Bracken. KrakenUniq build times are significantly longer. It takes more than 6 hours to build the smallest 8 genome database; we were also not able to successfully build larger databases with KrakenUniq.

We also compared the runtimes of our new database building scheme to the sequential approach used in AFS-MetaCache1. The new scheme is significantly faster with

Table 4 Build times and memory consumption for the databases used for accuracy evaluation.

Tool	Name	Partitions (Size on Disk)	Peak Memory Consumption	Build Time
AFS-MetaCache2	AFS8	1 (18 GB)	26 GB	1 min 31 s
	AFS20	1 (42 GB)	64 GB	3 min 43 s
	AFS31	1 (60 GB)	92 GB	5 min 16 s
	AFS42	1 (66 GB)	99 GB	7 min 21 s
	AFS42-P2	2 (40 GB+33 GB)	85 GB	8 min 10 s
	AFS42-P4	4 (20 GB+15 GB+23 GB+17 GB)	63 GB	9 min 53 s
Kraken2 +Bracken	AFS8	1 (39 GB)	67 GB	13 min
	AFS20	1 (94 GB)	156 GB	30 min
	AFS31	1 (122 GB)	195 GB	49 min
	AFS42	1 (153 GB)	247 GB	71 min
KrakenUniq	AFS8	1 (85 GB)	124 GB	381 min
	AFS20	did not finish		

speedups between 6.9× and 5.9×; e.g. for AFS42 the database built time is reduced from 51 minutes to 7 minutes 21 seconds.

Querying speed and memory consumption for processing ONT read datasets are shown in Table 5. Depending on the reference genome set, AFS-MetaCache2 processes between 18.9 and 29.3 million reads per minute with a memory consumption between 22 GB and 67 GB and an average read length of (501 ± 104)bp. Using the same genome sets, Kraken2+Bracken is faster (between 28.1 and 41.1 million reads per minute) but consumes more memory (between 39 GB and 153 GB). KrakenUniq is again slowest (7.3 million reads per minute). AFS-MetaCache2 classifies Illumina reads at speeds of 113 to 143 million reads per minute (depending on the database) and Kraken2+Bracken process Illumina reads at 151 to 197 million reads per minute.

AFS-MetaCache2's support for partitioned reference databases has the advantage of consuming less memory and can also speed up querying since the window count statistics of smaller partitions can be processed much faster. The reduction in peak memory consumption for a single partition compared to the full database corresponds to the ratio of the partition size to the size of the full database, so querying AFS42-P4 takes at maximum 29 GB of memory for the largest partition with size 23 GB instead of 71 GB for the full AFS42 database with a size of 66 GB. Querying speed for ONT (Illumina) reads increases from 18.9 (113.1) million reads per minute for AFS42 to 24.2 (142.3) million reads for AFS42-P4.

Discussion

Determination and quantification of food ingredients are important issues in food bio-surveillance [8]. The potential presence of a large variety of food components establishes the need for a broad-scale screening method that allows for precise identification and quantification of ingredients, ideally spanning various kingdoms of life including plants, animals, fungi, and bacteria. Established technologies for analyzing foodstuff

Table 5 Average querying speed in million reads per minute and memory consumption for classifying the ONT and Illumina read datasets. Note that memory consumption is mainly dominated by the size of the loaded database.

Tool	Database	Peak Memory Consumption	Query Speed ONT	Illumina
AFS-MetaCache2	AFS8	22 GB	29.3 MR/min	143.8 MR/min
	AFS20	50 GB	21.4 MR/min	122.7 MR/min
	AFS31	67 GB	19.1 MR/min	113.9 MR/min
	AFS42	71 GB	18.9 MR/min	113.1 MR/min
	AFS42-P2[1]	51 GB	20.5 MR/min	121.5 MR/min
	AFS42-P4[1]	29 GB	24.2 MR/min	142.3 MR/min
Kraken2+Bracken	AFS8	39 GB	42.1 MR/min	197.0 MR/min
	AFS20	94 GB	34.3 MR/min	175.6 MR/min
	AFS31	122 GB	29.7 MR/min	154.6 MR/min
	AFS42	153 GB	28.1 MR/min	151.2 MR/min
KrakenUniq	AFS8	85 GB	7.3 MR/min	61.8 MR/min

[1]based on total runtime for sequential querying of partitions and merging of results

such as qPCR/ddPCR and (Meta-)Barcoding are typically limited to a set of target species within a single assay that need to be defined beforehand by the use of primers and can lead to the failure to detect certain species [26, 27]. Deep sequencing of total genomic DNA from biological samples followed by bioinformatic analyses based on comparisons to available reference genomes can overcome this limitation.

In this study we have applied our approach to a set of real-world reference samples, containing admixtures of food-relevant species (chicken, turkey, pork, beef, horse, sheep). The results demonstrate that AFS-MetaCache2 is able to reliably detect the main components. In comparison to the established metagenomics tools Kraken2+Bracken and KrakenUniq for abundance estimation, AFS-MetaCache2 is superior in terms of absolute deviation and false positive rates. As different types of tissue can contain different concentrations of DNA (matrix effect), deviations could possibly be further reduced by a subsequent normalization procedure that takes tissue ratios into account [12].

Our results further show that long read sequencing technologies like ONT can yield more accurate abundance results and less false positive mappings for foodstuff analysis compared to short read sequencing methods such as Illumina. This can be attributed to the increased read length which provides improved genomic continuity of which all tested tools can take advantage of to yield better read assignments even at the cost of higher sequencing error rates[2].

The ONT datasets used in our study not only provided longer read length than Illumina but also contained more reads per sample. Therefore, they feature a much higher genomic coverage. To investigate the impact of coverage on accuracy we produced randomly downsampled versions of each ONT and Illumina dataset with 50% and 10% of reads remaining. The mean deviations of the abundance results were at most 0.9% for

[2]Using read mapping we estimated the error rate of the used ONT data set as ≈2.5% and of the Illumina data sets as ≈0.1%

Illumina and at most 0.6% for ONT when comparing any of the down-sampled sets to their complete counterparts.

Another important aspect of the metagenomic approach is efficient scalability to large-scale databases containing complex eukaryotic reference genome indices. The new database partitioning scheme introduced with AFS-MetaCache2 leads to faster build times compared to prior versions as well as Kraken2+Bracken, and KrakenUniq, while consuming less main memory, especially during the query phase. An important parameter in our new scheme is the number of database partitions. Our results show that working with more partitions has two distinct advantages: (i) accuracy can be increased compared to using a single partition since our new method automatically separates reference genomes with a high k-mer feature overlap into different database partitions, which reduces the number of overflowing buckets and in turn improves overall accuracy; (ii) using more partitions also reduces main memory consumption.

In the face of ever growing genomic databases this is particularly important; e.g. analyzing complex food matrices might require hundreds or even thousands of large reference genomes for which an efficient and scalable database construction scheme is crucial. Even though our new construction scheme provides high efficiency and scalability, using several thousands of complex genomes would still require significant runtimes on common multi-core CPUs. A possible approach to accelerate this time-consuming task would be the usage of modern GPU accelerators. In prior work [28], we have already shown how this could be done with earlier versions of AFS-MetaCache2. However, our new database construction introduced with AFS-MetaCache2 is more complex. Accelerating it on multi-GPU systems will thus be an interesting direction of future research and will be part of our future work.

Conclusion

We have presented AFS-MetaCache2, a fast and precise screening and quantification tool for whole genome shotgun sequencing-based biosurveillance applications such as food testing with a corresponing publicly available implementation. It can scale efficiently towards large-scale reference databases containing complex eukaryotic genomes making it suitable for broad metagenomic screening applications. Our new database partitioning scheme leads to faster build times as well as more accurate abundance results and less false positive mappings compared to previous versions of AFS-MetaCache2, Kraken2+Bracken and KrakenUniq when running with the same reference genome sets. It also allows our approach to be run on memory-constrained systems by sequentially querying smaller partitions followed by a fast results-merging phase. Evaluation results further show that ONT sequencing technology can achieve higher accuracy compared to Illumina short read sequencing.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

AFS-MetaCache2 is available at https://github/muellan/metacache; a dedicated AFS manual page can be found at https://github.com/muellan/metacache/blob/master/docs/afs.md. The utilized sequencing read datasets have been submitted to ENA project PRJEB34001.

Competing interests

The authors declare that they have no competing interests.

Funding

Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 439669440 TRR319 RMaP TP 01.

Authors' contributions

AM implemented and tested the software and performed the experiments. SLH conducted the sequencing experiments. BS, AM, SLH and TH wrote the draft of the manuscript. BS and TH proposed and supervised the project. AM, AW, FK, SLH, TH and BS analyzed and discussed the results. AM, BS, SLH and TH edited the manuscript. The authors read and approved the final manuscript.

Acknowledgments

This work was partially funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 439669440 TRR319 RMaP TP C01.

References

[1] Köppel, R., Ruf, J., Rentsch, J.: Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, horse and sheep. European Food Research and Technology **232**, 151–155 (2011) https://doi.org/10.1007/s00217-010-1371-y

[2] Singh, M., Young, R.G., Hellberg, R.S., Hanner, R.H., Corradini, M.G., Farber, J.M.: Twenty-three years of PCR-based seafood authentication assay development: What have we learned? Comprehensive Reviews in Food Science and Food Safety **23**, 13401 (2024) https://doi.org/10.1111/1541-4337.13401

[3] Blaxter, M., Lewin, H.A., DiPalma, F., Challis, R., Silva, M., Durbin, R., Formenti, G., Franz, N., Guigo, R., Harrison, P.W., Hiller, M., Hoff, K.J., Howe, K., Jarvis, E.D., Lawniczak, M.K.N., Lindblad-Toh, K., Mathews, D.J.H., Martin, F.J., Mazzoni, C.J., McCartney, A.M., Mulder, N., Paez, S., Pruitt, K.D.,

Ras, V., Ryder, O.A., Shirley, L., Thibaud-Nissen, F., Warnow, T., Waterhouse, R.M., Hajibabaei, M., Wong, G.K.-S., Burke, T., HarrisLewin, H.A.L., M Silva, Mkn, L., Djh, M., Silva: The earth biogenome project phase II: illuminating the eukaryotic tree of life. Frontiers in Science **3**, 1514835 (2025) https://doi.org/10.3389/FSCI.2025.1514835

[4] Köppel, R., Ganeshan, A., Weber, S., Pietsch, K., Graf, C., Hochegger, R., Griffiths, K., Burkhardt, S.: Duplex digital pcr for the determination of meat proportions of sausages containing meat from chicken, turkey, horse, cow, pig and sheep. European Food Research and Technology **245**, 853–862 (2019) https://doi.org/10.1007/s00217-018-3220-3

[5] Vishnuraj, M.R., Devatkal, S., Vaithiyanathan, S., Kumar, R.U., Srinivas, C., Mendiratta, S.K.: Detection of giblets in chicken meat products using microRNA markers and droplet digital PCR assay. LWT **140**, 110798 (2021) https://doi.org/10.1016/J.LWT.2020.110798

[6] Ma, X.Y., Shao, Z.L., Yu, X.P., Wang, Z.L.: A droplet digital pcr-based approach for quantitative analysis of the adulteration of atlantic salmon with rainbow trout. Foods 2023, Vol. 12, Page 4309 **12**, 4309 (2023) https://doi.org/10.3390/FOODS12234309

[7] Dobrovolny, S., Uhlig, S., Frost, K., Schlierf, A., Nichani, K., Simon, K., Cichnamarkl, M., Hochegger, R.: Interlaboratory validation of a DNA metabarcoding assay for mammalian and poultry species to detect food adulteration. Foods **11**, 1108 (2022) https://doi.org/10.3390/FOODS11081108/S1

[8] Kappel, K., Gadelmeier, A., Denay, G., Gerdes, L., Graff, A., Hagen, M., Hassel, M., Huber, I., Näumann, G., Pavlovic, M., Pietsch, K., Stumme, B., Völkel, I., Westerdorf, S., Wöhlke, A., Hochegger, R., Brinks, E., Franz, C., Haase: Detection of adulterated meat products by a next-generation sequencing-based metabarcoding analysis within the framework of the operation OPSON X: a cooperative project of the German National Reference Centre for Authentic Food (NRZ-Authent) and the competent German food control authorities. Journal für Verbraucherschutz und Lebensmittelsicherheit **18**, 375–391 (2023) https://doi.org/10.1007/s00003-023-01437-w

[9] Gorini, T., Mezzasalma, V., Deligia, M., Mattia, F.D., Campone, L., Labra, M., Frigerio, J.: Check your shopping cart: DNA barcoding and mini-barcoding for food authentication. Foods **12**, 2392 (2023) https://doi.org/10.3390/FOODS12122392

[10] Andronache, J., Cichna-Markl, M., Dobrovolny, S., Hochegger, R.: Development of a DNA metabarcoding method for the identification of crustaceans (malacostraca) and cephalopods (coleoidea) in processed foods. Foods **14**, 1549 (2025) https://doi.org/10.3390/FOODS14091549/S1

[11] Ripp, F., Krombholz, C., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., Hankeln, T.: All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. BMC Genomics **15**, 639 (2014) https://doi.org/10.1186/1471-2164-15-639

[12] Hellmann, S.L., Ripp, F., Bikar, S.E., Schmidt, B., Köppel, R., Hankeln, T.: Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq). European Food Research and Technology **246**, 193–200 (2020) https://doi.org/10.1007/s00217-019-03404-y

[13] Kobus, R., Abuín, J.M., Müller, A., Hellmann, S.L., Pichel, J.C., Pena, T.F., Hildebrandt, A., Hankeln, T., Schmidt, B.: A big data approach to metagenomics for all-food-sequencing. BMC Bioinformatics **21**(1), 102 (2020)

[14] Kutuzova, S., Nielsen, M., Piera, P., Nissen, J.N., Rasmussen, S.: Taxometer: Improving taxonomic classification of metagenomics contigs. Nature Communications **15**(1), 8357 (2024)

[15] Kim, C., Pongpanich, M., Porntaveetus, T.: Unraveling metagenomics through long-read sequencing: A comprehensive review. Journal of Translational Medicine **22**(1), 111 (2024)

[16] Portik, D.M., Brown, C.T., Pierce-Ward, N.T.: Evaluation of taxonomic classification and profiling methods for long-read shotgun metagenomic sequencing datasets. BMC Bioinformatics **23**(1), 541 (2022)

[17] Govender, K.N., Eyre, D.W.: Benchmarking taxonomic classifiers with Illumina and nanopore sequence data for clinical metagenomic diagnostic applications. Microbial Genomics **8**(10), 000886 (2022)

[18] Van Uffelen, A., Posadas, A., Roosens, N.H.C., Marchal, K., De Keersmaecker, S.C.J., Kevin, V.: Benchmarking bacterial taxonomic classification using nanopore metagenomics data of several mock communities. Scientific Data **11**(1) (2024) https://doi.org/10.1038/s41597-024-03672-8

[19] Gehrig, J.L., Portik, D.M., Driscoll, M.D., Jackson, E., Chakraborty, S., Gratalo, D., Ashby, M., Valladares, R.: Finding the right fit: evaluation of short-read and long-read sequencing approaches to maximize the utility of clinical microbiome data. Microbial Genomics **8**, 000794 (2022)

[20] Macip, G., Soler-Comas, A., Palomeque, A., Motos, A., Llonch, B., Canseco-Ribas, J., Bueno-Freire, L., Calabretta, D., Kiarostami, K., Cabrera, R., *et al.*: Comparative analysis of illumina and oxford nanopore sequencing platforms for 16S rRNA profiling of respiratory microbial communities. Scientific Reports **15**(1), 33688 (2025)

[21] Wood, D.E., Lu, J., Langmead, B.: Improved metagenomic analysis with kraken

2. Genome biology **20**(1), 257 (2019)

[22] Lu, J., Breitwieser, F.P., Thielen, P., Salzberg, S.L.: Bracken: estimating species abundance in metagenomics data. PeerJ Computer Science **3**, 104 (2017)

[23] Breitwieser, F.P., Baker, D.N., Salzberg, S.L.: Krakenuniq: confident and fast metagenomics classification using unique k-mer counts. Genome biology **19**(1), 198 (2018)

[24] Pockrandt, C., Zimin, A.V., Salzberg, S.L.: Metagenomic classification with krakenuniq on low-memory computers. Journal of open source software **7**(80), 4908 (2022)

[25] Broder, A.Z.: On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), pp. 21–29 (1997). IEEE

[26] Macher, T.H., Schütz, R., Yildiz, A., Beermann, A.J., Leese, F.: Evaluating five primer pairs for environmental DNA metabarcoding of central european fish species based on mock communities. Metabarcoding and Metagenomics 7: e103856 **7**, 103856 (2023) https://doi.org/10.3897/MBMG.7.103856

[27] Preckel, L., Brünen-Nieweler, C., Denay, G., Petersen, H., Cichna-Markl, M., Dobrovolny, S., Hochegger, R.: Identification of mammalian and poultry species in food and pet food samples using 16s rDNA metabarcoding. Foods **10** (2021) https://doi.org/10.3390/FOODS10112875

[28] Kobus, R., Müller, A., Jünger, D., Hundt, C., Schmidt, B.: Metacache-gpu: ultra-fast metagenomic classification. In: Proceedings of the 50th International Conference on Parallel Processing, pp. 1–11 (2021)

Table 6 Relative abundances obtained with AFS-MetaCache2 using reference database AFS42-P2

	Dataset	Main Components					Animals					Plants	Uncl.	Σ FP[9]
	& Sequ.Tech.	Cattle	Horse	Pig	Sheep	Chicken	Turkey	W.Buf.	Goat	Donkey	Deer[1]	Other			
Illumina	KGefLyo A	0.73%		6.22%		29.93%	56.99%				0.23%	0.18%[2]	0.10%[4]	5.18%	0.51%
	KGefLyo B	2.02%		3.95%		55.38%	32.19%				0.24%	0.11%[2]		5.69%	0.35%
	KGefLyo C	3.20%		1.54%		74.48%	16.87%				0.23%			3.34%	0.23%
	KGefLyo D	3.91%		0.42%		85.61%	5.57%				0.29%			3.76%	0.29%
	KLyo A	16.27%		70.37%		0.66%	4.68%				1.29%			6.25%	1.29%
	KLyo B	35.91%		49.52%	0.18%	2.21%	2.94%	0.10%			2.76%			6.01%	2.86%
	KLyo C	54.67%		25.97%	0.23%	3.75%	1.13%	0.13%	0.13%		4.24%	0.11%[2]		9.33%	4.62%
	KLyo D	70.17%		10.39%	0.34%	4.76%	0.32%	0.20%	0.18%		4.89%			8.52%	5.27%
	Kal A	1.21%	51.98%	29.86%	9.56%				0.91%	2.14%	0.95%		0.53%[6]	2.65%	3.99%
	Kal B	9.82%	34.70%	48.33%	1.28%				0.14%	1.21%	1.30%		0.17%[6]	2.72%	2.65%
	Kal C	22.23%	22.28%	18.63%	25.31%				2.58%	0.79%	3.80%		0.21%[7]	3.90%	7.37%
	Kal D	31.49%	1.10%	6.87%	43.95%				3.41%		4.70%			6.09%	8.11%
	Kal E	47.75%	7.51%	0.94%	28.26%				3.12%	0.27%	4.90%			5.97%	8.28%
Oxford Nanopore	KGefLyo A	0.61%		5.63%		31.35%	54.24%				0.13%	0.12%[3]	0.16%[5]	7.49%	0.41%
	KGefLyo B	1.58%		2.64%		62.78%	28.35%				0.00%		0.12%[5]	4.17%	0.12%
	KGefLyo C	2.21%		1.08%		79.65%	14.10%				0.00%			2.63%	0.00%
	KGefLyo D	2.83%		0.34%		88.24%	4.73%				0.00%			3.56%	0.00%
	KLyo A	16.01%		70.76%		0.73%	4.98%				0.11%			7.01%	0.11%
	KLyo B	35.39%		52.78%		2.37%	3.22%		0.11%		0.22%			5.50%	0.33%
	KLyo C	57.35%		27.36%		4.13%	1.28%		0.19%		0.56%		0.11%[5]	8.59%	0.85%
	KLyo D	71.32%		12.26%		6.11%	0.40%	0.11%	0.22%		0.60%			8.68%	0.94%
	Kal A	0.89%	55.66%	26.19%	10.67%				0.30%	1.44%	0.00%		0.70%[6]	3.77%	1.74%
	Kal B	10.78%	34.04%	49.52%	0.97%					0.80%	0.00%		0.20%[6]	3.27%	0.80%
	Kal C	21.11%	20.89%	23.47%	25.92%				0.82%	0.55%	0.37%		0.34%[7]	6.13%	2.08%
	Kal D	20.18%	0.92%	5.84%	63.25%				1.79%		0.50%			7.04%	2.29%
	Kal E	46.12%	8.87%	0.74%	34.16%				1.14%	0.22%	0.65%		0.14%[8]	7.67%	2.16%

[1]sum of read percentages classified as Roe Deer, Red Deer or Reindeer

[2]percentage of reads classified as Guinea Pig

[3]percentage of reads classified as Pheasant

[4]percentage of reads classified as Wheat

[5]percentage of reads classified as Barley

[6]percentage of reads classified as Celery

[7]percentage of reads classified as Mustard

[8]percentage of reads classified as Sesame

[9]presumed false positives (sum of abundances of all taxa other than known main components)

3 Discussion and future directions

3.1 Preparing AFS for official surveillance: achievements in method design and evaluation

In recent years, the field of foodomics has increasingly recognized the importance of accurate and comprehensive species identification to ensure food safety, prevent fraud, and comply with stringent regulatory standards. The rapid advancement of high-throughput DNA sequencing technologies has revolutionized the ability to assess food authenticity and safety (Billington et al., 2022; Haiminen et al., 2019; Haynes et al., 2019; Imanian et al., 2022; Ripp et al., 2014). The All-Food-Seq (AFS) method provides a powerful, PCR-free, whole-genome metagenomic approach that overcomes many of the inherent limitations of traditional targeted assays (Liu et al., 2017; Ripp et al., 2014). By enabling the unbiased detection and quantification of species across complex biological matrices, AFS offers significant improvements in sensitivity, accuracy, and scope of application for food screening and fraud detection.

This thesis has presented a series of studies that apply the AFS approach to diverse sample types, including multiple calibration datasets, real meat and seafood products, and processed foods, with the overarching goal of validating and refining the method for practical use in food safety and quality assurance. The calibrated datasets provided a controlled environment to evaluate the fundamental quantitative precision of AFS, demonstrating its capacity to achieve absolute deviations and to reliably detect even trace amounts of unexpected species.

Chapter 2.1 and **chapter 2.4** describe an evaluation of the potential of AFS as a robust tool for identifying and quantifying food ingredients. The method was tested for detecting and quantifying species composition proportion using calibration samples, both for sausage- and sea food-style food matrices. In

addition, the method was applied to "real-world" consumer samples by analysing both Doner Kebab-style foods (Hellmann, Ripp, et al., 2020) and processed sea foods (Chapter 2.4).

A review of the method with its advantages and challenges can be found in **chapter 2.2**. This section was published at The Bavarian Health and Food Safety Authority (*Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit*, LGL) to directly address German food control authorities, and hence it is in German language (Hellmann, Kobus, et al., 2020).

An algorithmic enhancement of the AFS pipeline is described in the **chapter 2.3**: The introduction of a computationally efficient, alignment-free k-mer-based approach called MetaCache-AFS represents a significant improvement in metagenomic analysis for food testing and bio-surveillance. This approach makes AFS faster and more scalable, while retaining accuracy and precision (Kobus et al., 2020).

Chapter 2.5 advances AFS to a third-generation sequencing platform. Performing Oxford Nanopore Technologies (ONT) long read sequencing on the established calibration sausages reveals a significant increase in accuracy compared to short read NGS technology (Müller et al., 2025).

3.1.1 Quantitative performance: accuracy, sensitivity, and comparison to qPCR and ddPCR

Building on the initial calibration sausage experiments presented by Ripp et al., 2014, the quantitative evaluation of AFS was systematically extended to additional taxonomic groups and more complex matrices, using an expanded series of reference sausages prepared in the same controlled manner. In the original work, only two mixed-meat

calibration sausages containing defined proportions of several mammalian species were used to demonstrate that read counting on WGS data can recover gravimetric proportions with good precision, with discrimination down to roughly 0.5 - 1 %.

In subsequent calibration sausage studies, all reference mixtures that had previously been used to benchmark qPCR and ddPCR (Kal, KGeflLyo, and Kylo; Eugster et al., 2009; Köppel et al., 2011, 2019) were re-analysed with AFS, allowing a direct method comparison on exactly the same materials. For each sausage, the deviation between measured and gravimetric species proportions was summarised as the total absolute percentage error across all components. AFS already produced errors in the same range as qPCR, and reliably quantified components across the full range of tested proportions, including the regulatory relevant low-percentage range of 1 %. After applying matrix-specific calibration functions, AFS showed the lowest total deviation from the ground truth in two-thirds of all sausages: in 8 of 13 recipes, calibrated AFS outperformed both qPCR and ddPCR, whereas ddPCR was clearly best in only one case and only marginally better in three further samples (Hellmann, Ripp, et al., 2020).

A second series of calibration mixtures, the FishCal samples, extended this evaluation to fish matrices, which introduced a different set of biases than those observed in meat sausages. Most species showed deviations that depended on the sample composition; however, salmon stood out as a clear exception, being consistently and markedly overestimated of up to eight-fold. This bias is explained by its unusually large genome relative to the other fish species and by the potential use of triploid aquaculture specimen, which are commonly applied in farming because triploids are sterile and show improved growth rates over wildtype diploid salmon (Murray et al., 2018). Accordingly, genome-size normalization followed by matrix-specific calibration was required to obtain realistic proportions. In FishCal, the mean total quantification error across mixtures decreased from roughly 45 % in the raw AFS output to about 20 % after genome-size correction and further to about 14 % after matrix calibration (Chapter 2.4), demonstrating that these systematic biases can be effectively mitigated.

Genome sizes among food items exhibit substantial variability, ranging from species with relatively small genomes, such as rice and fugu (387 and 391 Mb, respectively), to those with intermediate genome sizes, exemplified by chicken and cattle (1.2 and 3.5 Gb, respectively), and extending to species with very large genomes, such as wheat and onion (16.9 and 17.5 Gb, respectively; Gregory, 2025; Henniges et al., 2023). Larger genome sizes and higher DNA content per cell can lead to substantial overrepresentation in sequencing data, if read counts are not properly normalized to genome size and ploidy levels (see Chapter 3.2.1). This work demonstrates that c-values obtained from public databases like *Animal Genome Size Database* and *Plant DNA c-values Database* (Gregory, 2025; Henniges et al., 2023) can be used to mathematically correct for errors caused by genome size inequalities. For species where no c-value is available, a closely related species may be used as an approximation, although substantial differences can persist, particularly in lineages such as teleost fishes, where whole-genome duplication events and lineage-specific accumulation of transposable elements have generated pronounced variation in genome size (Volff, 2004). In such situations, genome size can instead be inferred directly from species-specific k-mer frequency profiles, i.e. from the characteristic distribution of short sequence motifs in the raw reads, which provides a data-driven estimate of genome size and thus circumvents the uncertainties associated with proxy C-values (Vurture et al., 2017), as described in Chapter 2.4.

Consequently, genome-size normalization is necessary to correct for major differences in DNA yield per gram of tissue. Across different food matrices and preparation protocols, our studies showed that AFS quantification reliably detects species at approximately 0.5 % proportion (Chapter 2.1, 2.4, 2.5). Applying genome-size corrections substantially reduces residual quantification error, supporting the robustness of read-count-based WGS quantification. These results indicate that, accounting for matrix-specific factors that equally impact all DNA-based methods, AFS achieves quantitative performance that is systematically superior to conventional qPCR

and broadly equivalent to ddPCR. In practical terms, AFS reaches the same order of accuracy and sensitivity as current digital PCR assays for calibrated matrices, but does not require multiple target-specific assays and prior assumptions about which taxa are present.

3.1.2 Integrated trace-level allergen detection and emerging spoilage signals

Beyond the detection of main components, the Kal A-E samples also contain trace-level spike-ins of 11 allergenic plant species. At proportions below 0.1 %, empirical data shows that quantitative estimates are often unreliable and may be difficult to distinguish from false-positive findings. However, the presence of the respective plant taxa could be identified qualitatively with high confidence (Hellmann, Ripp, et al., 2020; Ripp et al., 2014). From a regulatory perspective, a legally binding threshold for undeclared ingredients does not exist, so decisions must be made on a case-by-case basis (European Parliament & Council, 2011). However, ingredients present at low levels may be tolerated as technically unavoidable traces. In the case of a harmful or allergenic component, the qualitative detection by AFS alone can therefore serve as an early-warning signal that prompts follow-up analyses with complementary methods to verify the finding.

In addition to these allergenic spike-ins, AFS also revealed signals that are informative for emerging spoilage processes, AFS detected high read counts of the meat-spoilage bacterium *Brochothrix thermosphacta* together with its specific *phage BL3*, with an even genome coverage that is consistent with an actively growing population rather than spurious database hits (Kobus et al., 2020). *Brochothrix spp.* are psychrotrophic spoilage organisms of chilled meat and seafood, where their growth causes off-odours and quality loss (Illikoud et al., 2019), indicating an early microbiological signal of beginning

spoilage. In contrast to the present samples, where the detected phages are most likely of natural origin, bacteriophages targeting *B. thermosphacta* and other spoilage or pathogenic bacteria are increasingly utilized as biotechnology tools for food bio-preservation: in pork tissue, application of *B. thermosphacta phages* reduced bacterial counts by approximately two orders of magnitude, delayed off-odour development and extended the sensory shelf life from 4 to 8 days (Greer & Dilts, 2002). Comprehensive reviews further highlight that such phage-based interventions are being developed as targeted bio-preservation and bio-sanitisation strategies across meat, dairy and other food matrices (Garvey, 2022; Gildea et al., 2022). The European Food Safety Authority has evaluated several bacteriophage applications in foods and processing environments and concluded that appropriately characterised phage preparations can be considered safe for use as processing aids when directed against specific bacterial targets and used under defined conditions (European Food Safety Authority, 2009, 2016). Within this regulatory framework, the concurrent detection of *B. thermosphacta* and its phage in AFS datasets not only informs on emerging spoilage processes but also exemplifies how similar phage-host systems may be harnessed for future bio-preservation strategies in complex, processed foods.

Taken together, these examples illustrate that untargeted WGS-based assays such as AFS offer the opportunity to move beyond static label verification towards a more integrated assessment of composition and product quality. While current applications demonstrate that a single AFS analysis can simultaneously address ingredient authenticity, undeclared allergens and early microbiological deterioration, the same framework offers substantial potential for future expansion into comprehensive, routine quality assessment of complex foods.

3.1.3 Unexpected and mislabelled ingredients in real-world doner kebab and seafood products

A central component of the AFS approach is its capacity to reveal unexpected ingredients in real-world food products. In addition to the controlled calibration sausages, AFS was therefore applied to a set of retail products to assess its performance under realistic conditions with heterogeneous matrices and potentially imperfect labelling. In our studies, two major groups of food samples were analysed: meat-based doner kebab samples and commercial seafood items, containing multiple surimi-style foods. In five doner kebab samples collected from snack bars in the Rhine-Main area (Hellmann, Ripp, et al., 2020), the analysis revealed heterogeneity in species composition within the dissected meat components: two samples contained predominantly beef, whereas three samples showed a mixture of mostly turkey (>70 %) with minor parts of beef. This is noteworthy because doner kebab sold in Germany are legally expected to consist solely of turkey *or* beef and must be labelled correspondingly. In the context of products marketed as beef-based, the high turkey proportions are clear deviations from the labelling. The detected non-compliances are consistent with mislabelling rates reported from other surveys: Systematic DNA-based audits of kebab and similar meat products have found high frequencies of undeclared poultry and mixed species, often exceeding 50 % of tested items (Di Pinto et al., 2015; Omran et al., 2019; Szyłak et al., 2023). Furthermore, the AFS analysis uncovered minor quantities of other non-declared ingredients: Three samples showed measurable levels of soybean DNA (ranging from 0.5 % to 0.8 %), and one sample contained maize DNA (1.8 %). These unexpected findings suggest potential issues such as cross-contamination during production or intentional adulteration, which may potentially be of critical concern given health risks (e.g., allergenic reactions) and ethical implications associated with mislabelling.

A seafood sample sold as "paella-style dish with chicken" contained the expected chicken and rice, but also substantial proportions of mussel and cod, together with pea,

paprika, onion, and minor amounts of several spices and yeast. Against the declaration, we identified kisslip cuttlefish (6.9 %) and longfin inshore squid (3.9 %) to be present in significant amounts. On the level of eukaryotes alone, the sample already covers about 1.6 billion years of evolutionary divergence (Figure 8), and this complexity is further amplified by the additional contribution of prokaryotic lineages. Identifying this broad spectrum of species within a sample, barcode-based methods would require at the very least four assays to cover this phylogenetically diverse taxonomic groups: 16S for bacteria, ITS for fungi, COI/12S for metazoa, and trnL/rbcL for plants. However, it is very unlikely that this limited marker set offers the resolution required to distinguish all taxa at genus or even species level. Many of these lineages either share barcode regions that are too conserved to separate closely related species (Hollingsworth et al., 2009; Raclariu et al., 2017), or they are assayed with very short mini-barcodes because of degraded DNA and short-read sequencing, which both truncate the marker and reduce the number of informative sites (Fontes et al., 2024; Shokralla et al., 2015; Staats et al., 2016). For example, a single marker assay is not sufficient to discriminate within the order of *Gadiformes* (Mottola, Piredda, et al., 2024), so that confident species-level assignments across all domains of life cannot be guaranteed. Consequently, achieving sufficient discriminatory power would require a substantially expanded marker panel and other studies indeed resort to using up to 12 markers for complex samples (Arulandhu et al., 2017).

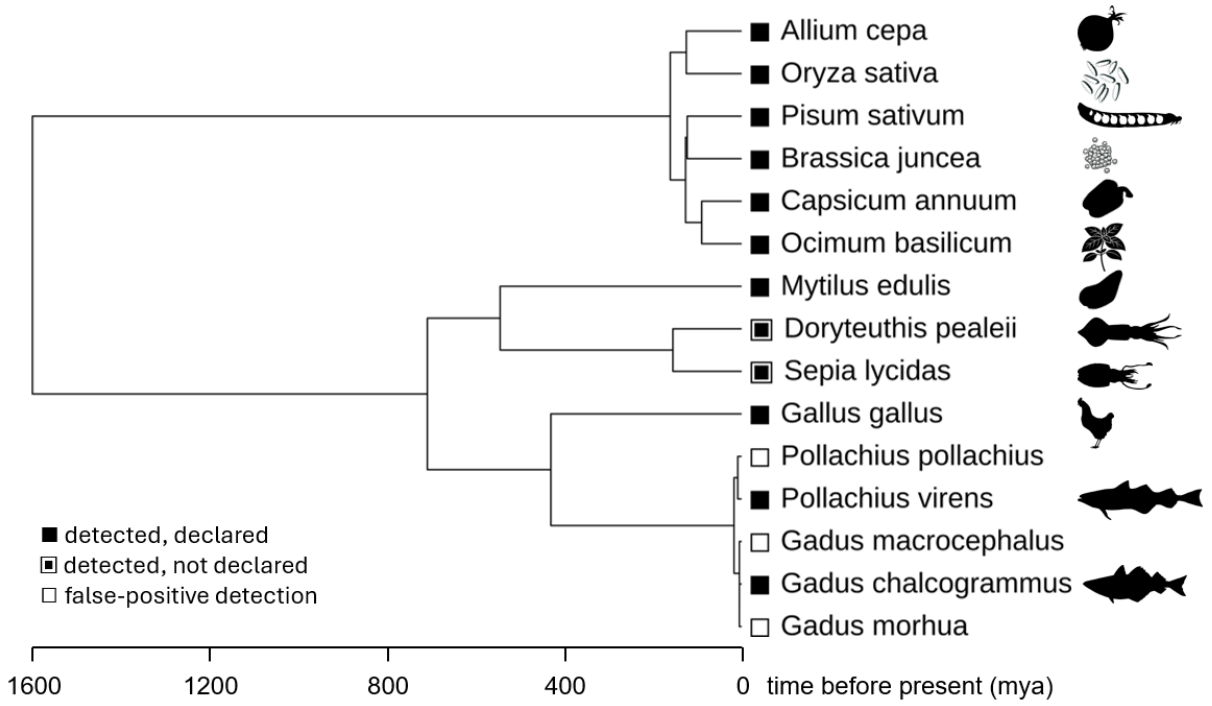

Figure 8: Time-calibrated phylogenetic context of taxa detected by AFS in paella sample. The tree shows the species identified by AFS taxonomic assignments and places them on a divergence-time scale (time before present, million years ago; mya). Symbols indicate interpretation relative to the declared recipe: ■ detected and declared ingredient, ▣ detected but not declared ingredient, and □ false-positive detection (phylogenetic data based on Gregory, 2025).

The analysis of surimi samples disclosed complex ingredient profiles: Despite having similar recipe profiles, the presence of unexpected species like Japanese threadfin bream *Nemipterus japoncus*, Atlantic horse mackerel *Trachurus trachurus* and yeast *Yarrowia alimentaria* highlight the comprehensive nature of the WGS approach, capable of uncovering both overt and subtle discrepancies between declared and actual contents. In one surimi sample imported from the Asian market, Japanese threadfin bream was identified as the main component, accounting for nearly 90 % of the product. The product was generically labelled as "whitefish" and commonly marketed as such in Asia. However, the species is not covered by the authorised "whitefish" categories in Germany (German Federal Office for Agriculture and Food, 2025), suggesting a case of mislabelling linked to regional differences in trade nomenclature rather than intentional adulteration. Such discrepancies between broad

commercial labels and the actual species used mirror findings from large-scale seafood barcoding campaigns across Europe, which report mislabelling rates of roughly 25 to 50 % in catering and retail, particularly for high-value or morphologically similar taxa (Feldmann et al., 2021; Filonzi et al., 2023; Haynes et al., 2019; Pardo et al., 2018). The same sample also contained signatures of *T. trachurus*, detectable only through BLAST of previously unclassified reads due to the absence of a reference genome. While its presence can be confirmed qualitatively, it cannot be quantified with confidence. However, the fraction of about 8 % unclassified reads provides an upper bound for its possible contribution, but cannot be interpreted as a precise quantitative estimate. This demonstrates both the sensitivity of the method to find unexpected species and its dependency on the completeness of reference databases (see Chapter 3.2.2). Another Surimi exhibited an unexpected dominance of nearly 16 % of the yeast *Yarrowia alimentaria*. This non-pathogenic, aerobic yeast is associated with high-protein, high-salt fermented and processed foods, where it can contribute to fermentation as well as spoilage (Liu et al., 2025; Nielsen et al., 2008). *Y. alimentaria* has not previously been reported for surimi or comparable seafood products. In the broader food sector, *Yarrowia spp.* are used as production hosts for aroma compounds, organic acids, polyalcohols, emulsifiers and surfactants (Zinjarde, 2014), and their biomass is explored as a protein- and fibre-rich dietary supplement with relevant vitamin and mineral content (Gottardi et al., 2021; Turck et al., 2019). *Y. alimentaria* cultures further achieve substantial protein hydrolysis, making this species a promising source of proteases with potential applications as meat tenderisers (Liu et al., 2025). Both an unintended occurrence as a food spoiler and a deliberate addition to modulate product properties are plausible explanations for this finding. Distinguishing between unintended spoilage and deliberate technological use would require targeted follow-up analyses; e.g. quantitative cultivation and qPCR across multiple batches or review of process documentation and manufacturer information.

These real-world case studies show that AFS not only recovers the declared main ingredients, but also systematically detects undeclared animal and plant components, including allergens, and revealed mislabelling in multiple cases. At the same time, the paella and surimi products indicate that AFS can uncover technologically or microbiologically relevant signals, such as process- or spoilage-associated yeasts, that inform on product quality and processing history. These examples also present interpretative limits, including dependence on incomplete reference databases, restricted quantitative confidence for low-abundance taxa, and the challenge of distinguishing technically unavoidable carry-over or environmental contamination from economically motivated adulteration.

3.1.4 k-mer minhashing and classification at scale

At the beginning of this thesis, the AFS method established a conceptional framework for species identification and quantification in complex biological mixtures by mapping short sequencing reads to a set of reference genomes using the alignment-based, Burrows-Wheeler transformation algorithm BWA aln (H. Li & Durbin, 2009). Although this approach provided a robust means of distinguishing between species, it was constrained by relatively long runtimes and a limited capacity to incorporate large or diverse reference databases, particularly when faced with the growing availability of newly sequenced genomes.

A pivotal enhancement in these updated methodologies was the replacement of BWA aln with the modern BWA mem and Bowtie2 as the primary mapping algorithms (Langmead & Salzberg, 2012; H. Li, 2013). This change led to a significant reduction in mapping time due to the more efficient algorithmic design. In addition, using Kal, KLyo and KGeflLyo calibration samples for reference, the newly implemented algorithms maintained a comparably or even higher sensitivity in read alignment (Table 2).

Table 2: Comparison of quantification performance of AFS mapping algorithms. Absolute errors summarized over all species components are shown for each sample. The lowest error for each sample is highlighted in bold.

	AFS with BWA aln	BWA mem	Bowtie2
Kal A	1.32	1.41	**1.25**
Kal B	1.80	1.37	**1.31**
Kal C	1.55	1.25	**0.85**
Kal D	**1.50**	2.22	2.00
Kal E	**0.87**	1.49	1.56
KGeflLyo A	5.30	5.28	**5.08**
KGeflLyo B	**6.97**	7.70	7.61
KGeflLyo C	5.90	**4.95**	5.12
KGeflLyo D	3.07	**2.37**	2.51
KLyo A	**1.85**	2.42	2.38
KLyo B	2.03	1.22	**1.04**
KLyo C	**2.73**	3.71	3.66
KLyo D	**1.02**	1.36	1.30
Sum Deviation	2.76	2.82	**2.74**

While the change of mapping algorithm reflects a minor improvement in both speed and precision, the transition to AFS-MetaCache represents a substantial paradigm shift (Kobus et al., 2020). Although the classical approach allowed for accurate species quantification through iterative mapping, it was inherently constrained by several limitations. First, the computational demand scaled linearly with the number of reference genomes, rendering the method less practical as genomic databases expand. Second, the necessity for iterative mapping rounds to accommodate mismatches and the reliance on heuristics for resolving multi-mapped reads often resulted in suboptimal handling of conserved genomic regions. Moreover, the subsequent use of BLAST-based metagenomic searches for unmapped reads further added to the computational overhead, while limiting the quantitative accuracy for unexpected species regions (Ripp et al., 2014).

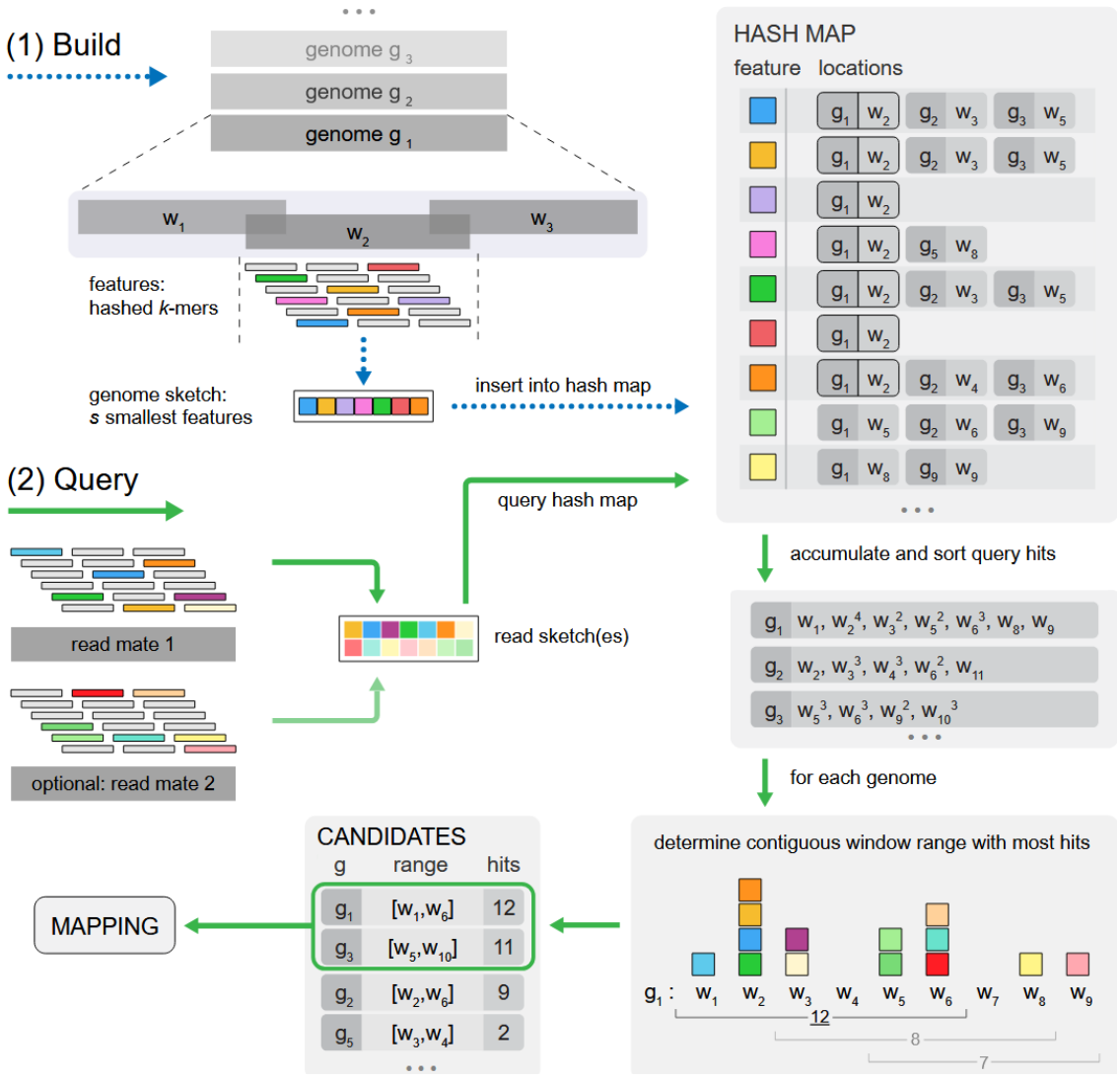

Figure 9: AFS-MetaCache workflow. 1) Build Phase (blue arrows): During database construction, each reference genome is divided into overlapping windows. Hash values for each k-mer are calculated and the smallest k-mers of each window are inserted into the database. 2) Query Phase (green arrows): For each read, hash values of each k-mer are calculated and the smallest are matched against the database. The returned numbers are used to count the number of this within each window. Target reference genomes are identified by high scores in the window count statistics. In case of tie between multiple genomes, the lowest common ancestor is returned.

In contrast, AFS-MetaCache capitalizes on an alignment-free, k-mer-centric framework that leverages minhashing techniques (Chapter 2.3). By reducing the dimensionality of the sequence data through subsampling of k-mers, the method achieves an order-of-magnitude decrease in processing requirements. The core innovation lies in constructing compact “sketches” of reference genome windows, which approximate

the full k-mer space via a representative subset (Figure 9). This accelerates read classification by enabling rapid comparisons based on Jaccard index, resulting in increased query speeds of over 400-fold (0.04 vs 17.1 mio reads/min).

Benchmarking AFS-MetaCache against other k-mer-based profilers such as Kraken2 (Wood & Salzberg, 2014), Kraken2+Bracken (J. Lu et al., 2017) and CLARK (Ounit et al., 2015) on the calibrator sausages confirmed that the algorithmic redesign not only increases speed and scalability but also improves classification quality. Across all calibrator sausages, AFS-MetaCache consistently produced the lowest sums of false-positive assignments (on average 0.7 - 1.1 % per sample vs. 1.3 - 4.4 % for CLARK, Kraken2 and Kraken2+Bracken), which was especially pronounced for closely related species pairs such as cattle vs. water buffalo and sheep vs. goat, where false-positives were reduced by up to almost an order of magnitude. Deviations of quantification between expected and measured meat proportions were lowest in about half of the samples and *on-par* with the other methods in the other half of the samples (Kobus et al., 2020). In addition, the method introduces an efficient partitioning scheme: By dividing large collections of reference genomes into manageable subsets, peak memory usage can be drastically reduced. As there is no maximum number of subsets, any number of reference genomes can be analysed in an AFS-MetaCache analysis, allowing our method to scale with the constantly growing availability of genomic reference data. To our knowledge, there is currently no other method capable of combining all available genomes into a single analysis. Therefore, AFS-MetaCache has a unique advantage over other existing methods and is setting a new standard for metagenomic screening in complex biological samples.

3.1.5 Can longer reads improve sensitivity? Advantages of third generation sequencing for taxonomic resolution and false-positive control

Long-read sequencing technologies, such as those offered by ONT and PacBio, provide significant advantages over Illumina short-read platforms in resolving complex genomic architectures (Albertsen, 2023; Hu et al., 2021; Kovaka et al., 2023; Nurk et al., 2022). These technologies facilitate advantages in the mapping of reads to difficult genomic regions (Ebbert et al., 2019), including repetitive elements (Nurk et al., 2022), structural variants (De Coster et al., 2019; Romagnoli et al., 2023), and regions with high heterozygosity (T. Zhang et al., 2022), which historically posed significant challenges for assembly and mapping with short-read data like Illumina reads. The ability to generate reads spanning tens to hundreds of kilobases enhances the resolution of these challenging regions, thereby overcoming longstanding limitations in genomic analysis (Ebbert et al., 2019; Ryan & Corvin, 2023). Although long-read platforms were initially characterized by high error rates, recent advancements in sequencing chemistry and the development of robust error-correction algorithms have reduced these rates for both PacBio and ONT to about 1 % (Table 3; Espinosa et al., 2024; Scarano et al., 2024).

Table 3: Overview of short- and long-read sequencing technologies and platforms: An exploration of distinctive features and advantages for use in AFS (based on Espinosa et al., 2024; Scarano et al., 2024).

	Illumina		**Pacific Biosciences**		**Oxford Nanopore Technologies**	
Sequencing principle	sequencing by synthesis, fluorescence		sequencing by synthesis, fluorescence		nanopores, eletrical current	
Accuracy	>99.9 %		>99 %		99 %	
Direct methylation detection	no		yes		yes	
Platform	NextSeq 2000	NovaSeq X series	Sequel II / IIe	Revio	GridION	PromethION
Data type	paired-end	paired-end	HiFi	HiFi	long read	long read
Read length	up to 2x300 bp	up to 2x150 bp	>20 kb	15-20 kb	>100 kb	>100 kb
Maximum throughput per flow cell	60-180 Gb	up to 8 Tb	30 Gb	90 Gb	48 Gb	50-200 Gb
Sequencing cost per Gb (USD)	30/20	2	43-86	11	72	10-25
Equipment cost (USD)	335,000	985,000	525,000	779,000	69,000	436,400

Already at an early stage during development of AFS, we observed that repetitive sequences could lead to false read mapping. For instance, bovine reads originating from repetitive regions were found to map not only to the bovine genome but also to genomes of closely related species such as buffalo, sheep, and goat. These reads, derived from Mammalian-wide interspersed repeats (MIRs) of approximately 260 bp, are highly conserved across the mammalian clade (Jurka et al., 1995; Krull et al., 2007; Smit & Riggs, 1995), and resulted in false-positive mapping and species identifications, thereby skewing the quantification process of the AFS analysis (Ripp et al., 2014).

By employing ONT long reads, we were able to extend the read length, thereby increasing taxonomic profiling resolution and reducing the overall false-positive

detection rate for all species over all samples (Chapter 2.5). The average read lengths achieved in this study was only 500 bp on average, primarily due to the highly fragmented nature of the input DNA extracted from boiled sausage samples (Dolch et al., 2020; Köppel et al., 2012). This degradation is inherent to the calibration samples and precludes the possibility of generating substantially longer reads. Despite only moderate increase in fragment length compared to short read sequencing, this translated into systematically better profiling outcomes: for all tested but one tested sample, ONT datasets achieved lower mean deviations from the known ingredient proportions than the corresponding Illumina datasets (Figure 10). In addition, ONT consistently produced markedly lower number of false-positive classifications, despite the much higher error rate of 2.6 %.

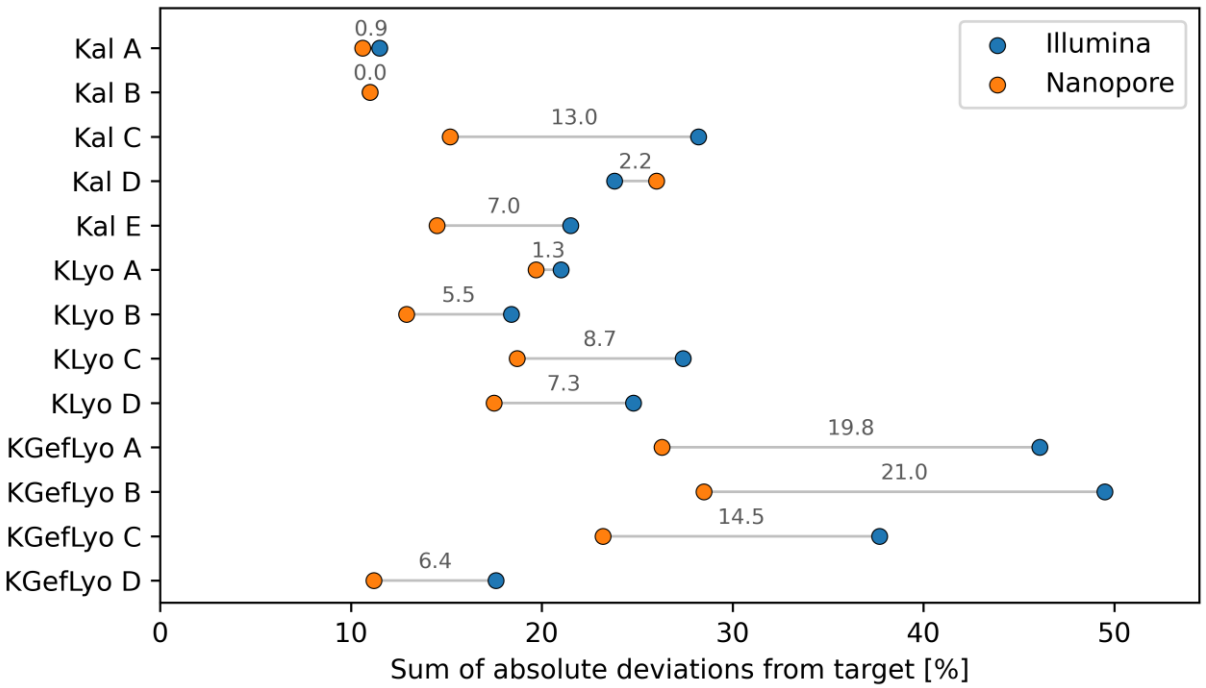

Figure 10: Comparison of quantification accuracy between Illumina short read sequencing and Oxford Nanopore Technologies long read sequencing across calibration samples. For each sample, the sum of absolute deviations from target (in %) is shown, where lower values indicate closer agreement with the known ingredient proportions. Points denote platform specific results (blue: short reads; orange: long reads), and the labels report the absolute difference between platforms.

Mechanistically, longer reads provide more independent discriminative sequence features per molecule, resulting in more k-mers across more windows, which stabilizes candidate ranking and reduces ambiguous matches driven by conserved or repetitive regions shared among related eukaryotic genomes. Even with read lengths that were "only" a few-fold longer (reflecting degradation in processed food), the added context was sufficient to reduce misclassification and to improve abundance estimates. This limitation of fragmented input DNA, often originating from harsh processing conditions during food production, poses a general challenge to all DNA-based identification methods (Dolch et al., 2020; Mano et al., 2017; Piskata et al., 2019), which can be mitigated to some extent by long read sequencing.

3.2 Limits of confidence: controlling false findings across matrices, taxa, and databases

The real-world samples demonstrate that, although AFS reliably quantifies declared main ingredients, reveals undeclared taxa across broad phylogenetic ranges and highlights cases of possible mislabelling or quality issues, key challenges remain before these capabilities can be used directly for routine enforcement decisions. The subsequent sections examine the main methodological and practical constraints that need to be addressed for AFS to be integrated into standard control workflows.

3.2.1 Why do read counts not equal biomass? Tissue DNA content, matrix dependence and ploidy as confounders

Matrix effects represent a critical challenge for all DNA-based identification methods by introducing biases during DNA extraction, amplification, and sequencing from complex food matrices. They arise when a sample's inherent biological, chemical, and

physical properties interfere with the quality and yield of extracted DNA (Andronache et al., 2025; Shokralla et al., 2015). The food matrix is formed by a complex mixture of diverse tissue types, comprising proteins, lipids, polysaccharides, polyphenols, and other secondary metabolites, and may compromise multiple species, each with its own characteristics. This complicates the recovery of high-quality DNA, which in turn introduces biases that can skew the downstream quantification of species proportions, e.g. by inhibiting enzymatic reactions (Lo & Shaw, 2018; Shokralla et al., 2015). Understanding these effects is essential because even minor discrepancies in DNA extraction or sequencing library preparation efficiency may lead to significant errors in species detection, potentially resulting in the over- or underestimation of ingredients, which in turn affects both authenticity verifications and food safety assessments.

Lipid-rich food matrices, such as fatty meat and fish or dairy products, often compromise DNA extraction by physical partitioning effects. During homogenization and cell lysis, lipids promote stable emulsions and poor phase separation, which can trap DNA at interphases and increase contaminant carry-over (Pirondini et al., 2010). In silica column and magnetic bead workflows, residual lipids may additionally impair binding and washing steps, thereby reducing yield and purity (Pirondini et al., 2010). By contrast, starch and other polysaccharides, common in flours, seeds, and many plant-derived foods, tend to persist as co-precipitants or high-molecular-weight contaminants. This increases viscosity, complicates pipetting and quantification, and can inhibit multiple downstream enzymes (Buljevíc et al., 2025; Rezadoost et al., 2016). Notably, the inhibitory strength varies among polysaccharides, and even comparatively neutral carbohydrates such as starch can cause practical failures, for example by forming compact pellets that are difficult to resuspend (Inglis Id et al., 2018). A common countermeasure is the use of high-salt lysis and precipitation conditions, which can reduce polysaccharide carry-over and facilitate removal (Rezadoost et al., 2016). Polyphenol-rich matrices, including herbs and spices and some fruits and vegetables, are dominated by oxidation-related mechanisms. Upon tissue disruption, phenols can

oxidize and form reactive products that covalently bind to DNA, forming complexes that are lost during cleanup and resulting in strongly reduced yields (Schenk et al., 2023). Polyvinylpyrrolidone is therefore frequently included to sequester phenolic compounds and limit these reactions (Schenk et al., 2023). Beyond these intrinsic matrix-derived inhibitors, food processing itself can further reduce DNA accessibility. Conditions such as elevated temperatures, extreme or fluctuating pH, and the presence of cross-linking agents can promote covalent cross-links between DNA and proteins or polysaccharides (Nayak et al., 2024; Ou et al., 2020). Such cross-linking can impede cleanup and reduce the availability of intact DNA during sequencing library preparation (Nayak et al., 2024), which can lead to systematic underestimation of affected species.

In practical food samples, all the factors discussed frequently co-occur and can interact: Taking the paella-style dish from (Chapter □) as an example, the food typically combines a starch-rich matrix from rice with substantial amounts of proteins and lipids from fish, seafood, and added oils, while bell pepper, spices and other plant-derived ingredients can contribute polyphenolic compounds. This combination can couple polysaccharide-driven viscosity and co-precipitation with lipid-driven emulsion formation and phase separation issues, and with polyphenol-related DNA loss and enzymatic inhibition. Such a composite food matrix generally necessitates multiple, complementary modifications to standard extraction workflows, including strong mechanical and enzymatic lysis, CTAB-based high-salt chemistry with phenolic-binding reagents to mitigate polysaccharides and polyphenols, and inhibitor-targeted clean-up; nonetheless, differential extractability between ingredients can persist and influence downstream quantification.

However, these chemical and physical factors do not fully explain the observed variability. Additionally, biological qualities of food matrices further enhance the complexity of reproducible DNA extraction. Food samples often exhibit heterogeneous textures and variable viscosities, which can lead to inconsistent cell lysis and uneven distribution of target DNA within the sample. A poorly homogenized sample may yield

DNA that is not representative of the entire matrix, resulting in sampling biases that affect quantification accuracy (Ripp et al., 2014, Chapter 1). Rigid cell walls can further limit lysis efficiency and thereby constrain DNA recovery (Schenk et al., 2023). Moreover, intrinsic differences in tissue architecture and cellular composition add another layer of variability: various tissues such as muscle, liver, adipose, and connective tissues differ significantly in cell density, nuclear DNA content, ploidy and mitochondrial abundance (Picard, 2021; Siuta et al., 2023). This is evident even within a single edible species when different tissues are compared: for Atlantic cod, DNA content measured in raw white muscle (478 µg/g wet weight) is markedly lower than in raw red muscle (870 µg/g), and far lower than in kidney (4,757 µg/g) or spleen (9,040 µg/g). A similar gradient is reported for common carp, where raw white muscle (672 µg/g) < raw red muscle (1,177 µg/g) < liver (4,221 µg/g) and reaches very high values in both kidney and spleen (>18,400 µg/g, respectively; Rehbein & Oehlenschläger, 2009). Consequently, when a food sample contains a mixture of tissues, the DNA contribution of each tissue does not necessarily correlate with its weight or volume in the final product. We observed this effect during the analysis of the calibration samples Kal A-E (Hellmann, Ripp, et al., 2020; Ripp et al., 2014), where pork was added both in form of lean meat and high-fat lard and rind (Eugster et al., 2009). As the latter are tissues rich in lipids, the DNA content per weight was reduced compared to muscle tissue and therefore resulting in a constant underestimation, both by AFS (Chapter 1 & 3) and PCR-based detection methods alike (Eugster et al., 2009; Köppel et al., 2012).

A third layer of complexity arises from tissue- and species-specific ploidy, which directly changes the amount of nuclear DNA per cell and can therefore bias read-count–based quantification if uncorrected. For example, wild-caught salmon are diploid, but in aquaculture triploidy is sometimes induced to produce sterile animals with enhanced growth rates; both beneficial traits for production management (Murray et al., 2018) This results in higher DNA amounts per weight of tissue for triploid salmon, which in turn leads to a bias in read counting-based quantification, if ignored. Polyploidy is also

widespread across crop plants, where it directly affects nuclear DNA content per cell. Prominent examples include hexaploid wheat (*Triticum aestivum*), tetraploid potato (*Solanum tuberosum*), and octoploid cultivated strawberry (*Fragaria×ananassa*) (Edger et al., 2019; Levy & Feldman, 2022; Xu et al., 2011). Their larger genome size and higher DNA content per cell can lead to substantial overrepresentation in sequencing data, if read counts are not normalized to genome size and ploidy levels, thereby rendering precise quantification impossible (Chapter □). For this normalization, information about ploidy for each species is required; for most food-relevant species, such information is available (Gregory, 2025; Henniges et al., 2023). In missing cases, it can be estimated from WGS data: ploidy can be inferred from allele-frequency histograms at biallelic SNPs in single-copy, nuclear loci, where expected peaks differ by ploidy: ~0.5 for diploid, ~0.33/0.67 for triploid, and ~0.25/0.5/0.75 for tetraploid (Viruel et al., 2019). This signal can be formally modelled and ploidy estimated using tools like *nQuire* (Weiß et al., 2018), which evaluates diploid, triploid, and tetraploid models from mapped short reads. To implement such an approach in a food-mixture context, two prerequisites must be met: (i) the inference operates on single-species data, requiring a read binning step after an initial AFS analysis; and (ii) the target species must reach sufficient nuclear coverage, typically >20x (Viruel et al., 2019). In cases of unfulfilled requirements, ploidy has to be determined using complementary methods, for example PCR-based genotyping of highly polymorphic short tandem repeat (STR) markers (Jacq, 2021).

Read-count-based quantification, as employed by AFS, is inherently matrix-dependent rather than a straightforward reflection of ingredient mass. Extraction yield, DNA integrity, and library-conversion efficiency can vary systematically between ingredients and thereby shift apparent species proportions. Importantly, this dependence is not unique to AFS: targeted DNA-quantification approaches such as qPCR and ddPCR likewise rely on matrix-adapted calibrators and/or conversion factors to translate DNA signals into weight-to-weight statements in complex foods (Köppel et al., 2012, 2019). Similarly, DNA metabarcoding read proportions are well-known to deviate from

biomass because amplification efficiencies are primer- and taxon-specific, and because tissue-dependent mitochondrial copy number can systematically skew read abundances (Giusti et al., 2024; Krehenwinkel et al., 2017; Piñol et al., 2019). Consequently, quantitative interpretation should be framed as DNA contribution under the applied workflow, not as direct biomass composition, unless calibration data or correction factors are available. In practice, robust AFS quantification benefits from standardized homogenization and extraction, calibration using relevant reference materials, and normalization by genome size and ploidy. In difficult matrices however, results near regulatory or decision thresholds should be interpreted conservatively and ideally supported by controls and orthogonal confirmation.

3.2.2 How reliable is species-level calling for close relatives? Influence of read length, genome relatedness, and cross-assignment

Detection of non-declared species is essential for identifying mislabelling. At the same time, the sharing of conserved regions across related reference genomes and resulting misassignment of sequence reads shows that species-level interpretation is in some cases prone to misinterpretation. This issue of taxonomically ambiguous hits caused by closely related species within the same genus or among recently diverged lineages was already evident in AFS validation work on meat reference materials, where near relatives were intentionally included to probe false-positive assignment. In these datasets, low-level spillover between closely related pairs such as sheep and goat as well as cattle and water buffalo can occur even when the true composition is known, reflecting the fact that a subset of short reads lacks sufficient discriminatory signal to be assigned unambiguously at species level (Ripp et al., 2014). In practice, such patterns can represent genuine multi-species content, but they can also arise as a systematic artifact of read classification when reference genomes share extensive conserved regions or repetitive sequences. Under routine analysis conditions, such false-positive

misassignments could result in false findings of non-compliance und unjustified rejections of the products by authorities.

In reference sausages, up to 1.7 % false-positive goat reads were observed in a sample containing 55 % sheep. To investigate this effect, *in silico* read mixtures were generated with known proportions and the related species absent, showing that false-positive reads scaled linearly with the real ingredient proportion and that shorter reads substantially increased classification errors (Hellmann, Ripp, et al., 2020). For example, a 100 % sheep dataset yielded 5.1 % false-positive goat assignments at 50 bp but only 2.7 % at 150 bp. In consequence, the limit of detection (LoD) in AFS is influenced by both genome relatedness in the reference database and sequence read length; with closely related genomes (e.g., sheep vs. goat and cattle vs. buffalo) included using 150 bp reads, the LoD was reported as 1.6 %, compared to 1.0 % when only distant species were tested (Hellmann, Ripp, et al., 2020). Because closely related species share conserved and low-complexity regions, short reads frequently map equally well to multiple reference genomes, forcing AFS to distribute evidence and thereby generating cross-assignments. Increasing read length reduces this ambiguity: longer reads are more likely to include species-informative variants and to span out of conserved or repetitive sequence into unique flanking regions, increasing classification by widening the score gap between the best and second-best matches (Pearman et al., 2020; Treangen & Salzberg, 2011).

The seafood case studies (Chapter □) illustrate the same mechanism in a taxon complex where recent divergence makes cross-assignment particularly pronounced. In multiple foods marketed as containing Alaska pollock, AFS reported both signals for Atlantic cod and Pacific cod in addition. However, several observations indicate that these secondary signals are largely a classification artifact rather than evidence of a multi-species ingredient mixture: Firstly, the same three-species constellation recurred across independent products and recipes. Secondly, the declared catch area (FAO 67) makes an Atlantic cod ingredient implausible as bycatch because it is absent from the

Northeast Pacific (Thünen Institute of Fisheries Ecology). Thirdly, both Pacific and Atlantic cod are more expensive than Alaska pollock, rendering economically motivated substitution unlikely (Pardo et al., 2018). Finally, the read distribution technically observed by AFS was consistently skewed to approximately a 4:2:1 ratio for Alaska pollock:Pacific cod:Atlantic cod; and the same pattern is observed for pure Alaska pollock samples. Consequently, at least the Atlantic cod fraction must represent cross-assignment among closely related reference genomes. The remaining partitioning within the *Gadus* complex cannot be interpreted robustly at species level. However, the recurrence of the same pattern in pure Alaska pollock samples and the lower price of this fish together support interpretation of the Atlantic and Pacific cod signals as a false-positive cross-assignment rather than a true ingredients.

To isolate bioinformatic ambiguity from sample-derived biases, *in silico* generated read mixtures were also used in this thesis to benchmark the intrinsic quantitative performance of the AFS pipeline under near-ideal conditions. Reads were simulated from the reference genomes of the respective species, so that the true input composition was exactly known. These datasets approximate an experimental scenario in which biological sources of bias, such as matrix effects, differences in DNA extraction efficiency, and intraspecific genetic variability, are deliberately excluded. This approach enables a focused examination of conserved genomic regions among closely related species and deviations between the known input composition, and the inferred proportions therefore provide a direct assessment of the classification performance and quantitative accuracy of AFS in taxonomically challenging scenarios. Under these conditions, the observed deviations in estimated proportions were as low as 0.25 ± 0.30 % across all species (Table 4). This level of agreement represents excellent quantitative performance, particularly because the mixtures contained three closely related *Gadidae* species, *Gadus morhua*, *Gadus chalcogrammus*, and *Melanogrammus aeglefinus*, which are difficult to discriminate on species level even with targeted methods (Filonzi et al., 2023). These findings highlight the potential of WGS based AFS

to deliver highly accurate species quantification once sample derived challenges are minimized experimentally or are adequately addressed through calibration experiments and appropriate normalization strategies.

3.2.3 Interpretation in the sub-percent grey zone: separating true traces from low-level artifacts

Signals in the sub-percent range lie in a grey zone where true trace content and methodological artifacts can no longer be cleanly separated. While the preceding section addresses systematic cross-assignment among recently diverged taxa, interpretational ambiguity at low levels remains even beyond such taxon complexes, because several independent error sources can generate small but reproducible read fractions. This is consequential in routine food control, where an interpretation of weak signals may trigger false non-compliance decisions. At the same time, dismissing them categorically however can obscure genuine traces, which is especially important in cases of allergen addition, where even admixtures well below 0.1 % can have harmful consequences for the consumer.

AFS relies on a reference database that contains multiple genomes. During analyses, this has the practical consequence that reads are not only assigned to the true ingredients present in a food, but also to numerous additional species represented in the database, all at low levels. To keep such background noise manageable, results are often filtered using a cutoff of 0.1 %, discarding all taxa below this threshold. However, this strategy can also remove true-positive detections when a real trace ingredient contributes fewer than 0.1 % of reads. Thus, low-level results require contextual interpretation rather than automatic filtering.

Table 4: Validation of AFS quantification on simulated fish datasets. For each simulation, predefined *target* species proportions were compared to the corresponding AFS-derived estimates (%). Columns list the investigated fish species; rows represent independent simulated mixtures covering single-species samples and multi-species compositions with varying relative abundances. Empty cells indicate species not included in the respective simulation.

	Gadus morhua		***Gadus chalcogrammus***		***Pollachius virens***		***Melanogrammus aeglefinus***		***Oncorhynchus mykiss***		***Salmo salar***		***Thunnus albacares***	
	target	AFS	target	AFS	target	AFS	target	AFS	target	AFS	target	AFS	target	AFS
simulation 1	0	0.01	100	99.99										
simulation 2	100	99.98	0	0.02										
simulation 3									50	49.98	50	50.02		
simulation 4									33.33	33.31	33.33	33.33	33.33	33.36
simulation 5	25	24.77			25	24.78			0	0.01	50	50.44		
simulation 6	25	25.8	25	24.09			50	50.11						
simulation 7	25	25.83	25	24.1	25	25.01	25	25.06						
simulation 8	12.5	12.98	25	24.19	12.5	12.43	12.5	12.44			12.5	12.64	25	25.32
simulation 9	12.5	12.83	12.5	11.96	12.5	12.42	25	24.87	6.25	6.31	25	25.28	6.25	6.33

Real-case screening data illustrate ambiguities at low levels. In doner kebab samples dominated by cattle, 0.1 - 0.4 % reads assigned to sheep/goat could be interpreted as candidate false-positives driven by shared conserved elements within Bovidae. However, comparison to expected false-positive background values suggested that in at least two samples the measured sheep/goat fractions were slightly higher than expected for an almost pure cattle matrix, so true trace amounts could not be excluded. Conversely, low chicken signals of 0.2 - 0.3 % in turkey-dominated samples were considered consistent with false-positive assignment. Overall, these examples show that observed values must be judged relative to empirically expected misassignment rates, rather than in isolation. The applied interpretation strategy benchmarks low-level detections against empirically expected cross-assignment rates arising from closely related taxa, to distinguish plausible trace content from background misclassification.

Another approach to detect more true trace ingredients is to reduce background by building the reference database more selectively using a two-step approach: First, the sample is screened qualitatively against an AFS database composed of mitochondrial genomes to identify candidate species present. Mitochondrial references are well suited for this initial step because (i) their small genome size (~14 – 20 kb) for animals (Boore, 1999) enables rapid searches across very large databases and (ii) mitochondria occur at high copy number per cell (Rath et al., 2024), which can improve detectability even when a species is present at low abundance. Based on this qualitative screen, a second AFS run is then performed using a custom nuclear-genome database containing only the species detected in the first step. By restricting the search space, this can markedly reduce background assignments and may increase sensitivity for trace-level components. This targeted approach can therefore lower spurious low-level hits, although it cannot fully eliminate them. In addition, this approach is more challenging for plant ingredients, as their often enormous mitochondrial genome can be up to three orders of magnitude higher than that of animals (Huang et al., 2024). At the same time, mitochondrial sequences can be highly conserved among taxa.

Therefore, even this two-step strategy does not guarantee reliable detection of very low-level ingredients, and ambiguous assignments at trace abundance may persist.

Very low percentage detections should often be treated as hypotheses that require contextual validation, rather than precise findings. Practical mitigation includes: (i) conservative taxonomic reporting where ambiguity is expected (as discussed under "close relationship"); (ii) the use of study- or run-specific false-positive expectations derived from simulations for relevant species pairs; and (iii) for results near decision thresholds, confirmation by orthogonal methods or re-analysis under stricter filtering and curated reference databases before drawing compliance conclusions.

3.2.4 Why deeper sequencing cannot fix missing references: limitations of reference-based screening

AFS is fundamentally reference-based and analytical results are therefore constrained by database composition and by the correctness and completeness of included reference genomes. Consequently, uncertainty in database coverage and reference quality propagates into both qualitative detection and quantitative read-fraction estimates.

A core constraint of whole-genome, reference-based food metagenomics is that detection is conditional on representation: the availability and quality of reference genomes are prerequisites that define the practical measurement limits of AFS. If a taxon is absent from the database, it cannot yield a taxon-specific signal. Reads from such taxa remain unclassified or, in worst case, are assigned to related taxa due to cross-assignment. Non-detection is therefore conditional on database coverage and cannot be interpreted as evidence of true absence. This dependence was observed in surimi analyses: in one sample, the main components could be classified, but a substantial read fraction remained unclassified. Subsequent qualitative analysis of

these reads suggested Atlantic horse mackerel (*Trachurus trachurus*) as an additional ingredient. Because no suitable *T. trachurus* reference genome was available for AFS, the finding had to remain qualitative and was reported as probable presence rather than a quantified ingredient fraction.

Unclassified reads should therefore be treated as an interpretable readout, not as residual noise. In AFS, any read that cannot be assigned reflects a mismatch between the observed data and the available reference database. At least four non-exclusive causes can contribute to unclassified reads: (i) the sample contains DNA from a taxon that is not present in the database , for example unexpected ingredients, non-food taxa (e.g. microorganisms, parasites), or missing reference genomes; (ii) the correct species is present in the database, but its reference assembly lacks parts of the genome (e.g. gaps, miss- or collapsed assemblies; Asalone et al., 2020; Peona et al., 2021), which leads to an underestimation of the species coupled with an elevated unclassified fraction; (iii) the correct species is present in the database, but the sample's genotype differs beyond recognition by the algorithm from the reference (e.g. high intraspecific diversity, different populations or breeds, or haplotypes and structural variants not represented in the reference Asalone et al., 2020; Bohling, 2020; Garrison et al., 2018) , which is indistinguishable from (ii), but originates from biological rather than technical properties; and (iv) reads are hard to map because of technical issues like elevated error profiles (Schirmer et al., 2016), fragmentation (W. Li & Freudenberg, 2014), index swapping (Costello et al., 2018), or residual sequencing adapters (Bolger et al., 2014). The final category is workflow-dependent and may be unavoidable to some extent, whereas (i) – (iii) primarily reflect database composition and completeness. For reporting, unclassified fractions should be stated and interpreted: elevated unclassified fractions are a diagnostic signal that can indicate missing taxa, unsuitable references, or overly stringent parameters. They can motivate follow up steps such as qualitative screening against alternative reference sets, parameter reassessment, or targeted database augmentation.

Beyond missing taxa and mismatched references, errors within reference genomes themselves can generate false-positive detections. During AFS analysis of a calibration sausage, several hundred reads were initially assigned to *Neisseria gonorrhoeae*, yielding an apparent fraction of approximately 0.04 % when the reference genome for strain TCDC NG08107 was included. Subsequent inspection indicated that the signal was inconsistent with true *Neisseria*: Some regions showed strong similarity to ruminant sequences, suggesting that this reference record contains bovine derived sequence and should be treated with caution (Ripp et al., 2014). An independent analysis found that this record comprises segments attributable to cattle and sheep (Merchant et al., 2014). Such issues occur at scale: (Breitwieser et al., 2019) reported human derived sequence in more than 2000 bacterial and archaeal assemblies, often enriched for high copy repeats such as Long interspersed nuclear elements (LINEs), Alu and other Short interspersed nuclear elements (SINE), and satellite sequences. In a read classification setting, these contaminants can yield recurrent low-level, false-positive signals that appear technically well supported.

A further example is the presence of Illumina adapter sequence in the reference genome of the common carp *Cyprinus carpio*. This artefact is attributed to insufficient quality control during assembly and subsequent deposition in public repositories. This issue has been reported in independent analyses (Etherington, 2014; S. Huang, 2021; Mann et al., 2023). Adapter-containing reads, which are commonly found within Illumina data sets, can therefore yield high scoring but biologically implausible matches to carp during analysis. The broader implication is that errors in public references can propagate through downstream pipelines and yield reproducible artefacts. Unexpected low-abundance hits should therefore be evaluated for plausibility, including inspection of read coverage distribution across the reference, identification of clustering in anomalous regions, and assessment of alternative explanations such as conserved elements or adapter-related matches.

Even in the absence of explicit contamination, public genomes differ in completeness and correctness, and this heterogeneity can bias both detection and quantification. Incomplete references reduce the fraction of assignable reads and can change the observed read fraction for a species. These effects are not resolved by deeper sequencing of the food samples: if sequence is missing from the reference, reads originating from that sequence remain unassigned or are diverted to other taxa regardless of coverage. In addition, genomes in databases usually represent a single assembly derived from one individual or strain. Intraspecific variation can therefore reduce classification confidence, particularly in regions affected by structural variation, and can contribute to unclassified reads that still show the target species among top hits in follow-up taxonomic analyses.

Reference quality should therefore be treated as a controlled variable. In this sense, the database becomes part of the measurement model: reference availability defines which hypotheses are testable, unclassified reads often reflect database mismatch, heterogeneous reference quality introduces systematic bias in detection and quantification, and contaminated references can generate reproducible false-positive findings. Database curation, transparent versioning, and conservative interpretation are therefore required, particularly for low-abundance signals.

3.3 From reference-limited to decision-ready: future directions for AFS in food control

The preceding sections highlight that AFS already delivers broad screening power, but that routine decision use is still constrained by reference availability, low abundance ambiguity, and database related artefacts. The following perspective therefore outlines practical development paths that can progressively increase interpretability and robustness toward decision ready workflows in food control.

3.3.1 Towards reference-complete AFS: reducing unclassified reads through reference expansion and high-quality assemblies

AFS can in theory detect any species' DNA present in a sample, but it requires that species' genome to be represented in reference databases. This is a critical limitation: public DNA barcode repositories (e.g. BOLD) contain sequence records for over 1.3 million species (Ratnasingham et al., 2024), whereas the number of species with whole-genome assemblies is only on the order of tens of thousands (Goldfarb et al., 2025). In practical terms, some food-relevant species still lack a reference genome sequence, especially minor crops, regional ingredients, or less-studied taxa. With roughly ~1.5 million described eukaryotic species in total (M. Blaxter et al., 2025), the coverage of genomic databases remains limited. If an ingredient or contaminant in a food sample has no genome available, WGS methods cannot confidently quantify it - the DNA reads either remain unclassified or may falsely match to the closest related genome present, as it was seen in the surimi case with Atlantic horse mackerel.

Despite this dependence on reference representation, the situation is improving rapidly because genome sequencing and assembly have become substantially cheaper and more scalable in recent years. Declining costs per unit of sequence data (Table 3),

higher platform throughput, and routine multiplexing enable a growing number of projects to generate whole-genome resources for an increasing number of species (Figure 11, blue), which for AFS translates directly into improved database coverage and a reduced fraction of unclassified reads. In parallel, reference quality is improving: long-read sequencing, combined with modern scaffolding approaches like Hi-C, linked reads and optical mapping and more mature assembly and polishing pipelines, now yields chromosome-level and haplotype-resolved assemblies rather than fragmented drafts (Figure 10, yellow; Kovaka et al., 2023; T. Zhang et al., 2022). For AFS, such improvements are practically important because higher contiguity and completeness reduce biases from missing sequence content, and haplotype resolution can mitigate reference bias in taxa with low interspecific divergence, thereby stabilizing read assignment and abundance estimates.

Large, international genome initiatives exemplify this shift toward systematic, high-quality reference generation. The *Earth BioGenome Project* provides the overarching framework, aiming to produce chromosome-level assemblies for all described eukaryotic species over a decade (M. Blaxter et al., 2025; Lewin et al., 2018). Beyond generating genomes, the *Earth BioGenome Project* establishes shared standards for sampling, ethical governance and benefit-sharing, and coordinates interoperable data infrastructures across participating nodes - and, critically, it reflects a broader trend toward reference quality management as an explicit objective rather than an implicit by-product of assembly production. This quality-management trend includes specimen validation and vouchering, harmonized metadata, transparent provenance, and increasingly routine screening for contamination and technical artefacts prior to public deposition. For AFS, these practices are not cosmetic: they reduce the probability that erroneous references propagate into false-positive detections, and they increase the interpretability of both assigned and unclassified read fractions by improving traceability to specific databases and specimen lineages. Within this umbrella, the *Vertebrate Genomes Project* (Rhie et al., 2021) and others demonstrate what is

technically achievable: by applying long-read sequencing, Hi-C scaffolding and stringent assembly quality criteria, the projects produce near-error-free, haplotype-resolved genomes that serve as methodological benchmarks and training sets for assembly, annotation, structural-variant analysis and comparative genomics. At the regional level, initiatives like the *Darwin Tree of Life* (M. L. Blaxter, 2022) and *African BioGenome Project* (Ebenezer et al., 2022) illustrate how national and continental efforts can scale biodiversity genomics. These consortia integrate genome sequencing with upstream activities such as specimen validation, vouchering and biobanking, and with downstream efforts including data standardisation, open-access dissemination and the development of community workflows. Together with taxon-focused efforts (e.g., large-scale plant and arthropod genome programmes), these initiatives do not

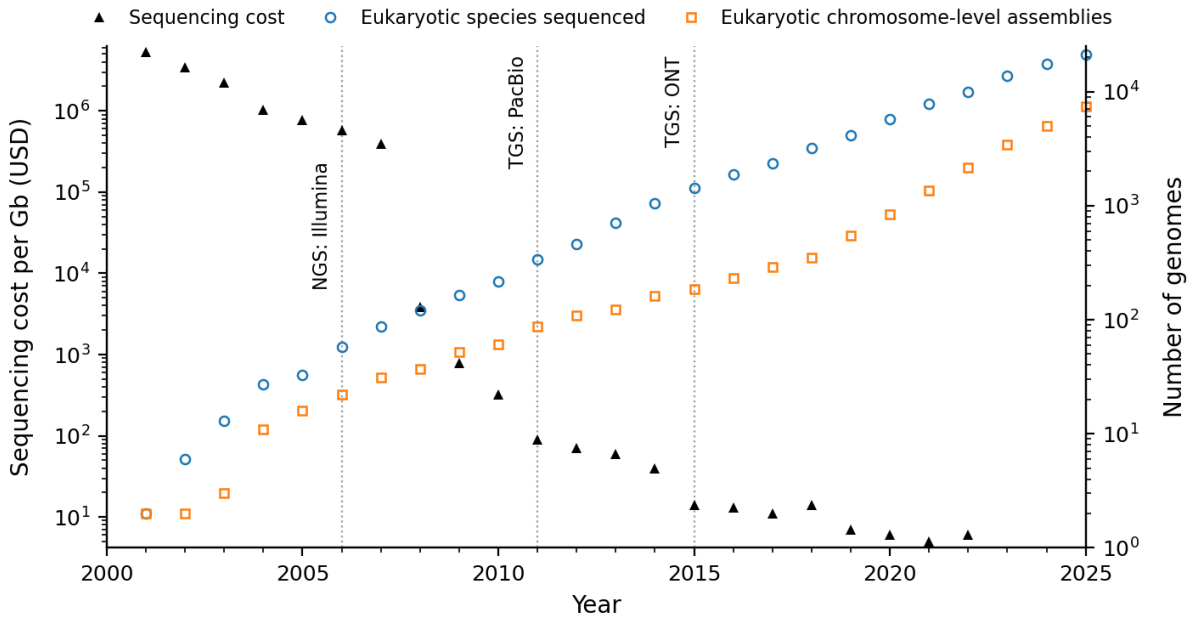

Figure 11: Decline in DNA sequencing cost and growth of public eukaryotic genome resources (2001–2025). Cost per Gb in USD (black triangles, left y-axis), cumulative number of eukaryotic species represented by at least one publicly available genome assembly (blue circles, right y-axis), and the cumulative number of species with chromosome-level or complete assemblies (orange squares, right y-axis) are shown. Vertical dotted lines mark the introduction of major sequencing technology eras (NGS: Illumina; TGS: PacBio; TGS: Oxford Nanopore Technologies).

merely increase the number of available genomes, but also establish reproducible pipelines and quality criteria that directly strengthen the reliability of reference-based analyses such as AFS.

Coordinated genome initiatives and the parallel rise of explicit reference quality management are shifting both the scale and reliability of available resources. As coverage expands and assembly quality improves, AFS is expected to yield fewer unclassified reads, reduced systematic misassignment to nearest-neighbour genomes, and more stable quantitative estimates. Ongoing expansion in reference coverage and continued maturation of quality-management practices are expected to progressively shift AFS from a method limited by references toward one primarily limited by biological signal characteristics and sample complexity.

3.3.2 From read counts to multi-layer genomic evidence: resolving close-relative ambiguity

Because eukaryotic genomes contain extensive conserved and repetitive sequence, computational read partitioning across multiple candidate references inevitably produces systematic low-abundance false-positives. To disentangle true multi-species signals from purely bioinformatic ambiguity, this thesis therefore processed *in silico* datasets generated from single species, 100 % samples and processed them with the AFS pipeline. Despite the absence of any second species in the input, low-level assignments to closely related genomes consistently emerged as systematic read cross-assignment. Simulations showed that false-positive assignments scale approximately linearly with the true proportion of the present ingredient (Hellmann, Ripp, et al., 2020).

As an additional, independent plausibility check, genome-wide coverage profiles (read depth along the reference) can be inspected to discriminate true-positives from false

classifications: genuine organisms typically exhibit broadly distributed, comparatively even coverage across their reference genome, whereas false-positive matches driven by conserved regions shared between genomes tend to produce sparse, highly localized coverage peaks (Figure 12). An example is provided by the microbiota signal in the *KLyo C* sample, where *Brochothrix thermosphacta* shows an even coverage pattern consistent with a true-positive, while *Actinoalloteichus sp.* displays a highly uneven profile with distinct peaks consistent with a false-positive assignment. This criterion can only be applied however, when sequencing depth is sufficient to cover a substantial fraction of the candidate genome. For illustration, rice (*Oryza sativa*) has a genome size of 387 Mb, which is comparatively small among food-relevant eukaryotic genomes. Even modest genome-wide depth of 5× mean coverage would require >12 million 150 bp reads for rice alone. In routine AFS, by contrast, typically less than 5 million reads are generated for an entire sample and are additionally partitioned across multiple ingredients, such that eukaryotic ingredients rarely reach the genome-wide depth required for coverage-profile-based discrimination. Consequently, coverage-based true-/false-positive discrimination is primarily practical for microorganisms with genomes on the order of only a few Mb, where comparable read numbers cover the genome sufficiently.

Where coverage-based diagnostics are impossible, ambiguity at low abundance must be addressed by locus-restricted evidence rather than by genome-wide read-depth patterns. The untargeted WGS approach inherently provides additional information that can be utilized in those taxon complexes where proportion estimates are limited by sequence similarity and cross-assignment. A plausible strategy is therefore a marker-based confirmation that tests for discriminative features expected only under true multi-species content. Genome-wide single nucleotide polymorphisms (SNPs) are abundant, stable markers that can support cultivar/breed discrimination and provenance inference when representative population reference panels exist (Vignal et al., 2002). In food authentication, this is particularly relevant for high-value ingredients

in which the analytical objective often shifts from species identification to inference of variety/breed and population background, as illustrated by cacao origin assignment using population-specific genetic signatures (Bhattacharjee et al., 2023). Importantly, the same diagnostic SNP logic can be repurposed to resolve sister-species pairs that remain ambiguous under genome-wide analysis. Conceptually, SNP-based inference in AFS might be more robust for distinguishing close relatives than whole-genome analysis, because it can be restricted to explicitly diagnostic loci rather than relying on all sequence, including conserved regions that drive ambiguity. However, reliable SNP-based assignment typically requires sufficient coverage at informative loci and population-aware reference datasets, which currently do not exist for all food relevant species pairs.

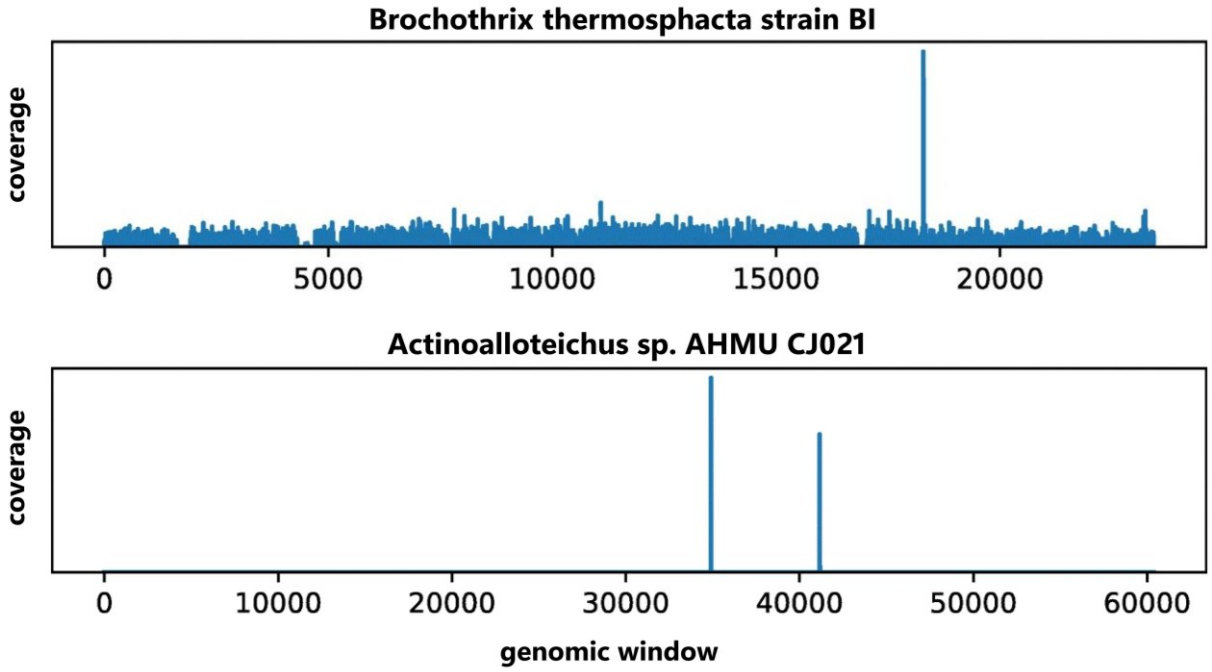

Figure 12: Genome-wide coverage profiles to distinguish true-positives from bioinformatic misassignment. Read depth per genomic window is shown along each reference. *Brochothrix thermosphacta* (top) exhibits distributed, comparatively even coverage across the genome, consistent with a genuine presence, whereas *Actinoalloteichus sp.* (bottom) shows near-zero background coverage with a few sharp peaks only, indicative of a false-positive assignment driven by conserved regions (adapted from Kobus et al., 2020).

Highly polymorphic repeat-based markers provide a complementary method. Simple sequence repeats (SSR) or short tandem repeats (1-6 bp repeats; STR) mutate

predominantly via repeat-number changes and thus exhibit high polymorphism among individuals and populations, enabling compact genetic fingerprinting panels for traceability and breed/cultivar assignment (Ellegren, 2004; Vieira et al., 2016). For example, multiplex SSR genotyping for cattle-breed traceability reported 180 alleles across 16 loci in six breeds; using the six most polymorphic loci, differentiation between all breeds was achievable, while the full 16-locus panel provided individual-level resolution (J. Zhao et al., 2017). Because WGS enables genome-wide SSR discovery, AFS reads can, in principle, be mapped to pre-validated loci to challenge closely related species calls and to flag assignments that lack marker support.

Taxon-specific repeats and Transposable Element (TE) insertions add a further presence/absence-style marker layer. As many TE families undergo lineage- or species-specific expansions, insertion patterns can serve as sensitive "fingerprints" when targets are screened (Konkel et al., 2010; Kramerov & Vassetzky, 2005). This principle is routinely exploited in targeted assays; for instance, a porcine SINE-based qPCR assay combines high copy number (sensitivity) with species specificity through locus screening (C. Zhang et al., 2015). In AFS, this information can be used for screening against e.g. an species-specific repeatome database, to confirm true-positive and identify false-positive classifications.

Collectively, these marker layers (SNPs, SSR/STRs, taxon-specific repeats and TE insertions) can operationally function as an orthogonal confirmatory tier in AFS analysis: in case of low-level close-relative assignments, explicit tests for discriminative features can be applied. This could include (i) defining diagnostic SNP panels for sister-species pairs and evaluating allele support at these loci, (ii) exploiting highly variable SSR/STR loci as compact fingerprints whose concordant allele patterns would be difficult to explain by diffuse mapping ambiguity alone, and (iii) using screened taxon-specific repeats and TE insertion targets as sensitive presence/absence evidence. Operationally, such marker layers would not replace AFS, but could act as an additional layer of information on species pairs where genome-wide read classification has a

limitation based on sequence similarity. However, the extent to which these approaches improve specificity without sensitivity losses, how they behave across matrices, processing states, and mixture complexities, needs targeted validation before they can be considered reliable components of an AFS decision framework.

3.3.3 Beyond sequence: modification as decision support for AFS

Beyond nucleotide sequence, long-read platforms (PacBio and ONT) provide the capability to detect DNA base modifications directly, without prior chemical treatment. These techniques thereby enable exploitation of epigenetic information as an additional layer in modification-enhanced AFS. DNA methylation, predominantly 5-methylcytosine (5mC) at CpG dinucleotides, is central to genome integrity and transposable-element silencing (Bird, 2002; Deniz et al., 2019; Yoder et al., 1997). Furthermore, DNA methylation plays a pivotal role in the regulation of numerous biological processes such as cell differentiation (Reik et al., 2001), development (Z. D. Smith & Meissner, 2013), and the maintenance of cellular identity (Fisher, 2002). Additionally, this epigenetic mechanism has been implicated in an organism's response to environmental influences (Feil & Fraga, 2012) and in speciation (Jablonka & Raz, 2009). In addition to differences between species, every tissue and cell type within a species possesses a unique methylation exhibiting unique differentially methylated regions (DMRs) specific to these cells (Drouilhet et al., 2022; Rivera & Ren, 2013; Roadmap Epigenomics Consortium et al., 2015; Zhou et al., 2020). Numerous studies have shown that samples of the same tissue type from different individuals tend to cluster more closely than those of different tissues from the same individual, showing that DNA methylation is tissue-specific (Baker et al., 2023; Drouilhet et al., 2022; Ju et al., 2023; A. T. Lu et al., 2023; Schultz et al., 2015). As these DMRs represent an epigenetic fingerprint of an individual tissue of a species, they can in principle also be utilized as biomarkers for discrimination of even closely related species in food matrices.

Proof-of-concept for leveraging methylation as a discriminative signal has been demonstrated outside the food domain. A human array-based methylation assay was adapted to non-human primates and mouse and showed that genome-wide methylation profiles are highly reproducible and can separate samples by biological class, distinguishing tissue types within and between species; in the same study, 5-hydroxymethylcytosine (5hmC) measurements in brain tissue revealed species-specific levels with a consistent hierarchy (human > rhesus ≫ mouse), underscoring that modified-base landscapes can encode robust species-dependent patterns (Chopra et al., 2014). Comparable conclusions arise from evolutionary plant epigenomics, where whole-genome methylomes across *Brassicaceae* showed that interspecific methylation differences are strongly influenced by genome organization, in particular by lineage-specific expansion or contraction of repeats and transposable elements; notably, many between-species differences cluster in hypervariable, repeat-rich regions (Seymour et al., 2014), which can diverge rapidly, even among relatively closely related taxa.

In food-related applications, the most robust progress to date has primarily concerned methylation-derived fingerprints for tracing geographic origin, whereas species-level identification was of secondary interest. Single-base resolution methylomes compiled across different livestock have demonstrated separable, context-associated signatures, including population- or location-associated patterns in clonal marbled crayfish, producer-associated signatures in shrimp, river-origin and rearing-environment associations in salmon, and farm-dependent patterns in chicken (Venkatesh et al., 2023). Collectively, these results support the premise that methylation carries structured, application-relevant information for traceability. At the same time, partial environmental and time dependence implies that routine deployment would require control of confounders and appropriate reference designs.

For modification-enhanced AFS, a concrete implementation strategy would be the construction of curated reference databases of methylation signatures for multiple food-relevant tissues per species. Such reference methylomes would capture the

distribution of 5mC across CpG-rich regulatory regions, repetitive elements, and other genomic domains and could draw on public resources like MethBank and NGSmethDB (Lebrón et al., 2017; R. Li et al., 2017) and on experimental methylome datasets (Klughammer et al., 2023; Meissner et al., 2005), while likely requiring targeted curation for the specific sister-species pairs that are problematic at the sequence level. Operationally, a two-fold logic is plausible: (i) conventional AFS read classification provides primary assignments and estimates, and (ii) for ambiguous near-relative assignments, methylation profiles derived from long reads are queried against the reference methylation signature database to test whether diagnostic methylation patterns support the presence of the candidate species. This approach is conceptually attractive because it introduces an orthogonal signal that is not identical to nucleotide sequence similarity, and could therefore add specificity in cases where sequence-level discrimination is ambiguous.

Beyond species discrimination, DMRs further offer a potential tissue-identity layer that could address matrix-effect-driven quantification biases. If tissue identity could be inferred from methylation fingerprints, organ-specific differences in DNA yield (e.g., lower yields in adipose relative to muscle) could, in principle, be incorporated into downstream interpretation to mitigate tissue-driven distortions of proportion estimates. A proof-of-principle has shown that tissue identity can be inferred from methylation fingerprints using methylation-sensitive profiling, enabling robust discrimination of multiple organs in both salmon and cattle; specifically, methylation-sensitive amplified polymorphism (a restriction-fragment-based approach conceptually related to RFLP) supported discrimination of brain, eye, heart, kidney, liver, and muscle, indicating that between-organ methylation differences can be sufficiently strong to act as diagnostic signals (Rodríguez López et al., 2012). Accordingly, modification-enhanced AFS could, in principle, provide not only species evidence but also information on which tissues contributed DNA to a processed sample. Complementing epigenetic fingerprints potentially provide the necessary

information for the identification of the tissues present in a sample. In turn, with this information available, it could in principle be possible to correct the differences in DNA amounts per weight of tissue, which would serve as an invaluable improvement towards solving quantification issues caused by matrix effects.

3.3.4 Perspective and near-term roles for AFS

AFS provides an untargeted, primer-independent view of the DNA present in a sample and thereby avoids primer-driven amplification bias. In a single workflow, AFS analyses nuclear and organellar DNA alongside microbial and viral nucleic acids, enabling simultaneous screening of animals, plants, fungi, protists, bacteria, archaea, and viruses and facilitates the discovery of undeclared ingredients and potential contaminants.

Despite this breadth, three factors currently limit large-scale deployment of AFS for routine screening. First, AFS is reference-dependent. Species can only be identified and quantified if suitable genomes are present in the database; missing references increase unclassified reads and misassignment to close relatives and may require qualitative follow-up. Second, per-sample costs remain higher in many settings, as WGS approaches typically require more sequencing depth than marker-based metabarcoding. Third, routine use requires bioinformatic capacity for standardised processing, quality control, and interpretation.

However, these constraints are dynamic rather than fundamental. Reference collections are expanding rapidly and assembly quality continues to improve; sequencing costs are declining; and laboratory protocols and pipelines are becoming more efficient, collectively reducing database- and cost-related bottlenecks. The bioinformatic requirements are broadly comparable to those already established for metabarcoding in many laboratories and can be further enhanced by training programmes for regulatory laboratories and competent authorities. Interpretation, however, will

continue to require conservative decision rules: read fractions are not direct biomass proxies because matrix effects persist, a limitation that also affects other DNA-based assays. As reference resources become more comprehensive and sequencing becomes more affordable, the balance is likely to shift further towards AFS as a scalable option for broad food surveillance.

A pragmatic near-term implementation is therefore a staged evidence workflow in which AFS provides broad, untargeted screening, complemented by orthogonal confirmation where consequences warrant it. Challenging authenticity questions are most robustly addressed by converging evidence from complementary strategies, so that method-specific weaknesses are compensated and the resulting conclusions remain defensible under scrutiny.

4 Summary

All-food-sequencing (AFS) is a non-targeted, whole-genome DNA-based screening method for simultaneous qualitative and quantitative species diagnosis in complex foods. No a priori knowledge or primers are required, and a single analysis can potentially detect animal and plant components as well as fungi, bacteria, and other accompanying microbiota.

Quantitative performance of AFS was evaluated using controlled calibration samples, including defined meat and fish mixtures. In direct comparison with established assays, AFS delivered quantification that outperformed conventional qPCR and matched ddPCR, while retaining a key practical advantage: broad, primer-independent screening and quantification of multiple ingredients within a single workflow, rather than many separate target specific assays. The calibration series also identified two systematic factors that influence quantitative accuracy. First, differences in genome size can shift read proportions, but this bias was shown to be correctable by genome size normalization. Second, matrix effects of the food composition can alter DNA extraction yield between ingredients, so the highest quantitative agreement is achieved with matrix-specific calibration - a constraint shared by all DNA-based methods.

Application for real food products showed how AFS behaves under practical conditions with regard to heterogeneity and incomplete or inaccurate labelling of species components. Doner kebab samples showed significant deviations and several cases in which the predominant meat type did not correspond to the advertised expectation. In addition, other ingredients were detected that suggest unintentional admixture or deliberate substitution. In the case of seafood and surimi products, the declared main ingredients were generally confirmed. In addition, the analysis revealed undeclared taxa, including additional seafood and plant components in mixed dishes, some with potential allergen relevance. Beyond the main ingredient composition, AFS provided

early warning signals, including allergen-relevant plant admixtures and microbial patterns that indicate incipient spoilage.

The analysed samples also revealed practical limitations of AFS that are crucial for regulatory surveillance: ambiguous classification within closely related species, dependence on reference genomes, and the need to interpret low-level findings conservatively and verify them with targeted follow-up measures. Specificity of read classification was improved by both algorithm choice and sequencing technology. K-mer-based classification and database partitioning enabled screening against significantly larger genome reference collections. In a comparison between Illumina short reads and Oxford Nanopore long reads on calibration sausages, long-read sequencing improved quantification accuracy and reduced the number of false-positives despite higher error rates in long read, as longer reads provide more discriminative information for resolving conserved regions that otherwise drive ambiguous classification.

AFS shows the potential to provide reliable formulation information and cross-domain early warning signals as a universal, primer-free WGS screening method, provided that reference dependency and taxonomic ambiguity are recognized as current limitations and integrated into the decision on verification or follow-up measures. Therefore, it is promising for routine screening in official food monitoring in the future.

5 Zusammenfassung

All-Food-Sequencing (AFS) ist ein ungezieltes, gesamt-genomisches DNA-basiertes Screeningverfahren zur gleichzeitigen qualitativen und quantitativen Artendiagnose in komplexen Lebensmitteln. Es sind weder *a priori* Kenntnisse noch Primer erforderlich und eine einzige Analyse kann potenziell tierische und pflanzliche Bestandteile ebenso wie Pilze, Bakterien und weitere begleitende Mikrobiota erfassen.

Die quantitative Leistung der AFS wurde anhand kontrollierter Kalibrierungsproben bewertet, darunter definierte Fleisch- und Fischmischungen. Im direkten Vergleich mit etablierten Verfahren erzielte AFS eine Quantifizierung, die herkömmliche qPCR übertraf und mit ddPCR gleichauf lag, mit einem wichtigen praktischen Vorteil: ein breites, Primer-unabhängiges Screening und eine Quantifizierung mehrerer Inhaltsstoffe in einer einzigen Analyse anstelle vieler separater, spezifischer Assays. Die Kalibrierungsserie identifizierte auch zwei systematische Faktoren, die die quantitative Genauigkeit beeinflussen. Zum einen können Unterschiede in der Genomgröße die Read-Anteile systematisch verschieben; aber diese Verzerrung erwies sich durch Normalisierung der Genomgröße als korrigierbar. Zum anderen können Matrixeffekte der Lebensmittelzusammensetzung die DNA-Extraktionsausbeute zwischen den Inhaltsstoffen verändern, sodass eine exakte Quantifizierung von einer matrixspezifischen Kalibrierung profitiert – eine Einschränkung, die alle DNA-basierten Methoden gemeinsam haben.

Die Anwendung mit realen Lebensmittelprodukten zeigte, wie sich AFS unter praxisnahen Bedingungen hinsichtlich Heterogenität und unvollständiger oder ungenauer Kennzeichnung der Zutaten verhält. Die Döner Kebab-Proben wiesen erhebliche Abweichungen und mehrere Fälle auf, in denen die vorherrschende Fleischsorte nicht der beworbenen Erwartung entsprach. Zusätzlich wurden weitere Bestandteile nachgewiesen, die den Verdacht unbeabsichtigter Beimischung oder gezielter Substitution nahelegen. Bei den Meeresfrüchte- und Surimi-Produkten

wurden die deklarierten Hauptbestandteile im Allgemeinen bestätigten. Darüber hinaus machte die Analyse wiederholt nicht deklarierte Taxa sichtbar, darunter zusätzliche Seafood-Anteile in Mischgerichten sowie pflanzliche Signale mit potenzieller Allergenrelevanz. Über die reine Rezeptur hinaus lieferte AFS noch zusätzliche Hinweise. Es wurden Spurenbefunde detektiert, die als Frühwarnsignale dienen können, darunter allergenrelevante Pflanzenbeimischungen sowie mikrobielle Muster, die auf beginnenden Verderb hindeuten.

Diese Proben zeigten auch praktische Limitierungen der Methode auf, die für eine behördliche Überwachung entscheidend sind: uneindeutige Zuordnung innerhalb nah verwandter Spezies, die Abhängigkeit von Referenzgenomen sowie die Notwendigkeit, Befunde im Spurenbereich konservativ zu interpretieren und gezielt mit Folgemaßnahmen zu verifizieren.

Die Spezifität der Read-Klassifizierung wurde dabei sowohl durch die Wahl des Algorithmus als auch durch die Sequenzierungstechnologie verbessert. Die K-mer-basierte Klassifizierung und Partitionierung der Datenbank ermöglichte ein Screening gegen deutlich größere genomische Referenzsammlungen. Im Vergleich zwischen Illumina Short-Reads und Oxford Nanopore Long-Reads bei Kalibrierungswürsten verbesserte die Long-Read-Sequenzierung die Quantifizierungsgenauigkeit und reduzierte die Anzahl falsch-positiver Befunde trotz höherer Fehlerraten bei den Long-Reads, da längere Reads mehr diagnostische Unterschiede für die Auflösung konservierter Regionen liefern, die andernfalls zu einer mehrdeutigen Klassifizierung führen würden.

AFS zeigt das Potenzial, als universelles, Primer-freies WGS-Screening sowohl verlässliche Rezepturhinweise als auch domänenübergreifende Frühwarnsignale zu liefern, sofern Referenzabhängigkeit und taxonomische Mehrdeutigkeit als gegenwärtig vorhandene Limitierungen anerkannt und in die Entscheidung über

Verifikation oder Folgemaßnahmen integriert werden, wodurch es künftig für Routine-Screenings in der amtlichen Lebensmittelüberwachung interessant werden kann.

References

Adenuga, B. M., Spychaj, A., & Montowska, M. (2025). A new nuclear marker for quantitative analysis of wild boar and domestic pig meat in game meat products using PLAG1 zinc finger gene. *Scientific Reports 2025 15:1*, *15*(1), 20454-. https://doi.org/10.1038/s41598-025-05167-x

Afzaal, M., Saeed, F., Hussain, M., Shahid, F., Siddeeg, A., & Al-Farga, A. (2022). Proteomics as a promising biomarker in food authentication, quality and safety: A review. *Food Science & Nutrition*, *10*(7), 2333. https://doi.org/10.1002/FSN3.2842

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

Albertsen, M. (2023). Long-read metagenomics paves the way toward a complete microbial tree of life. *Nature Methods 2023 20:1*, *20*(1), 30–31. https://doi.org/10.1038/s41592-022-01726-6

Allio, R., Donega, S., Galtier, N., & Nabholz, B. (2017). Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Molecular Biology and Evolution*, *34*(11), 2762–2772. https://doi.org/10.1093/MOLBEV/MSX197

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Andeta, A. F., Vandeweyer, D., Woldesenbet, F., Eshetu, F., Hailemicael, A., Woldeyes, F., Crauwels, S., Lievens, B., Ceusters, J., Vancampenhout, K., & Van Campenhout, L. (2018). Fermentation of enset (Ensete ventricosum) in the Gamo highlands of Ethiopia: Physicochemical and microbial community dynamics. *Food Microbiology*, *73*, 342–350. https://doi.org/10.1016/J.FM.2018.02.011

Andreani, N. A., & Fasolato, L. (2017). Pseudomonas and Related Genera. *The Microbiological Quality of Food: Foodborne Spoilers*, 25–59. https://doi.org/10.1016/B978-0-08-100502-6.00005-4

Andronache, J., Cichna-Markl, M., Dobrovolny, S., & Hochegger, R. (2025). Development of a DNA Metabarcoding Method for the Identification of Crustaceans (Malacostraca) and Cephalopods (Coleoidea) in Processed Foods. *Foods*, *14*(9), 1549. https://doi.org/10.3390/FOODS14091549/S1

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). https://doi.org/10.1093/GIGASCIENCE/GIX080

Asalone, K. C., Ryan, K. M., Yamadi, M., Cohen, A. L., Farmer, W. G., George, D. J., Joppert, C., Kim, K., Mughal, M. F., Said, R., Toksoz-Exley, M., Bisk, E., & Bracht, J. R. (2020). Regional sequence expansion or collapse in heterozygous genome assemblies. *PLOS Computational Biology*, *16*(7), e1008104. https://doi.org/10.1371/JOURNAL.PCBI.1008104

Baker, E. C., San, A. E., Cilkiz, K. Z., Littlejohn, B. P., Cardoso, R. C., Ghaffari, N., Long, C. R., Riggs, P. K., Randel, R. D., Welsh, T. H., & Riley, D. G. (2023). Inter-Individual Variation in DNA Methylation Patterns across Two Tissues and Leukocytes in Mature Brahman Cattle. *Biology*, *12*(2), 252. https://doi.org/10.3390/BIOLOGY12020252

Balkir, P., Kemahlioglu, K., & Yucel, U. (2021). Foodomics: A new approach in food quality and safety. *Trends in Food Science & Technology*, *108*, 49–57. https://doi.org/10.1016/J.TIFS.2020.11.028

Batovska, J., Cogan, N. O. I., Lynch, S. E., & Blacket, M. J. (2017). Using next-generation sequencing for DNA barcoding: Capturing allelic variation in ITS2. *G3: Genes, Genomes, Genetics*, *7*(1), 19–29. https://doi.org/10.1534/G3.116.036145

Bell, K. L., Petit, R. A., Cutler, A., Dobbs, E. K., Macpherson, J. M., Read, T. D., Burgess, K. S., & Brosi, B. J. (2021). Comparing whole-genome shotgun sequencing and DNA metabarcoding approaches for species identification and quantification of pollen species mixtures. *Ecology and Evolution*, *11*(22), 16082–16098. https://doi.org/10.1002/ECE3.8281

Bhattacharjee, R., Luseni, M. M., Agre, P. A., Lava Kumar, P., & Grenville-Briggs, L. J. (2023). *Genetic diversity and population structure of cacao (Theobroma cacao L.) germplasm from Sierra Leone and Togo based on KASP- SNP genotyping*. https://doi.org/10.21203/RS.3.RS-3345739/V1

Billington, C., Kingsbury, J. M., & Rivas, L. (2022). Metagenomics Approaches for Improving Food Safety: A Review. *Journal of Food Protection*, *85*(3), 448–464. https://doi.org/10.4315/JFP-21-301

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes Dev.*, *16*(1), 6–21. https://doi.org/10.1101/gad.947102

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *359*(1444), 669. https://doi.org/10.1098/RSTB.2003.1447

Blaxter, M. L. (2022). Sequence locally, think globally: The Darwin Tree of Life Project. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4), e2115642118. https://doi.org/10.1073/PNAS.2115642118

Blaxter, M., Lewin, H. A., DiPalma, F., Challis, R., da Silva, M., Durbin, R., Formenti, G., Franz, N., Guigo, R., Harrison, P. W., Hiller, M., Hoff, K. J., Howe, K., Jarvis, E. D., N Lawniczak, M. K., Lindblad-Toh, K., H Mathews, D. J., Martin, F. J., Mazzoni, C. J., … Silva, da. (2025). The Earth BioGenome Project Phase II: illuminating the eukaryotic tree of life. *Frontiers in Science*, *3*, 1514835. https://doi.org/10.3389/FSCI.2025.1514835

Bohling, J. (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution*, *10*(14), 7585–7601. https://doi.org/10.1002/ECE3.6483

Böhme, K., Calo-Mata, P., Barros-Velázquez, J., & Ortea, I. (2019). Review of Recent DNA-Based Methods for Main Food-Authentication Topics. *Journal of Agricultural and Food Chemistry*, *67*(14), 3854–3864. https://doi.org/10.1021/ACS.JAFC.8B07016

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/BIOINFORMATICS/BTU170

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, *27*(8), 1767. https://doi.org/10.1093/NAR/27.8.1767

Breitwieser, F. P., Pertea, M., Zimin, A. V., & Salzberg, S. L. (2019). Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Research*, *29*(6), 954–960. https://doi.org/10.1101/GR.245373.118

Bruno, A., Sandionigi, A., Agostinetto, G., Bernabovi, L., Frigerio, J., Casiraghi, M., & Labra, M. (2019). Food Tracking Perspective: DNA Metabarcoding to Identify Plant Composition in Complex and Processed Food Products. *Genes 2019, Vol. 10, Page 248*, *10*(3), 248. https://doi.org/10.3390/GENES10030248

Buljevíc, N., Preiner, D., Šikuten, I., & Tomaz, I. (2025). Comparison of Different DNA Isolation Methods from Grapevine (Vitis vinifera) Leaves. *Separations 2025, Vol. 12, Page 316*, *12*(11), 316. https://doi.org/10.3390/SEPARATIONS12110316

Bundesamt für Verbraucherschutz und Lebensmittelsicherheit. (2014). Amtliche Sammlung von Untersuchungsverfahren nach § 64 LFGB, § 35 Vorl. Tabakgesetz, § 28 b Gentechnikgesetz. *Beuth Verlag GmbH*.

Burger, R. (2012). *EHEC O104:H4 in Germany 2011:* Large outbreak of bloody diarrhea and haemolytic uraemic syndrome by shiga toxin-producing E. coli via contaminated food. https://doi.org/10.25646/1235

Byappanahalli, M. N., Nevers, M. B., Korajkic, A., Staley, Z. R., & Harwood, V. J. (2012). Enterococci in the environment. *Microbiology and Molecular Biology Reviews : MMBR*, *76*(4), 685–706. https://doi.org/10.1128/MMBR.00023-12

Cai, Y., Li, X., Lv, R., Yang, J., Li, J., He, Y., & Pan, L. (2014). Quantitative Analysis of Pork and Chicken Products by Droplet Digital PCR. *BioMed Research International*, *2014*. https://doi.org/10.1155/2014/810209

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*. https://doi.org/10.1186/1471-2105-10-421

Cankar, K., Stebih, D., Dreo, T., Žel, J., & Gruden, K. (2006). Critical points of DNA quantification by real-time PCR - Effects of DNA extraction method and sample matrix on quantification of genetically modified organisms. *BMC Biotechnology*, *6*(1), 1–15. https://doi.org/10.1186/1472-6750-6-37

Centre for Biodiversity Genomics, U. of G. (2021). Secretariat of the Convention on Biological Diversity The Global Taxonomy Initiative 2020: A Step-by-Step Guide for DNA Barcoding. *Technical Series No. 94. Secretariat of the Convention on Biological Diversity, Montreal, 66 Pages*.

Cermakova, E., Lencova, S., Mukherjee, S., Horka, P., Vobruba, S., Demnerova, K., & Zdenkova, K. (2023). Identification of Fish Species and Targeted Genetic Modifications Based on DNA Analysis: State of the Art. *Foods*, *12*(1), 228. https://doi.org/10.3390/FOODS12010228/S1

Chang, C. H., Lin, H. Y., Ren, Q., Lin, Y. S., & Shao, K. T. (2016). DNA barcode identification of fish products in Taiwan: Government-commissioned authentication cases. *Food Control*, *66*, 38–43. https://doi.org/10.1016/J.FOODCONT.2016.01.034

Chopra, P., Papale, L. A., White, A. T. J., Hatch, A., Brown, R. M., Garthwaite, M. A., Roseboom, P. H., Golos, T. G., Warren, S. T., & Alisch, R. S. (2014). Array-based assay detects genome-wide 5-mC and 5-hmC in the brains of humans, non-human primates, and mice. *BMC Genomics*, *15*(1), 131-. https://doi.org/10.1186/1471-2164-15-131

Costello, M., Fleharty, M., Abreu, J., Farjoun, Y., Ferriera, S., Holmes, L., Granger, B., Green, L., Howd, T., Mason, T., Vicente, G., Dasilva, M., Brodeur, W., DeSmet, T., Dodge, S., Lennon, N. J., & Gabriel, S. (2018). Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics*, *19*(1), 332-. https://doi.org/10.1186/S12864-018-4703-0

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*, 123–132. https://doi.org/10.1016/J.TRAC.2016.02.026

Dawan, J., & Ahn, J. (2022). Application of DNA barcoding for ensuring food safety and quality. *Food Science and Biotechnology*, *31*(11), 1355. https://doi.org/10.1007/S10068-022-01143-7

De Coster, W., De Rijk, P., De Roeck, A., De Pooter, T., D'Hert, S., Strazisar, M., Sleegers, K., & Van Broeckhoven, C. (2019). Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Research*, *29*(7), 1178–1187. https://doi.org/10.1101/GR.244939.118

Deniz, Ö., Frost, J. M., & Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews. Genetics*, *20*(7), 417–431. https://doi.org/10.1038/S41576-019-0106-6

Di Pinto, A., Bottaro, M., Bonerba, E., Bozzo, G., Ceci, E., Marchetti, P., Mottola, A., & Tantillo, G. (2015). Occurrence of mislabeling in meat products using DNA-based assay. *Journal of Food Science and Technology*, *52*(4), 2479–2484. https://doi.org/10.1007/S13197-014-1552-Y

Dobrovolny, S., Uhlig, S., Frost, K., Schlierf, A., Nichani, K., Simon, K., Cichna-markl, M., & Hochegger, R. (2022). Interlaboratory Validation of a DNA Metabarcoding Assay for Mammalian and Poultry Species to Detect Food Adulteration. *Foods*, *11*(8), 1108. https://doi.org/10.3390/FOODS11081108

Dolch, K., Andrée, S., & Schwägele, F. (2020). Comparison of Real-Time PCR Quantification Methods in the Identification of Poultry Species in Meat Products. *Foods 2020, Vol. 9, Page 1049*, *9*(8), 1049. https://doi.org/10.3390/FOODS9081049

Drouilhet, L., Moreno, C., Plisson-Petit, F., Marcon, D., Fabre, S., & Hazard, D. (2022). Variability in Global DNA Methylation Rate Across Tissues and Over Time in Sheep. *Frontiers in Genetics*, *13*, 791283. https://doi.org/10.3389/FGENE.2022.791283

Ebbert, M. T. W., Jensen, T. D., Jansen-West, K., Sens, J. P., Reddy, J. S., Ridge, P. G., Kauwe, J. S. K., Belzil, V., Pregent, L., Carrasquillo, M. M., Keene, D., Larson, E., Crane, P., Asmann, Y. W., Ertekin-Taner, N., Younkin, S. G., Ross, O. A., Rademakers, R., Petrucelli, L., & Fryer, J. D. (2019). Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biology*, *20*(1). https://doi.org/10.1186/S13059-019-1707-2

Ebenezer, T. G. E., Muigai, A. W. T., Nouala, S., Badaoui, B., Blaxter, M., Buddie, A. G., Jarvis, E. D., Korlach, J., Kuja, J. O., Lewin, H. A., Majewska, R., Mapholi, N., Maslamoney, S., Mbo'o-Tchouawou, M., Osuji, J. O., Seehausen, O., Shorinola, O., Tiambo, C. K., Mulder, N., … Djikeng, A. (2022). Africa: sequence 100,000 species to safeguard biodiversity. *Nature 2022 603:7901*, *603*(7901), 388–392. https://doi.org/10.1038/d41586-022-00712-4

Edger, P. P., Poorten, T. J., VanBuren, R., Hardigan, M. A., Colle, M., McKain, M. R., Smith, R. D., Teresi, S. J., Nelson, A. D. L., Wai, C. M., Alger, E. I., Bird, K. A., Yocca, A. E., Pumplin, N., Ou, S., Ben-Zvi, G., Brodt, A., Baruch, K., Swale, T., … Knapp, S. J. (2019). Origin and evolution of the octoploid strawberry genome. *Nature Genetics 2019 51:3*, *51*(3), 541–547. https://doi.org/10.1038/s41588-019-0356-4

Egerton, S., Culloty, S., Whooley, J., Stanton, C., & Ross, R. P. (2018). The gut microbiota of marine fish. *Frontiers in Microbiology*, *9* (MAY), 343795. https://doi.org/10.3389/FMICB.2018.00873

Elbrecht, V., Braukmann, T. W. A., Ivanova, N. V., Prosser, S. W. J., Hajibabaei, M., Wright, M., Zakharov, E. V., Hebert, P. D. N., & Steinke, D. (2019). Validation of COI metabarcoding primers for terrestrial arthropods. *PeerJ*, *7*(10). https://doi.org/10.7717/PEERJ.7745

Ellegren, H. (2004). Microsatellites: simple sequences with complex evolution. *Nature Reviews. Genetics*, *5*(6), 435–445. https://doi.org/10.1038/NRG1348

Espinosa, E., Bautista, R., Larrosa, R., & Plata, O. (2024). Advancements in long-read genome sequencing technologies and algorithms. *Genomics*, *116*(3), 110842. https://doi.org/10.1016/J.YGENO.2024.110842

Esser, M., Brinkmann, M., & Hecker, M. (2024). Solving freshwater conservation challenges through next-generation sequencing approaches. *Environmental Science: Advances*, *3*(9), 1181–1196. https://doi.org/10.1039/D4VA00112E

Eugster, A., Ruf, J., Rentsch, J., & Köppel, R. (2009). Quantification of beef, pork, chicken and turkey proportions in sausages: Use of matrix-adapted standards and comparison of single versus multiplex PCR in an interlaboratory trial. *European Food Research and Technology*, *230*(1), 55–61. https://doi.org/10.1007/s00217-009-1138-5

European Food Safety Authority. (2009). The use and mode of action of bacteriophages in food production - Endorsed for public consultation 22 January 2009 - Public consultation 30 January – 6 March 2009. *EFSA Journal*, *7*(5), 1076. https://doi.org/10.2903/J.EFSA.2009.1076

European Food Safety Authority. (2016). Evaluation of the safety and efficacy of ListexTM P100 for reduction of pathogens on different ready-to-eat (RTE) food products. *EFSA Journal*, *14*(8), e04565. https://doi.org/10.2903/J.EFSA.2016.4565

European Parliament & Council. (2011). *REGULATION (EU) No 1169/2011*.

European Parliament & Council. (2013a). *2013/99/EU:* Outcome of the coordinated control plan with a view to establish the prevalence of fraudulent practices in the marketing of certain foods.

European Parliament & Council. (2013b). *2013/2091(INI):* REPORT on the food crisis, fraud in the food chain and the control thereof | A7-0434/2013 | European Parliament. https://www.europarl.europa.eu/doceo/document/A-7-2013-0434_EN.html

European Parliament & Council. (2023). *COMM(23)01788[1]*. https://food.ec.europa.eu/food-safety/eu-agri-food-fraud-network/eu-coordinated-actions/honey-2021-2022_en#about-the-eu-action

Europol. (2014). Thousands of tonnes of fake food and drink seized in Interpol-Europol operation | Europol. https://www.europol.europa.eu/media-press/newsroom/news/thousands-of-tonnes-of-fake-food-and-drink-seized-in-interpol-europol-operation

Everstine, K., Spink, J., & Kennedy, S. (2013). Economically Motivated Adulteration (EMA) of Food: Common Characteristics of EMA Incidents. *Journal of Food Protection*, *76*(4), 723–735. https://doi.org/10.4315/0362-028X.JFP-12-399

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/BIOINFORMATICS/BTW354

Fanelli, V., Mascio, I., Miazzi, M. M., Savoia, M. A., De Giovanni, C., & Montemurro, C. (2021). Molecular Approaches to Agri-Food Traceability and Authentication: An Updated Review. *Foods 2021, Vol. 10, Page 1644*, *10*(7), 1644. https://doi.org/10.3390/FOODS10071644

Federal Office of Consumer Protection and Food Safety. (2022). BVL - OPSON XI (2021/2022) – Fehldeklaration und Fremdwasserzusatz bei Fischen, Krebs- und Weichtieren im Fokus der Untersuchungen. https://www.bvl.bund.de/DE/Arbeitsbereiche/01_Lebensmittel/03_Verbraucher/16_Food_Fraud/06_OPSON_Operationen/OPSON-XI/OPSON_11_node.html

Feil, R., & Fraga, M. F. (2012). Epigenetics and the environment: emerging patterns and implications. *Nat. Rev. Genet.*, *13*(2), 97–109. https://doi.org/10.1038/nrg3142

Feldmann, F., Ardura, A., Blanco-Fernandez, C., & Garcia-Vazquez, E. (2021). DNA Analysis Detects Different Mislabeling Trend by Country in European Cod Fillets. *Foods 2021, Vol. 10, Page 1515*, *10*(7), 1515. https://doi.org/10.3390/FOODS10071515

Fernandes, T. J. R., Amaral, J. S., & Mafra, I. (2021). DNA barcode markers applied to seafood authentication: an updated review. *Critical Reviews in Food Science and Nutrition*, *61*(22), 3904–3935. https://doi.org/10.1080/10408398.2020.1811200

Ferreira, A. O., Azevedo, O. M., Barroso, C., Duarte, S., Egas, C., Fontes, J. T., Ré, P., Santos, A. M. P., & Costa, F. O. (2024). Multi-marker DNA metabarcoding for precise species identification in ichthyoplankton samples. *Scientific Reports 2024 14:1*, *14*(1), 1–14. https://doi.org/10.1038/s41598-024-69963-7

Filipa-Silva, A., Castro, R., Rebelo, M., Mota, M. J., Almeida, A., Valente, L. M. P., & Gomes, S. (2024). Enhancing the authenticity of animal by-products: harmonization of DNA extraction methods from novel ingredients. *Frontiers in Chemistry*, *12*, 1350433. https://doi.org/10.3389/FCHEM.2024.1350433

Filonzi, L., Ardenghi, A., Rontani, P. M., Voccia, A., Ferrari, C., Papa, R., Bellin, N., & Nonnis Marzano, F. (2023). Molecular Barcoding: A Tool to Guarantee Correct Seafood Labelling and Quality and Preserve the Conservation of Endangered Species. *Foods 2023, Vol. 12, Page 2420*, *12*(12), 2420. https://doi.org/10.3390/FOODS12122420

Fisher, A. G. (2002). Cellular identity and lineage choice. *Nat. Rev. Immunol.*, *2*(12), 977–982. https://doi.org/10.1038/nri958

Fontes, J. T., Katoh, K., Pires, R., Soares, P., & Costa, F. O. (2024). Benchmarking the discrimination power of commonly used markers and amplicons in marine fish (e)DNA (meta)barcoding. *Metabarcoding and Metagenomics 8: E128646*, *8*, e128646. https://doi.org/10.3897/MBMG.8.128646

Galimberti, A., De Mattia, F., Losa, A., Bruni, I., Federici, S., Casiraghi, M., Martellos, S., & Labra, M. (2013). DNA barcoding as a new tool for food traceability. *Food Research International*, *50*(1), 55–63. https://doi.org/10.1016/J.FOODRES.2012.09.036

Garrison, E., Sirén, J., Novak, A. M., Hickey, G., Eizenga, J. M., Dawson, E. T., Jones, W., Garg, S., Markello, C., Lin, M. F., Paten, B., & Durbin, R. (2018). Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nature Biotechnology*, *36*(9), 875–881. https://doi.org/10.1038/NBT.4227

Garvey, M. (2022). Bacteriophages and Food Production: Biocontrol and Bio-Preservation Options for Food Safety. *Antibiotics 2022, Vol. 11, Page 1324*, *11*(10), 1324. https://doi.org/10.3390/ANTIBIOTICS11101324

German Federal Institute for Risk Assessment. (2023). *World Food Safety Day: 600 million reasons for good kitchen hygiene*. World Food Safety Day: 600 Million Reasons for Good Kitchen Hygiene. https://www.bfr.bund.de/en/press_information/2023/10/world_food_safety_day__600_million_reasons_for_good_kitchen_hygiene-311338.html

German Federal Office for Agriculture and Food. (2025). *Fischerei - Verzeichnis der Handelsbezeichnungen für Erzeugnisse der Fischerei und Aquakultur (deutsch-lateinisch)*. https://www.ble.de/SharedDocs/Downloads/DE/Fischerei/Fischwirtschaft/HandelsbezeichnungDLat.html

Gildea, L., Ayariga, J. A., & Robertson, B. K. (2022). Bacteriophages as Biocontrol Agents in Livestock Food Production. *Microorganisms 2022, Vol. 10, Page 2126*, *10*(11), 2126. https://doi.org/10.3390/MICROORGANISMS10112126

Giusti, A., Malloggi, C., Magagna, G., Filipello, V., Armani, A., Lombardia dell, della, Romagna, E., Ubertini, B., & Correspondence Alice Giusti, I. (2024). Is the metabarcoding ripe enough to be applied to the authentication of foodstuff of animal origin? A systematic review. *Comprehensive Reviews in Food Science and Food Safety*, *23*(1), e13256. https://doi.org/10.1111/1541-4337.13256

Goldfarb, T., Kodali, V. K., Pujar, S., Brover, V., Robbertse, B., Farrell, C. M., Oh, D. H., Astashyn, A., Ermolaeva, O., Haddad, D., Hlavina, W., Hoffman, J., Jackson, J. D., Joardar, V. S., Kristensen, D., Masterson, P., McGarvey, K. M., McVeigh, R., Mozes, E., … Murphy, T. D. (2025). NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Research*, *53*(D1), D243–D257. https://doi.org/10.1093/NAR/GKAE1038

Gorini, T., Mezzasalma, V., Deligia, M., De Mattia, F., Campone, L., Labra, M., & Frigerio, J. (2023). Check Your Shopping Cart: DNA Barcoding and Mini-Barcoding for Food Authentication. *Foods*, *12*(12), 2392. https://doi.org/10.3390/FOODS12122392

Gottardi, D., Siroli, L., Vannini, L., Patrignani, F., & Lanciotti, R. (2021). Recovery and valorization of agri-food wastes and by-products using the non-conventional yeast Yarrowia lipolytica. *Trends in Food Science & Technology*, *115*, 74–86. https://doi.org/10.1016/J.TIFS.2021.06.025

Greer, G. G., & Dilts, B. D. (2002). Control of Brochothrix thermosphacta spoilage of pork adipose tissue using bacteriophages. *Journal of Food Protection*, *65*(5), 861–863. https://doi.org/10.4315/0362-028X-65.5.861

Gregory, T. R. (2025). *Animal Genome Size Database*.

Griffiths, A. M., Sotelo, C. G., Mendes, R., Pérez-Martín, R. I., Schröder, U., Shorten, M., Silva, H. A., Verrez-Bagnis, V., & Mariani, S. (2014). Current methods for seafood authenticity testing in Europe: Is there a need for harmonisation? *Food Control*, *45*, 95–100. https://doi.org/10.1016/J.FOODCONT.2014.04.020

Guevara, R., Gianotti, M., Roca, P., & Oliver, J. (2011). Age and sex-related changes in rat brain mitochondrial function. *Cellular Physiology and Biochemistry : International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology*, *27*(3–4), 201–206. https://doi.org/10.1159/000327945

Gui, S., Wei, W., Jiang, C., Luo, J., Chen, L., Wu, S., Li, W., Wang, Y., Li, S., Yang, N., Li, Q., Fernie, A. R., & Yan, J. (2022). A pan-Zea genome map for enhancing maize improvement. *Genome Biology*, *23*(1), 178. https://doi.org/10.1186/S13059-022-02742-7

Guo, N., Chen, H. X., Zhang, L. P., Zhang, J. Y., Yang, L. Y., & Li, L. (2020). Infection and molecular identification of ascaridoid nematodes from the important marine food fish Japanese threadfin bream Nemipterus japonicus (Bloch) (Perciformes: Nemipteridae) in China. *Infection, Genetics and Evolution*, *85*, 104562. https://doi.org/10.1016/J.MEEGID.2020.104562

Haider, N., Nabulsi, I., & Al-Safadi, B. (2012). Identification of meat species by PCR-RFLP of the mitochondrial COI gene. *Meat Science*, *90*(2), 490–493. https://doi.org/10.1016/J.MEATSCI.2011.09.013

Haiminen, N., Edlund, S., Chambliss, D., Kunitomi, M., Weimer, B. C., Ganesan, B., Baker, R., Markwell, P., Davis, M., Huang, B. C., Kong, N., Prill, R. J., Marlowe, C. H., Quintanar, A., Pierre, S., Dubois, G., Kaufman, J. H., Parida, L., & Beck, K. L. (2019). Food authentication from shotgun sequencing reads with an application on high protein powders. *Npj Science of Food 2019 3:1*, *3*(1), 1–11. https://doi.org/10.1038/s41538-019-0056-6

Hajibabaei, M., Singer, G. A. C., Hebert, P. D. N., & Hickey, D. A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics : TIG*, *23*(4), 167–172. https://doi.org/10.1016/J.TIG.2007.02.001

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

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. https://doi.org/10.1098/RSPB.2002.2218

Hebert, P. D. N., Ratnasingham, S., & DeWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B: Biological Sciences*, *270*(Suppl 1), S96. https://doi.org/10.1098/RSBL.2003.0025

Heid, C. A., Stevens, J., Livak, K. J., & Williams, P. M. (1996). Real time quantitative PCR. *Genome Research*, *6*(10), 986–994. https://doi.org/10.1101/GR.6.10.986

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data 2018 5:1*, *5*(1), 1–7. https://doi.org/10.1038/sdata.2018.156

Hellmann, S. L., Kobus, R., Schmidt, B., Bikar, S., Köppel, R., & Hankeln, T. (2020). All-Food-Seq: Next Generation Sequencing-basiertes Screeningverfahren zur quantifizierbaren Speziesidentifikation in prozessierten Lebensmitteln. In *8. Fachtagung Gentechnik - Neue molekularbiologische Techniken und deren Herausforderungen für die Analytik - Band 12 Gentechnik für Umwelt- und Verbraucherschutz* (pp. 74–88). Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit.

Hellmann, S. L., Ripp, F., Bikar, S. E., Schmidt, B., Köppel, R., & Hankeln, T. (2020). Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq). *European Food Research and Technology*, *246*(1), 193–200. https://doi.org/10.1007/s00217-019-03404-y

Helyar, S. J., Lloyd, H. A. D., De Bruyn, M., Leake, J., Bennett, N., & Carvalho, G. R. (2014). Fish Product Mislabelling: Failings of Traceability in the Production Chain and Implications for Illegal, Unreported and Unregulated (IUU) Fishing. *PLOS ONE*, *9*(6), e98691. https://doi.org/10.1371/JOURNAL.PONE.0098691

Henniges, M. C., Johnston, E., Pellicer, J., Hidalgo, O., Bennett, M. D., & Leitch, I. J. (2023). The Plant DNA C-Values Database: A One-Stop Shop for Plant Genome Size Data. *Methods in Molecular Biology*, *2703*, 111–122. https://doi.org/10.1007/978-1-0716-3389-2_9

Higgins, J., Santos, B., Khanh, T. D., Trung, K. H., Duong, T. D., Doai, N. T. P., Khoa, N. T., Ha, D. T. T., Diep, N. T., Dung, K. T., Phi, C. N., Thuy, T. T., Tuan, N. T., Tran, H. D., Trung, N. T., Giang, H. T., Nhung, T. K., Tran, C. D., Lang, S. V., … De Vega, J. J. (2021). Resequencing of 672 Native Rice Accessions to Explore Genetic Diversity and Trait Associations in Vietnam. *Rice*, *14*(1), 1–16. https://doi.org/10.1186/S12284-021-00481-0

Higuchi, R., Fockler, C., Dollinger, G., & Watson, R. (1993). Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Bio/Technology (Nature Publishing Company)*, *11*(9), 1026–1030. https://doi.org/10.1038/NBT0993-1026

Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., Fazekas, A. J., Graham, S. W., James, K. E., Kim, K. J., John Kress, W., Schneider, H., van AlphenStahl, J., Barrett, S. C. H., van den Berg, C., Bogarin, D., … Little, D. P. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(31), 12794–12797. https://doi.org/10.1073/PNAS.0905845106

Hou, Y., Chen, S., Zheng, Y., Zheng, X., & Lin, J. M. (2023). Droplet-based digital PCR (ddPCR) and its applications. *TrAC Trends in Analytical Chemistry*, *158*, 116897. https://doi.org/10.1016/J.TRAC.2022.116897

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. *Human Immunology*, *82*(11), 801–811. https://doi.org/10.1016/J.HUMIMM.2021.02.012

Huang, K., Xu, W., Hu, H., Jiang, X., Sun, L., Zhao, W., Long, B., Fan, S., Zhou, Z., Mo, P., Jiang, X., Tian, J., Deng, A., Xie, P., & Wang, Y. (2024). The Mitochondrial Genome of Cathaya argyrophylla Reaches 18.99 Mb: Analysis of Super-Large Mitochondrial Genomes in Pinaceae. https://arxiv.org/abs/2410.07006v1

Huck, C. W., Pezzei, C. K., & Huck-Pezzei, V. A. (2016). An industry perspective of food fraud. *Current Opinion in Food Science*, *10*, 32–37. https://doi.org/10.1016/J.COFS.2016.07.004

Hughes, L. C., Ortí, G., Huang, Y., Sun, Y., Baldwin, C. C., Thompson, A. W., Arcila, D., Betancur, R., Li, C., Becker, L., Bellora, N., Zhao, X., Li, X., Wang, M., Fang, C., Xie, B., Zhoui, Z., Huang, H., Chen, S., … Shi, Q. (2018). Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(24), 6249–6254. https://doi.org/10.1073/PNAS.1719358115

Iammarino, M., Marino, R., & Albenzio, M. (2017). How meaty? Detection and quantification of adulterants, foreign proteins and food additives in meat products. *International Journal of Food Science & Technology*, *52*(4), 851–863. https://doi.org/10.1111/IJFS.13350

Illikoud, N., Rossero, A., Chauvet, R., Courcoux, P., Pilet, M. F., Charrier, T., Jaffrès, E., & Zagorec, M. (2019). Genotypic and phenotypic characterization of the food spoilage bacterium Brochothrix thermosphacta. *Food Microbiology*, *81*, 22–31. https://doi.org/10.1016/J.FM.2018.01.015

Imanian, B., Donaghy, J., Jackson, T., Gummalla, S., Ganesan, B., Baker, R. C., Henderson, M., Butler, E. K., Hong, Y., Ring, B., Thorp, C., Khaksar, R., Samadpour, M., Lawless, K. A., MacLaren-Lee, I., Carleton, H. A., Tian, R., Zhang, W., & Wan, J. (2022). The power, potential, benefits, and challenges of implementing high-throughput sequencing in food safety systems. *Npj Science of Food 2022 6:1*, *6*(1), 35-. https://doi.org/10.1038/s41538-022-00150-6

Inglis Id, P. W., De Castro, M., Pappas, R., Resende, L. V, & Grattapaglia, D. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. https://doi.org/10.1371/journal.pone.0206085

Interpol. (2016). *Operation OPSON V*. https://www.europol.europa.eu/sites/default/files/documents/report_opson_v.pdf

Islam, A., Halder, J., Rahman, A. M., Ud-Daula, A., Uddin, S., Hossain, M. K., Jahan, N., Alim, A., Bhuyan, A. A., Rubaya, Hasan, M., & Alam, J. (2021). Meat origin differentiation by polymerase chain reaction-restriction fragment length polymorphism. *International Journal of Food Properties*, *24*(1), 1022–1033. https://doi.org/10.1080/10942912.2021.1953068

Iwamoto, M., Ayers, T., Mahon, B. E., & Swerdlow, D. L. (2010). Epidemiology of seafood-associated infections in the United States. *Clinical Microbiology Reviews*, *23*(2), 399–411. https://doi.org/10.1128/CMR.00059-09

Jablonka, E. V. A., & Raz, G. A. L. (2009). Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Q. Rev. Biol.*, *84*(2), 131–176. https://doi.org/10.1086/598822

Jacq, C. (2021). NOFSal-MP10: A single hypervariable STR 12-plex for accurate verification of triploidy in Atlantic salmon (Salmo salar L.). *Aquaculture*, *541*, 736823. https://doi.org/10.1016/J.AQUACULTURE.2021.736823

Josefsen, M. H., Andersen, S. C., Christensen, J., & Hoorfar, J. (2015). Microbial food safety: Potential of DNA extraction methods for use in diagnostic metagenomics. *Journal of Microbiological Methods*, *114*, 30–34. https://doi.org/10.1016/J.MIMET.2015.04.016

Ju, X., Wang, Z., Cai, D., Bello, S. F., & Nie, Q. (2023). DNA methylation in poultry: a review. *Journal of Animal Science and Biotechnology*, *14*(1), 1–10. https://doi.org/10.1186/S40104-023-00939-9

Jurka, J., Zietkiewicz, E., & Labuda, D. (1995). Ubiquitous mammalian-wide interspersed repeats (MIRs) are molecular fossils from the mesozoic era. *Nucleic Acids Research*, *23*(1), 170–175. https://doi.org/10.1093/NAR/23.1.170

Kappel, K., Gadelmeier, A., Denay, G., Gerdes, L., Graff, A., Hagen, M., Hassel, M., Huber, I., Näumann, G., Pavlovic, M., Pietsch, K., Stumme, B., Völkel, I., Westerdorf, S., Wöhlke, A., Hochegger, R., Brinks, E., Franz, C., & Haase, Ilka. (2023). Detection of adulterated meat products by a next-generation sequencing-based metabarcoding analysis within the framework of the operation OPSON X: a cooperative project of the German National Reference Centre for Authentic Food (NRZ-Authent) and the competent German food control authorities. *Journal Fur Verbraucherschutz Und Lebensmittelsicherheit*, *18*(4), 375–391. https://doi.org/10.1007/S00003-023-01437-W

Kim, J. S., Chung, H., Park, B., Veerappan, K., & Kim, Y. K. (2024). Chloroplast genome sequencing and divergence analysis of 18 Pyrus species: insights into intron length polymorphisms and evolutionary processes. *Frontiers in Genetics*, *15*, 1468596. https://doi.org/10.3389/FGENE.2024.1468596

Kirpičnikov, V. S. (1987). *Genetische Grundlagen der Fischzüchtung*. Dt. Landwirtschaftsverl.

Klughammer, J., Romanovskaia, D., Nemc, A., Posautz, A., Seid, C. A., Schuster, L. C., Keinath, M. C., Lugo Ramos, J. S., Kosack, L., Evankow, A., Printz, D., Kirchberger, S., Ergüner, B., Datlinger, P., Fortelny, N., Schmidl, C., Farlik, M., Skjærven, K., Bergthaler, A., … Bock, C. (2023). Comparative analysis of genome-scale, base-resolution DNA methylation profiles across 580 animal species. *Nature Communications 2023 14:1*, *14*(1), 1–23. https://doi.org/10.1038/s41467-022-34828-y

Kobus, R., Abuín, J. M., Müller, A., Hellmann, S. L., Pichel, J. C., Pena, T. F., Hildebrandt, A., Hankeln, T., & Schmidt, B. (2020). A big data approach to metagenomics for all-food-sequencing. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-3429-6

Koch. (2024). Neuer Honigskandal: 80 Prozent des Honigs sind gefälscht. https://www.wochenblatt-dlv.de/maerkte/neuer-honigskandal-80-prozent-honigs-gefaelscht-578354

Konkel, M. K., Walker, J. A., & Batzer, M. A. (2010). LINEs and SINEs of primate evolution. *Evolutionary Anthropology: Issues, News, and Reviews*, *19*(6), 236–249. https://doi.org/10.1002/EVAN.20283

Köppel, R., Eugster, A., Ruf, J., & Rentsch, J. (2012). Quantification of Meat Proportions by Measuring DNA Contents in Raw and Boiled Sausages Using Matrix-Adapted Calibrators and Multiplex Real-Time PCR. *Journal of AOAC International*, *95*(2), 494–499. https://doi.org/10.5740/jaoacint.11-115

Köppel, R., Ganeshan, A., Weber, S., Pietsch, K., Graf, C., Hochegger, R., Griffiths, K., & Burkhardt, S. (2019). Duplex digital PCR for the determination of meat proportions of sausages containing meat from chicken, turkey, horse, cow, pig and sheep. *European Food Research and Technology*, *245*(4), 853–862. https://doi.org/10.1007/s00217-018-3220-3

Köppel, R., Rentsch, J., Ruf, J., Eugster, A., Graf, C., Felderer, N., Pietsch, K., & Ilg, E. (2014). Results of an International Interlaboratory Trial to Determine Twelve Allergens Using Real-time PCR- and ELISA-based Assays. *Chimia*, *68*(10), 721–725. https://doi.org/10.2533/CHIMIA.2014.721

Köppel, R., Ruf, J., & Rentsch, J. (2011). Multiplex real-time PCR for the detection and quantification of DNA from beef, pork, horse and sheep. *European Food Research and Technology*, *232*(1), 151–155. https://doi.org/10.1007/s00217-010-1371-y

Kovaka, S., Ou, S., Jenike, K. M., & Schatz, M. C. (2023). Approaching complete genomes, transcriptomes and epi-omes with accurate long-read sequencing. *Nature Methods 2023 20:1*, *20*(1), 12–16. https://doi.org/10.1038/s41592-022-01716-8

Kramerov, D. A., & Vassetzky, N. S. (2005). Short retroposons in eukaryotic genomes. *International Review of Cytology*, *247*, 165–221. https://doi.org/10.1016/S0074-7696(05)47004-7

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias in qualitative and quantitative arthropod metabarcoding. *Scientific Reports 2017 7:1*, *7*(1), 17668-. https://doi.org/10.1038/s41598-017-17333-x

Krull, M., Petrusma, M., Makalowski, W., Brosius, J., & Schmitz, J. (2007). Functional persistence of exonized mammalian-wide interspersed repeat elements (MIRs). *Genome Research*, *17*(8), 1139–1145. https://doi.org/10.1101/GR.6320607

Kubista, M., Andrade, J. M., Bengtsson, M., Forootan, A., Jonák, J., Lind, K., Sindelka, R., Sjöback, R., Sjögreen, B., Strömbom, L., Ståhlberg, A., & Zoric, N. (2006). The real-time polymerase chain reaction. *Molecular Aspects of Medicine*, *27*(2–3), 95–125. https://doi.org/10.1016/J.MAM.2005.12.007

Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

Lanubile, A., Stagnati, L., Marocco, A., & Busconi, M. (2024). DNA-based techniques to check quality and authenticity of food, feed and medicinal products of plant origin: A review. *Trends in Food Science & Technology*, *149*, 104568. https://doi.org/10.1016/J.TIFS.2024.104568

Lebrón, R., Gómez-Martín, C., Carpena, P., Bernaola-Galván, P., Barturen, G., Hackenberg, M., & Oliver, J. L. (2017). NGSmethDB 2017: enhanced methylomes and differential methylation. *Nucleic Acids Research*, *45*(D1), D97–D103. https://doi.org/10.1093/NAR/GKW996

Leitch IJ, Johnston E, Pellicer J, Hidalgo O, & Bennett MD. (2019). *Plant DNA C-values Database (Release 7.1)*.

Levy, A. A., & Feldman, M. (2022). Evolution and origin of bread wheat. *The Plant Cell*, *34*(7), 2549. https://doi.org/10.1093/PLCELL/KOAC130

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., Durbin, R., Edwards, S. V., Forest, F., Gilbert, M. T. P., Goldstein, M. M., Grigoriev, I. V., Hackett, K. J., Haussler, D., Jarvis, E. D., Johnson, W. E., Patrinos, A., Richards, S., Castilla-Rubio, J. C., … Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(17), 4325–4333. https://doi.org/10.1073/PNAS.1720115115

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://arxiv.org/abs/1303.3997v2

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, R., Liang, F., Li, M., Zou, D., Sun, S., Zhao, Y., Zhao, W., Bao, Y., Xiao, J., & Zhang, Z. (2017). MethBank 3.0: a database of DNA methylomes across a variety of species. *Nucleic Acids Research*, *46* (Database issue), D288. https://doi.org/10.1093/NAR/GKX1139

Li, W., & Freudenberg, J. (2014). Mappability and read length. *Frontiers in Genetics*, *5*(NOV), 1–1. https://doi.org/10.3389/FGENE.2014.00381

Liou, P., Banda, A., Isaacs, R. B., & Hellberg, R. S. (2020). Labeling compliance and species authentication of fish fillets sold at grocery stores in Southern California. *Food Control*, *112*, 107137. https://doi.org/10.1016/J.FOODCONT.2020.107137

Liu, Y., Gao, X., Zang, M., Sun, B., Zhang, S., Xie, P., & Liu, X. (2025). Insights into Microbial Community and Its Enzymatic Profiles in Commercial Dry-Aged Beef. *Foods*, *14*(3), 529. https://doi.org/10.3390/FOODS14030529

Liu, Y., Ripp, F., Koeppel, R., Schmidt, H., Lukas Hellmann, S., Weber, M., Krombholz, C. F., Schmidt, B., & Hankeln, T. (2017). AFS: identification and quantification of species composition by metagenomic sequencing. *Bioinformatics*, *January*, btw822. https://doi.org/10.1093/bioinformatics/btw822

Lo, Y. T., & Shaw, P. C. (2018). DNA-based techniques for authentication of processed food and food supplements. *Food Chemistry*, *240*, 767–774. https://doi.org/10.1016/J.FOODCHEM.2017.08.022

Lorusso, L., Shum, P., Piredda, R., Mottola, A., Maiello, G., Cartledge, E. L., Neave, E. F., Di Pinto, A., & Mariani, S. (2024). Mismanagement and poor transparency in the European processed seafood supply revealed by DNA metabarcoding. *Food Research International*, *194*, 114901. https://doi.org/10.1016/J.FOODRES.2024.114901

Lu, A. T., Fei, Z., Haghani, A., Robeck, T. R., Zoller, J. A., Li, C. Z., Lowe, R., Yan, Q., Zhang, J., Vu, H., Ablaeva, J., Acosta-Rodriguez, V. A., Adams, D. M., Almunia, J., Aloysius, A., Ardehali, R., Arneson, A., Baker, C. S., Banks, G., … Horvath, S. (2023). Universal DNA methylation age across mammalian tissues. *Nature Aging 2023 3:9*, *3*(9), 1144–1166. https://doi.org/10.1038/s43587-023-00462-6

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ. Computer Science*, *3*(1). https://doi.org/10.7717/PEERJ-CS.104

Ma, X. Y., Shao, Z. L., Yu, X. P., & Wang, Z. L. (2023). A Droplet Digital PCR-Based Approach for Quantitative Analysis of the Adulteration of Atlantic Salmon with Rainbow Trout. *Foods 2023, Vol. 12, Page 4309*, *12*(23), 4309. https://doi.org/10.3390/FOODS12234309

Ma, Y., Li, Y., Jiang, C., Zheng, L., Liu, S., & Zhao, L. (2022). High-quality chromosome-level genome assembly of Pacific cod, Gadus macrocephalus. *Frontiers in Marine Science*, *9*. https://doi.org/10.3389/fmars.2022.1067526

Macher, T. H., Schütz, R., Yildiz, A., Beermann, A. J., & Leese, F. (2023). Evaluating five primer pairs for environmental DNA metabarcoding of Central European fish species based on mock communities. *Metabarcoding and Metagenomics 7: E103856*, *7*, e103856-. https://doi.org/10.3897/MBMG.7.103856

Macqueen, D. J., & Johnston, I. A. (2014). A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1778). https://doi.org/10.1098/RSPB.2013.2881

Mallott, E. K., Garber, P. A., & Malhi, R. S. (2018). trnL outperforms rbcL as a DNA metabarcoding marker when compared with the observed plant component of the diet of wild white-faced capuchins (Cebus capucinus, Primates). *PLoS ONE*, *13*(6), e0199556. https://doi.org/10.1371/JOURNAL.PONE.0199556

Mano, J., Nishitsuji, Y., Kikuchi, Y., Fukudome, S. ichi, Hayashida, T., Kawakami, H., Kurimoto, Y., Noguchi, A., Kondo, K., Teshima, R., Takabatake, R., & Kitta, K. (2017). Quantification of DNA fragmentation in processed foods using real-time PCR. *Food Chemistry*, *226*, 149–155. https://doi.org/10.1016/J.FOODCHEM.2017.01.064

Maquet, A., Lievens, A., Paracchini, V., Kaklamanos, G., De la Calle Guntinas, M. B., Garlant, L., Papoci, S., Pietretti, D., Ždiniakova, T., Breidbach, A., Omar, O. J., Boix, S. A., Dimitrova, T., & Ulberth, F. (2021). Results of an EU wide coordinated control plan to establish the prevalence of fraudulent practices in the marketing of herbs and spices. https://doi.org/10.2760/309557

Masuyama, M., Iida, R., Takatsuka, H., Yasuda, T., & Matsuki, T. (2005). Quantitative change in mitochondrial DNA content in various mouse tissues during aging. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1723*(1–3), 302–308. https://doi.org/10.1016/J.BBAGEN.2005.03.001

Meissner, A., Gnirke, A., Bell, G. W., Ramsahoye, B., Lander, E. S., & Jaenisch, R. (2005). Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic Acids Research*, *33*(18), 5868. https://doi.org/10.1093/NAR/GKI901

Merchant, S., Wood, D. E., & Salzberg, S. L. (2014). Unexpected cross-species contamination in genome sequencing projects. *PeerJ*, *2014*(1), e675. https://doi.org/10.7717/PEERJ.675

Miller, F. J., Rosenfeldt, F. L., Zhang, C., Linnane, A. W., & Nagley, P. (2003). Precise determination of mitochondrial DNA copy number in human skeletal and cardiac muscle by a PCR-based assay: lack of change of copy number with age. *Nucleic Acids Research*, *31*(11). https://doi.org/10.1093/NAR/GNG060

Mottola, A., Intermite, C., Piredda, R., Lorusso, L., Ranieri, L., Carpino, S., Celano, G. V., & Di Pinto, A. (2024). DNA Metabarcoding Approach as a Potential Tool for Supporting Official Food Control Programs: A Case Study. *Foods*, *13*(18), 2941. https://doi.org/10.3390/FOODS13182941

Mottola, A., Piredda, R., Lorusso, L., Ranieri, L., Intermite, C., Barresi, C., Galli, C., & Di Pinto, A. (2024). Decoding Seafood: Multi-Marker Metabarcoding for Authenticating Processed Seafood. *Foods*, *13*(15), 2382. https://doi.org/10.3390/FOODS13152382

Muflihah, Hardianto, A., Kusumaningtyas, P., Prabowo, S., & Hartati, Y. W. (2023). DNA-based detection of pork content in food. *Heliyon*, *9*(3), e14418. https://doi.org/10.1016/J.HELIYON.2023.E14418

Müller, A., Wichmann, A., Kallenborn, F., Hellmann, S. L., Hankeln, T., & Schmidt, B. (2025). Improved Metagenomic Analysis for All-Food-Sequencing with AFS-MetaCache2: Illumina vs. Nanopore. *BioRxiv*, 2025.12.18.694891. https://doi.org/10.64898/2025.12.18.694891

Murray, D. S., Kainz, M. J., Hebberecht, L., Sales, K. R., Hindar, K., & Gage, M. J. G. (2018). Comparisons of reproductive function and fatty acid fillet quality between triploid and diploid farm Atlantic salmon (Salmo salar). *Royal Society Open Science*, *5*(8). https://doi.org/10.1098/RSOS.180493

Nayak, S., Nayak, P., Saha, S., Patnaik, L., Nayak, S., Pradhan, S. P., Sharma, S. N., & Muduli, N. (2024). Formaldehyde as Preservative for Fish and Seafood: A Boon or a Bane. *Environmental Quality Management*, *34*(2), e22357. https://doi.org/10.1002/TQEM.22357

Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Wainwright, P. C., Friedman, M., & Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(34), 13698–13703. https://doi.org/10.1073/PNAS.1206625109

Nielsen, D. S., Jacobsen, T., Jespersen, L., Koch, A. G., & Arneborg, N. (2008). Occurrence and growth of yeasts in processed meat products – Implications for potential spoilage. *Meat Science*, *80*(3), 919–926. https://doi.org/10.1016/J.MEATSCI.2008.04.011

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/SCIENCE.ABJ6987

O'Mahony, P. J. (2013). Finding horse meat in beef products--a global problem. *QJM: Monthly Journal of the Association of Physicians*, *106*(6), 595–597. https://doi.org/10.1093/QJMED/HCT087

Omran, G. A., Tolba, A. O., El-Sharkawy, E. E. E. D., Abdel-Aziz, D. M., & Ahmed, H. Y. (2019). Species DNA-based identification for detection of processed meat adulteration: is there a role of human short tandem repeats (STRs)? *Egyptian Journal of Forensic Sciences*, *9*(1), 15-. https://doi.org/10.1186/S41935-019-0121-Y

Ou, J., Zheng, J., Huang, J., Ho, C. T., & Ou, S. (2020). Interaction of Acrylamide, Acrolein, and 5-Hydroxymethylfurfural with Amino Acids and DNA. *Journal of Agricultural and Food Chemistry*, *68*(18), 5039–5048. https://doi.org/10.1021/ACS.JAFC.0C01345

Ounit, R., Wanamaker, S., Close, T. J., & Lonardi, S. (2015). CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics*, *16*(1), 236-. https://doi.org/10.1186/S12864-015-1419-2

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403–1414. https://doi.org/10.1111/1462-2920.13023

Pardo, M. Á., Jiménez, E., Viðarsson, J. R., Ólafsson, K., Ólafsdóttir, G., Daníelsdóttir, A. K., & Pérez-Villareal, B. (2018). DNA barcoding revealing mislabeling of seafood in European mass caterings. *Food Control*, *92*, 7–16. https://doi.org/10.1016/J.FOODCONT.2018.04.044

Park, S., Lee, J., Kim, J., Kim, D., Lee, J. H., Pack, S. P., & Seo, M. (2023). Benchmark study for evaluating the quality of reference genomes and gene annotations in 114 species. *Frontiers in Veterinary Science*, *10*, 1128570. https://doi.org/10.3389/FVETS.2023.1128570

Pearman, W. S., Freed, N. E., & Silander, O. K. (2020). Testing the advantages and disadvantages of short- And long- read eukaryotic metagenomics using simulated reads. *BMC Bioinformatics*, *21*(1), 220-. https://doi.org/10.1186/S12859-020-3528-4

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q., Irestedt, M., & Suh, A. (2021). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263–286. https://doi.org/10.1111/1755-0998.13252

Picard, M. (2021). Blood mitochondrial DNA copy number: What are we counting? *Mitochondrion*, *60*, 1–11. https://doi.org/10.1016/J.MITO.2021.06.010

Piferrer, F., Beaumont, A., Falguière, J. C., Flajšhans, M., Haffray, P., & Colombo, L. (2009). Polyploid fish and shellfish: Production, biology and applications to aquaculture for performance improvement and genetic containment. *Aquaculture*, *293*(3–4), 125–156. https://doi.org/10.1016/J.AQUACULTURE.2009.04.036

Piñol, J., Senar, M. A., & Symondson, W. O. C. (2019). The choice of universal primers and the characteristics of the species mixture determine when DNA metabarcoding can be quantitative. *Molecular Ecology*, *28*(2), 407–419. https://doi.org/10.1111/MEC.14776

Pirondini, A., Bonas, U., Maestri, E., Visioli, G., Marmiroli, M., & Marmiroli, N. (2010). Yield and amplificability of different DNA extraction procedures for traceability in the dairy food chain. *Food Control*, *21*(5), 663–668. https://doi.org/10.1016/J.FOODCONT.2009.10.004

Piskata, Z., Servusova, E., Babak, V., Nesvadbova, M., & Borilova, G. (2019). The Quality of DNA Isolated from Processed Food and Feed via Different Extraction Procedures. *Molecules*, *24*(6), 1188. https://doi.org/10.3390/MOLECULES24061188

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*(15), 2511–2542. https://doi.org/10.1080/10408398.2013.873767

Preckel, L., Brünen-Nieweler, C., Denay, G., Petersen, H., Cichna-Markl, M., Dobrovolny, S., & Hochegger, R. (2021). Identification of Mammalian and Poultry Species in Food and Pet Food Samples Using 16S rDNA Metabarcoding. *Foods 2021, Vol. 10, Page 2875*, *10*(11), 2875. https://doi.org/10.3390/FOODS10112875

Quan, P. L., Sauzade, M., & Brouzes, E. (2018). dPCR: A Technology Review. *Sensors 2018, Vol. 18, Page 1271*, *18*(4), 1271. https://doi.org/10.3390/S18041271

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–395. https://doi.org/10.1038/S41586-018-0273-1

Raclariu, A. C., Heinrich, M., Ichim, M. C., & de Boer, H. (2017). Benefits and Limitations of DNA Barcoding and Metabarcoding in Herbal Product Authentication. *Phytochemical Analysis*, *29*(2), 123. https://doi.org/10.1002/PCA.2732

Rath, S. P., Gupta, R., Todres, E., Wang, H., Jourdain, A. A., Ardlie, K. G., Calvo, S. E., & Mootha, V. K. (2024). Mitochondrial genome copy number variation across tissues in mice and humans. *Proceedings of the National Academy of Sciences of the United States of America*, *121*(33), e2402291121. https://doi.org/10.1073/PNAS.2402291121

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System: Barcoding. *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., Wei, C., Chan, D., Agda, J., Agda, J., Ballesteros-Mejia, L., Boutou, H. A., El Bastami, Z. M., Ma, E., Manjunath, R., Rea, D., Ho, C., Telfer, A., McKeowan, J., Rahulan, M., Steinke, C., Dorsheimer, J., Milton, M., & Hebert, P. D. N. (2024). BOLD v4: A Centralized Bioinformatics Platform for DNA-Based Biodiversity Data. *Methods in Molecular Biology (Clifton, N.J.)*, *2744*, 403–441. https://doi.org/10.1007/978-1-0716-3581-0_26

Rehbein, H., & Oehlenschläger, J. (2009). Fishery Products: Quality, Safety and Authenticity. *Fishery Products: Quality, Safety and Authenticity*, 1–477. https://doi.org/10.1002/9781444322668

Reik, W., Dean, W., & Walter, J. (2001). Epigenetic reprogramming in mammalian development. *Science*, *293*(5532), 1089–1093. https://doi.org/10.1126/science.1063443

Rezadoost, M. H., Kordrostami, M., & Kumleh, H. H. (2016). An efficient protocol for isolation of inhibitor-free nucleic acids even from recalcitrant plants. *3 Biotech*, *6*(1), 61. https://doi.org/10.1007/S13205-016-0375-0

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature 2021 592:7856*, *592*(7856), 737–746. https://doi.org/10.1038/s41586-021-03451-0

Rieder, J., Kapopoulou, A., Bank, C., & Adrian-Kalchhauser, I. (2023). Metagenomics and metabarcoding experimental choices and their impact on microbial community characterization in freshwater recirculating aquaculture systems. *Environmental Microbiome*, *18*(1), 1–21. https://doi.org/10.1186/S40793-023-00459-Z

Ripp, F., Krombholz, C., Liu, Y., Weber, M., Schäfer, A., Schmidt, B., Köppel, R., & Hankeln, T. (2014). All-Food-Seq (AFS): a quantifiable screen for species in biological samples by deep DNA sequencing. *BMC Genomics*, *15*(1), 639. https://doi.org/10.1186/1471-2164-15-639

Rivera, C. M., & Ren, B. (2013). Mapping human epigenomes. *Cell*, *155*(1), 39–55. https://doi.org/10.1016/j.cell.2013.09.011

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature 2015 518:7539*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Robert Koch Institute. (2019). Listeriose-Ausbruch mit Listeria monocytogenes Sequenz-Cluster-Typ 2521 (Sigma1) in Deutschland. *Epidemiologisches Bulletin*, *41*. https://www.rki.de/DE/Content/Infekt/EpidBull/Archiv/2019/41/Art_02.html

Rodríguez López, C. M., Morán, P., Lago, F., Espiñeira, M., Beckmann, M., & Consuegra, S. (2012). Detection and quantification of tissue of origin in salmon and veal products using methylation sensitive AFLPs. *Food Chemistry*, *131*(4), 1493–1498. https://doi.org/10.1016/J.FOODCHEM.2011.09.120

Romagnoli, S., Bartalucci, N., & Vannucchi, A. M. (2023). Resolving complex structural variants via nanopore sequencing. *Frontiers in Genetics*, *14*, 1213917. https://doi.org/10.3389/FGENE.2023.1213917

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547. https://doi.org/10.1016/J.GECCO.2019.E00547

Ryan, N. M., & Corvin, A. (2023). Investigating the dark-side of the genome: a barrier to human disease variant discovery? *Biological Research*, *56*(1), 42. https://doi.org/10.1186/S40659-023-00455-0

Sayers, E. W., Bolton, E. E., Brister, J. R., Canese, K., Chan, J., Comeau, D. C., Connor, R., Funk, K., Kelly, C., Kim, S., Madej, T., Marchler-Bauer, A., Lanczycki, C., Lathrop, S., Lu, Z., Thibaud-Nissen, F., Murphy, T., Phan, L., Skripchenko, Y., … Sherry, S. T. (2022). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *50*(D1), D20–D26. https://doi.org/10.1093/NAR/GKAB1112

Scarano, C., Veneruso, I., De Simone, R. R., Di Bonito, G., Secondino, A., & D'Argenio, V. (2024). The Third-Generation Sequencing Challenge: Novel Insights for the Omic Sciences. *Biomolecules 2024, Vol. 14, Page 568*, *14*(5), 568. https://doi.org/10.3390/BIOM14050568

Schenk, J. J., Becklund, L. E., Carey, S. J., & Fabre, P. P. (2023). What is the "modified" CTAB protocol? Characterizing modifications to the CTAB DNA extraction protocol. *Applications in Plant Sciences*, *11*(3), e11517. https://doi.org/10.1002/APS3.11517

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N., & Quince, C. (2016). Illumina error profiles: Resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics*, *17*(1), 125-. https://doi.org/10.1186/S12859-016-0976-Y

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Bolchacova, E., Voigt, K., Crous, P. W., Miller, A. N., Wingfield, M. J., Aime, M. C., An, K. D., Bai, F. Y., Barreto, R. W., Begerow, D., Bergeron, M. J., Blackwell, M., … Schindel, D. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), 6241–6246. https://doi.org/10.1073/PNAS.1117018109

Schultz, M. D., He, Y., Whitaker, J. W., Hariharan, M., Mukamel, E. A., Leung, D., Rajagopal, N., Nery, J. R., Urich, M. A., Chen, H., Lin, S., Lin, Y., Jung, I., Schmitt, A. D., Selvaraj, S., Ren, B., Sejnowski, T. J., Wang, W., & Ecker, J. R. (2015). Human body epigenome maps reveal noncanonical DNA methylation variation. *Nature 2015 523:7559*, *523*(7559), 212–216. https://doi.org/10.1038/nature14465

Senyuva, H. Z., Jones, I. B., Sykes, M., & Baumgartner, S. (2019). A critical review of the specifications and performance of antibody and DNA-based methods for detection and quantification of allergens in foods. *Food Additives & Contaminants: Part A*, *36*(4), 507–547. https://doi.org/10.1080/19440049.2019.1579927

Seymour, D. K., Koenig, D., Hagmann, J., Becker, C., & Weigel, D. (2014). Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. *PLOS Genetics*, *10*(11), e1004785. https://doi.org/10.1371/JOURNAL.PGEN.1004785

Shaffer, M. R., Andruszkiewicz Allan, E., Van Cise, A. M., Parsons, K. M., Shelton, A. O., & Kelly, R. P. (2025). Observation Bias in Metabarcoding. *Molecular Ecology Resources*, *25*(7), e14119. https://doi.org/10.1111/1755-0998.14119

Shelton, A. O., Gold, Z. J., Jensen, A. J., D'Agnese, E., Andruszkiewicz Allan, E., Van Cise, A., Gallego, R., Ramón-Laca, A., Garber-Yonts, M., Parsons, K., & Kelly, R. P. (2022). Toward quantitative metabarcoding. *Ecology*, *104*(2). https://doi.org/10.1002/ECY.3906

Sherman, R. M., Forman, J., Antonescu, V., Puiu, D., Daya, M., Rafaels, N., Boorgula, M. P., Chavan, S., Vergara, C., Ortega, V. E., Levin, A. M., Eng, C., Yazdanbakhsh, M., Wilson, J. G., Marrugo, J., Lange, L. A., Williams, L. K., Watson, H., Ware, L. B., … Salzberg, S. L. (2018). Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nature Genetics 2018 51:1*, *51*(1), 30–35. https://doi.org/10.1038/s41588-018-0273-y

Shokralla, S., Hellberg, R. S., Handy, S. M., King, I., & Hajibabaei, M. (2015). A DNA Mini-Barcoding System for Authentication of Processed Fish Products. *Scientific Reports 2015 5:1*, *5*(1), 1–11. https://doi.org/10.1038/srep15894

Sidstedt, M., Rådström, P., & Hedman, J. (2020). PCR inhibition in qPCR, dPCR and MPS—mechanisms and solutions. *Analytical and Bioanalytical Chemistry 2020 412:9*, *412*(9), 2009–2023. https://doi.org/10.1007/S00216-020-02490-2

Singh, M., Young, R. G., Hellberg, R. S., Hanner, R. H., Corradini, M. G., & Farber, J. M. (2024). Twenty-three years of PCR-based seafood authentication assay development: What have we learned? *Comprehensive Reviews in Food Science and Food Safety*, *23*(4), e13401. https://doi.org/10.1111/1541-4337.13401

Siuta, J., Dobosz, A., Kawecki, J., & Dobosz, T. (2023). DNA Content of Various Fluids and Tissues of the Human Body. *Genes 2024, Vol. 15, Page 17*, *15*(1), 17. https://doi.org/10.3390/GENES15010017

Smit, A. F. A., & Riggs, A. D. (1995). MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Research*, *23*(1), 98–102. https://doi.org/10.1093/NAR/23.1.98

Smith, D. R. (2015). Mutation Rates in Plastid Genomes: They Are Lower than You Might Think. *Genome Biology and Evolution*, *7*(5), 1227. https://doi.org/10.1093/GBE/EVV069

Smith, Z. D., & Meissner, A. (2013). DNA methylation: Roles in mammalian development. *Nat. Rev. Genet.*, *14*(3), 204–220. https://doi.org/10.1038/nrg3354

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. https://doi.org/10.1007/S00216-016-9595-8

Streeter, K., & Katouli, M. (2016). Pseudomonas aeruginosa: A review of their Pathogenesis and Prevalence in Clinical Settings and the Environment. *Infection, Epidemiology and Medicine*, *2*(1), 25–32. https://doi.org/10.18869/MODARES.IEM.2.1.25

Szyłak, A., Kostrzewa, W., Bania, J., & Tabiś, A. (2023). Do You Know What You Eat? Kebab Adulteration in Poland. *Foods*, *12*(18), 3380. https://doi.org/10.3390/FOODS12183380

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

Teletchea, F., Maudet, C., & Hänni, C. (2005). Food and forensic molecular identification: update and challenges. *Trends in Biotechnology*, *23*(7), 359–366. https://doi.org/10.1016/J.TIBTECH.2005.05.006

Thünen Institute of Fisheries Ecology. Nordostpazifik - Fischbestände (Northeast Pacific - Fish stocks*)*. Retrieved May 29, 2024, from https://www.fischbestaende-online.de/fao-fanggebiete/nordostpazifik

Tigrero-Vaca, J., Díaz, B., Gu, G., & Cevallos-Cevallos, J. M. (2025). Next-generation sequencing applications in food science: fundamentals and recent advances. *Frontiers in Bioengineering and Biotechnology*, *13*, 1638957. https://doi.org/10.3389/FBIOE.2025.1638957

Treangen, T. J., & Salzberg, S. L. (2011). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, *13*(1), 36. https://doi.org/10.1038/NRG3117

Turck, D., Castenmiller, J., de Henauw, S., Hirsch-Ernst, K., Kearney, J., Maciuk, A., Mangelsdorf, I., McArdle, H. J., Naska, A., Pelaez, C., Pentieva, K., Siani, A., Thies, F., Tsabouri, S., Vinceti, M., Cubadda, F., Engel, K., Frenzel, T., Heinonen, M., … Knutsen, H. K. (2019). Safety of Yarrowia lipolytica yeast biomass as a novel food pursuant to Regulation (EU) 2015/2283. *EFSA Journal*, *17*(2), e05594. https://doi.org/10.2903/J.EFSA.2019.5594

Ulca, P., Balta, H., Çağin, I., & Senyuva, H. Z. (2013). Meat species identification and Halal authentication using PCR analysis of raw and cooked traditional Turkish foods. *Meat Science*, *94*(3), 280–284. https://doi.org/10.1016/J.MEATSCI.2013.03.008

Valentini, A., Pompanon, F., & Taberlet, P. (2009). DNA barcoding for ecologists. *Trends in Ecology & Evolution*, *24*(2), 110–117. https://doi.org/10.1016/J.TREE.2008.09.011

Venkatesh, G., Tönges, S., Hanna, K., Ng, Y. L., Whelan, R., Andriantsoa, R., Lingenberg, A., Roy, S., Nagarajan, S., Fong, S., Raddatz, G., Böhl, F., & Lyko, F. (2023). Context-dependent DNA methylation signatures in animal livestock. *Environmental Epigenetics*, *9*(1). https://doi.org/10.1093/EEP/DVAD001

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. de F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology*, *39*(3), 312. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Vignal, A., Milan, D., SanCristobal, M., & Eggen, A. (2002). A review on SNP and other types of molecular markers and their use in animal genetics. *Genetics, Selection, Evolution : GSE*, *34*(3). https://doi.org/10.1186/1297-9686-34-3-275

Viruel, J., Conejero, M., Hidalgo, O., Pokorny, L., Powell, R. F., Forest, F., Kantar, M. B., Soto Gomez, M., Graham, S. W., Gravendeel, B., Wilkin, P., & Leitch, I. J. (2019). A Target Capture-Based Method to Estimate Ploidy From Herbarium Specimens. *Frontiers in Plant Science*, *10*, 467104. https://doi.org/10.3389/FPLS.2019.00937

Volff, J. N. (2004). Genome evolution and biodiversity in teleost fish. *Heredity 2005 94:3*, *94*(3), 280–294. https://doi.org/10.1038/sj.hdy.6800635

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/BIOINFORMATICS/BTX153

Walters, W., Hyde, E. R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J. A., Jansson, J. K., Caporaso, J. G., Fuhrman, J. A., Apprill, A., & Knight, R. (2016). Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *MSystems*, *1*(1). https://doi.org/10.1128/MSYSTEMS.00009-15

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., & Burbano, H. A. (2018). nQuire: A statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, *19*(1), 122-. https://doi.org/10.1186/S12859-018-2128-Z

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3). https://doi.org/10.1186/GB-2014-15-3-R46

Worm, M., Jappe, U., Kleine-Tebbe, J., Schäfer, C., Reese, I., Saloga, J., Treudler, R., Zuberbier, T., Waßmann, A., Fuchs, T., Dölle, S., Raithel, M., Ballmer-Weber, B., Niggemann, B., & Werfel, T. (2014). Food allergies resulting from immunological cross-reactivity with inhalant allergens: Guidelines from the German Society for Allergology and Clinical Immunology (DGAKI), the German Dermatology Society (DDG), the Association of German Allergologists (AeDA) and the Society for Pediatric Allergology and Environmental Medicine (GPA). *Allergo Journal International*, *23*(1), 1. https://doi.org/10.1007/S40629-014-0004-6

Xu, X., Pan, S., Cheng, S., Zhang, B., Mu, D., Ni, P., Zhang, G., Yang, S., Li, R., Wang, J., Orjeda, G., Guzman, F., Torres, M., Lozano, R., Ponce, O., Martinez, D., De La Cruz, G., Chakrabarti, S. K., Patil, V. U., … Visser, R. G. F. (2011). Genome sequence and analysis of the tuber crop potato. *Nature 2011 475:7355*, *475*(7355), 189–195. https://doi.org/10.1038/nature10158

Yao, L., Lu, J., Qu, M., Jiang, Y., Li, F., Guo, Y., Wang, L., & Zhai, Y. (2020). Methodology and application of PCR-RFLP for species identification in tuna sashimi. *Food Science & Nutrition*, *8*(7), 3138–3146. https://doi.org/10.1002/FSN3.1552

Yoder, J. A., Walsh, C. P., & Bestor, T. H. (1997). Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*, *13*(8), 335–340. https://doi.org/10.1016/s0168-9525(97)01181-5

Zhang, C., Fang, X., Qiu, H., & Li, N. (2015). A Short Interspersed Nuclear Element (SINE)-Based Real-Time PCR Approach to Detect and Quantify Porcine Component in Meat Products. *Journal of AOAC International*, *98*(5), 1471–1473. https://doi.org/10.5740/JAOACINT.15-056

Zhang, T., Zhou, J., Gao, W., Jia, Y., Wei, Y., & Wang, G. (2022). Complex genome assembly based on long-read sequencing. *Briefings in Bioinformatics*, *23*(5). https://doi.org/10.1093/BIB/BBAC305

Zhang, X., Wang, T., Ji, J., Wang, H., Zhu, X., Du, P., Zhu, Y., Huang, Y., & Chen, W. (2020). The distinct spatiotemporal distribution and effect of feed restriction on mtDNA copy number in broilers. *Scientific Reports*, *10*(1), 3240. https://doi.org/10.1038/S41598-020-60123-1

Zhao, C., Wang, D., Teng, J., Yang, C., Zhang, X., Wei, X., & Zhang, Q. (2023). Breed identification using breed-informative SNPs and machine learning based on whole genome sequence data and SNP chip data. *Journal of Animal Science and Biotechnology*, *14*(1), 1–13. https://doi.org/10.1186/S40104-023-00880-X

Zhao, J., Zhu, C., Xu, Z., Jiang, X., Yang, S., & Chen, A. (2017). Microsatellite markers for animal identification and meat traceability of six beef cattle breeds in the Chinese market. *Food Control*, *78*, 469–475. https://doi.org/10.1016/J.FOODCONT.2017.03.017

Zhou, Y., Liu, S., Hu, Y., Fang, L., Gao, Y., Xia, H., Schroeder, S. G., Rosen, B. D., Connor, E. E., Li, C. J., Baldwin, R. L., Cole, J. B., Van Tassell, C. P., Yang, L., Ma, L., & Liu, G. E. (2020). Comparative whole genome DNA methylation profiling across cattle tissues reveals global and tissue-specific methylation patterns. *BMC Biology*, *18*(1), 1–17. https://doi.org/10.1186/S12915-020-00793-5

Zinjarde, S. S. (2014). Food-related applications of Yarrowia lipolytica. *Food Chemistry*, *152*, 1–10. https://doi.org/10.1016/J.FOODCHEM.2013.11.117

Some figures were created with https://BioRender.com

Abbreviations

12S (rRNA)	12-Svedberg (ribosomal RNA)
16S (rRNA)	16-Svedberg (ribosomal RNA)
18S (rRNA)	18-Svedberg (ribosomal RNA)
5hmC	5-hydroxymethylcytosine
5mC	5-methylcytosine
AFS	All-Food-Sequencing
ASV	Amplicon Sequence Variant
BLAST	Basic Local Alignment Search Tool
BOLD	Barcode of Life Data System
BWA	Burrows-Wheeler Aligner
CLARK	CLAssifier based on Reduced K-mers
COI	Cytochrome C Oxidase I
CPU	Central Processing Unit
ct(DNA)	chloroplast (DNA)
CTAB	Cetyltrimethylammoniumbromid
CYTB	Cytochrome b
ddPCR	Droplet digital polymerase chain reaction
DMR	differentially methylated regions
FASER	Food Authentication from SEquencing Reads
Gb	Gigabases
ITS	internal transcribed spacer
kb	kilobases
LINEs	Long interspersed nuclear elements
LoD	Limit of Detection
matK	Maturase K
Mb	Megabases
MIRs	Mammalian-wide interspersed repeats
mt(DNA)	mitochondrial (DNA)
MYA	million years ago
NCBI	National Center for Biotechnology Information
NGS	Next-Generation Sequencing
nDNA	nuclear DNA
NGS	Next-Generation Sequencing
nr/nt	non-redundant nucleotide collection

ONT	Oxford Nanopore Technologies
OTU	Operational Taxonomic Units
PCR	Polymerase chain reaction
PacBio	Pacific Biosciences
qPCR	Quantitative polymerase chain reaction
RAM	Random-Access Memory
rbcL	Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase large subunit
RuBisCO	Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase
rRNA	ribosomal RNA
SINEs	Short interspersed nuclear elements
SNP	single nucleotide polymorphism
SSR	simple sequence repeats
STR	short tandem repeats
TGS	Third-Generation Sequencing
TE	Transposable Element
trnL	Leucine-tRNA
WGS	whole-genome shotgun

Acknowledgements

Declaration of use of Artificial Intelligence

Task	**Tools**
Code prototyping and scripting assistance	Gemini, GPT
Data cleaning and table harmonization	Gemini, GPT
Debugging and error-troubleshooting	Gemini, GPT, perplexity
Consistency checks across chapters and figures	Gemini, GPT, perplexity
Grammar, style, and consistency editing	Gemini, GPT, LanguageTool
Translation and english phrasing support	DeepL, GPT, LanguageTool
Outline development, structural planning, brainstorming	Gemini, GPT, NotebookLM
Literature discovery	Elicit, Gemini, GPT, perplexity
Summarization of background material	Gemini, GPT, NotebookLM
Scientific writing refinement	Gemini, GPT, LanguageTool

Eidesstattliche Erklärung

VERSICHERUNG
für das Gesuch um Zulassung zur Promotion in den Fachbereichen 17 – 22 der Johannes Gutenberg-Universität Mainz

Name: Hellmann, Sören Lukas

Hiermit versichere ich gemäß § 11, Abs. 3d der Promotionsordnung vom 22.12.2003:
(zutreffendes ist angekreuzt.)

☒ Ich habe die heute als Dissertation vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel (Literatur, Apparaturen, Material) in der Arbeit angegeben.

☒ Ich habe oder hatte die jetzt als Dissertation vorgelegte Arbeit nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.

☐ Ich hatte die heute als Dissertation vorgelegte Arbeit als Prüfungsarbeit für folgende Prüfung eingereicht:
Bezeichnung der Prüfung: _____

Prüfungsstelle: _____

☒ Ich hatte weder die jetzt als Dissertation vorgelegte Arbeit noch Teile einer Abhandlung bei einer anderen Fakultät bzw. einem anderen Fachbereich als Dissertation eingereicht.

☐ Ich hatte die folgende Abhandlung mit nachstehenden Ergebnis eingereicht:

Titel der Abhandlung:

_____

Fakultät bzw. Fachbereich und Hochschule:

_____

Ergebnis bzw. Beurteilung: _____

Mainz, den 05.01.2026 _____

(Unterschrift)

Curriculum vitae

Wissenschaftliche Veröffentlichungen

Publikationen

- Müller, A., Wichmann, A., Kallenborn, F., **Hellmann, S.L.**, Hankeln, T., & Schmidt, B. (2025). Improved Metagenomic Analysis for All-Food-Sequencing with AFS-MetaCache2: Illumina vs. Nanopore. BioRxiv, 2025.12.18.694891. https://doi.org/10.64898/2025.12.18.694891
- Poetzsch, G., Jelacic, L., Dammer, L., **Hellmann, S.L.**, Balling, M., Andrade-Navarro, M., Avivi, A., Shams, I., Bicker, A., & Hankeln, T. (2025). Adaptation of the Spalax galili transcriptome to hypoxia may underlie the complex phenotype featuring longevity and cancer resistance. npj Aging, 11(1). https://doi.org/10.1038/s41514-025-00206-3
- Seidel, M., Hamley-Bennett, C., Reeksting, B. J., Bagga, M., **Hellmann, S.L.**, Hoffmann, T. D., Kraemer, C., Ofițeru, I. D., Paine, K., & Gebhard, S. (2025). Metabolic Insights Into Microbially Induced Calcite Formation by Bacillaceae for Application in Bio-Based Construction Materials. Environmental Microbiology, 27(4). https://doi.org/10.1111/1462-2920.70093
- Petersen, M., **Hellmann, S.L.**, Böker, F., Brand, F., Sharma, A. K., Paleico, M. L., Köster, J., & Meesters, C. (2025). Teaching Reproducible Data Analysis for HPC Users — The Snakemake Teaching Alliance. Electronic Communications of the EASST, 83. https://doi.org/10.14279/eceasst.v83.2600
- Schmidt, H., Mauer, K., Glaser, M., Sayyaf Dezfuli, B., **Hellmann, S.L.**, Silva Gomes, A. L., Butter, F., Wade, R. C., Hankeln, T., & Herlyn, H. (2022). Identification of antiparasitic drug targets using a multi-omics workflow in the acanthocephalan model. BMC Genomics, 23(1). https://doi.org/10.1186/s12864-022-08882-1
- Mauer, K. M., Schmidt, H., Dittrich, M., Fröbius, A. C., **Hellmann, S.L.**, Zischler, H., Hankeln, T., & Herlyn, H. (2021). Genomics and transcriptomics of epizoic Seisonidea (Rotifera, syn. Syndermata) reveal strain formation and gradual gene loss with growing ties to the host. BMC Genomics, 22(1). https://doi.org/10.1186/s12864-021-07857-y
- Gutiérrez, Y., Fresch, M., **Hellmann, S.L.**, Hankeln, T., Scherber, C., & Brockmeyer, J. (2021). A multifactorial proteomics approach to sex-specific effects of diet composition and social environment in an omnivorous insect. Ecology and Evolution, 11(13), 8623–8639. https://doi.org/10.1002/ece3.7676

- **Hellmann, S.L.**, Kobus, R., Schmidt, B., Bikar, S., Köppel, R., Hankeln, T. (2020). All-Food-Seq: Next-Generation Sequencing-basiertes Screeningverfahren zur quantifizierbaren Speziesidentifikation in prozessierten Lebensmitteln. 8. Fachtagung Gentechnik – Neue molekularbiologische Techniken und deren Herausforderungen für die Analytik (Band 12).
- Kobus, R., Abuín, J. M., Müller, A., **Hellmann, S.L.**, Pichel, J. C., Pena, T. F., Hildebrandt, A., Hankeln, T., & Schmidt, B. (2020). A big data approach to metagenomics for all-food-sequencing. BMC Bioinformatics, 21(1). https://doi.org/10.1186/s12859-020-3429-6
- Mauer, K., **Hellmann, S.L.**, Groth, M., Fröbius, A. C., Zischler, H., Hankeln, T., & Herlyn, H. (2020). The genome, transcriptome, and proteome of the fish parasite Pomphorhynchus laevis (Acanthocephala). PLOS ONE, 15(6), e0232973. https://doi.org/10.1371/journal.pone.0232973
- Schmidt, H., **Hellmann, S.L.**, Waldvogel, A.-M., Feldmeyer, B., Hankeln, T., & Pfenninger, M. (2020). A High-Quality Genome Assembly from Short and Long Reads for the Non-biting Midge Chironomus riparius (Diptera). G3: Genes|Genomes|Genetics, 10(4), 1151–1157. https://doi.org/10.1534/g3.119.400710
- **Hellmann, S.L.**, Ripp, F., Bikar, S.-E., Schmidt, B., Köppel, R., & Hankeln, T. (2019). Identification and quantification of meat product ingredients by whole-genome metagenomics (All-Food-Seq). European Food Research and Technology, 246(1), 193–200. https://doi.org/10.1007/s00217-019-03404-y
- Liu, Y., Ripp, F., Köppel, R., Schmidt, H., **Hellmann, S.L.**, Weber, M., Krombholz, C. F., Schmidt, B., & Hankeln, T. (2017). AFS: identification and quantification of species composition by metagenomic sequencing. Bioinformatics, 33(9), 1396–1398. https://doi.org/10.1093/bioinformatics/btw822
- Oppold, A.-M., Schmidt, H., Rose, M., **Hellmann, S.L.**, Dolze, F., Ripp, F., Weich, B., Schmidt-Ott, U., Schmidt, E., Kofler, R., Hankeln, T., & Pfenninger, M. (2017). Chironomus riparius (Diptera) genome sequencing reveals the impact of minisatellite transposable elements on population divergence. Molecular Ecology, 26(12), 3256–3275. https://doi.org/10.1111/mec.14111

Workshops

- Vandendorpe, J., Lindstädt, B., & **Hellmann, S.L.** (2024). Workshop on Electronic Lab Notebooks (ELNs). Zenodo. https://doi.org/10.5281/zenodo.13318741
- Vandendorpe, J., & **Hellmann, S.L.** (2024). Demo with eLabFTW. TIB AV-Portal / NFDI4Microbiota. https://doi.org/10.5446/68306