foods

MDPI

Review

Step-by-Step Metagenomics for Food Microbiome Analysis: A Detailed Review

Jan Sadurski *, Magdalena Polak-Berecka, Adam Staniszewski and Adam Waśko

Department of Biotechnology, Microbiology and Human Nutrition, Faculty of Food Science and Biotechnology, University of Life Sciences in Lublin, 20-704 Lublin, Poland; magdalena.polak-berecka@up.lublin.pl (M.P.-B.); adam.staniszewski@up.lublin.pl (A.S.); adam.wasko@up.lublin.pl (A.W.)

* Correspondence: jan.sadurski@up.lublin.pl

Abstract: This review article offers a comprehensive overview of the current understanding of using metagenomic tools in food microbiome research. It covers the scientific foundation and practical application of genetic analysis techniques for microbial material from food, including bioinformatic analysis and data interpretation. The method discussed in the article for analyzing microorganisms in food without traditional culture methods is known as food metagenomics. This approach, along with other omics technologies such as nutrigenomics, proteomics, metabolomics, and transcriptomics, collectively forms the field of foodomics. Food metagenomics allows swift and thorough examination of bacteria and potential metabolic pathways by utilizing foodomic databases. Despite its established scientific basis and available bioinformatics resources, the research approach of food metagenomics outlined in the article is not yet widely implemented in industry. The authors believe that the integration of next-generation sequencing (NGS) with rapidly advancing digital technologies such as artificial intelligence (AI), the Internet of Things (IoT), and big data will facilitate the widespread adoption of this research strategy in microbial analysis for the food industry. This adoption is expected to enhance food safety and product quality in the near future.

Keywords: bioinformatics; foodomics; metagenomics; microbiome; food

check for updates

Citation: Sadurski, J.; Polak-Berecka, M.; Staniszewski, A.; Waśko, A. Step-by-Step Metagenomics for Food Microbiome Analysis: A Detailed Review. *Foods* **2024**, *13*, 2216. https://doi.org/10.3390/foods13142216

Academic Editor: Gabriele Rocchetti

Received: 11 June 2024
Revised: 11 July 2024
Accepted: 12 July 2024
Published: 14 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Consumer awareness has been on the rise in recent decades, leading to shifts in dietary preferences. This has resulted in an increased demand for more precise food research, as traditional microbiological culturing methods are limited in fully identifying microorganisms in food. These methods can only provide partial identification based on morphological and biochemical characteristics of culturable microorganisms, and they fall short in capturing unculturable microorganisms. Concerns have been raised regarding the incomplete microbial biodiversity picture presented by these techniques. Nevertheless, they continue to be the most commonly employed approaches in the food sector for gauging microbiological safety by identifying and describing the food's microbiota.

In the 21st century, culture-independent methods have emerged, allowing the bypassing of microbiological culturing and its constraints [1]. These procedures, which are centered on nucleic acid assessment, involve PCR-DGGE (polymerase chain reaction–denaturing gradient gel electrophoresis), T-RFLP (terminal restriction fragment length polymorphism), and FISH (fluorescence in situ hybridization) [2]. The landscape of food research has been transformed by "omics" technologies, which produce extensive data on the collective attributes of a sample regarding microorganism structure, function, and growth dynamics. High-throughput screening (HTS) is integral to these technologies, enabling rapid and extensive measurements. Next-generation sequencing (NGS) typifies HTS technology.

NGS has notably decreased analysis costs, accelerated sequencing speed, and enhanced result quality. It is categorized into second- and third-generation sequencing

methods. Technological advancements have steered a shift towards sequencing bacterial DNA as a mainstream approach in food microbiology research. NGS allows parallel mass sequencing of short sequence reads and individual long fragment sequencing. The emergence of HTS technology has spurred notable scientific development, culminating in the creation of four primary omics disciplines (genomics, transcriptomics, proteomics, and metabolomics) and their sub-disciplines (epigenomics, lipidomics, metallomics, etc.). By merging data from diverse omics areas, a new research discipline named "foodomics" has arisen, harnessing vast repositories of information [3].

"Foodomics" emerged as a distinct field in 2009, with a primary focus on investigating food and nutrition through omics technologies. The main goal of foodomics is to enhance food quality, thereby improving consumer health. The omics technologies employed in foodomics encompass nutrigenomics, proteomics, metabolomics, transcriptomics, and genomics [4]. Nutrigenomics specifically delves into how nutrients influence gene expression and aims to elucidate the interactions between bioactive food components and the genome at a molecular level, ultimately impacting gene expression. Nutrigenomics and nutrigenetics are often used interchangeably due to their close relationship [5]. Whereas nutrigenetics examines the correlations between single nucleotide polymorphisms (SNPs) and an individual's response to dietary intake, nutrigenomics utilizes nutrigenetics to present a holistic perspective of an individual's metabolism, with the intention of tailoring diets, preventing diseases, and mitigating life-threatening risks. This comprehensive understanding is facilitated by high-throughput sequencing techniques enabling a thorough examination of gene variations across the entire genome [6,7].

Proteomics delves into the study of proteins and their interactions within the cellular environment, which mirrors the constantly changing state of cells, tissues, and organisms. This field plays a crucial role in the identification of disease markers, as well as in the detection and accurate quantification of proteins, thereby enhancing our comprehension of disease causation. In contrast, metabolomics systematically identifies and measures all metabolites present in an organism.

Transcriptomics and genomics concentrate on the examination of nucleic acids found in biological specimens [8]. Transcriptomics specifically investigates gene expression at the RNA level, furnishing insights into the genetic makeup and functionality of genes across the entire genome to elucidate the molecular processes involved in specific biological functions. Genomics, on the other hand, is employed to conduct a thorough examination of an organism's genetic material, aiding in the identification of species present in food, the determination of the abundance of such microorganisms, and the detection of contaminants such as foodborne pathogens [9]. The amalgamation of these omics technologies in the field of foodomics facilitates a comprehensive understanding of the composition, safety, and nutritional characteristics of food, thereby contributing to advancements in both food science and consumer health.

An analysis was conducted on publications related to food metagenomics from 2018–2023, using VosViewer (Figure 1) [10]. The bibliographic data of articles were retrieved from the Web of Science database. The analysis involved filtering results based on the keywords "food", "metagenomics", and "quality". A co-occurrence network was established using 357 records from the past 5 years. The frequency of occurrences served as samples, resulting in 28 terms after excluding closely related ones. The input data for the network included full records and references cited. The size of each occurrence in the diagram corresponds to its frequency in publications. The color spectrum reflects the average normalized number of citations received by documents containing a specific term. Notably, articles featuring "shotgun metagenomics" and "database" garnered the highest number of citations, indicating researchers' interest in research methodologies within the realm of "food quality". Similarly, articles discussing untargeted sequencing details such as "amplicon sequencing", "whole genome sequencing", and "alignment" received significant attention, contrasting with the lower citation rate for articles mentioning "16S ribosomal RNA", a targeted sequencing method. The limited occurrences and citations for

terms such as "machine learning" and "extraction" suggest a research gap in the field of food metagenomics.

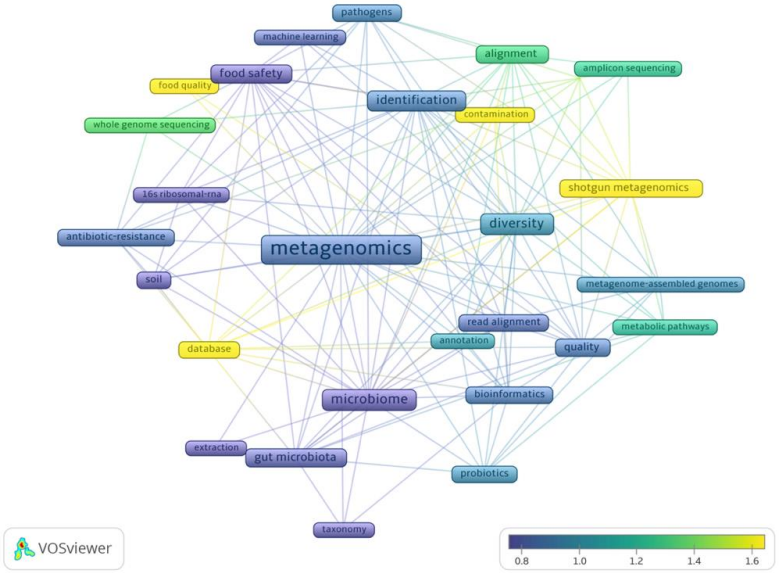

Figure 1. Co-occurrence of selected keywords in articles (2018–2023) using VOSviewer [10].

The objective of this review is to outline the individual processes involved in the analysis of food metagenomics, emphasizing its key aspects and presenting modern bioinformatic solutions.

2. Metagenomics

Metagenomics, a particular discipline within genomics frequently utilized in the field of food research, integrates genomic techniques with the principle of meta-analysis, which involves amalgamating and extrapolating findings from diverse studies through statistical approaches. The central emphasis of metagenomics lies in the sequencing and examination of the complete microbiota DNA, accompanied by metatranscriptomics that relies on cDNA (complementary DNA) sequencing [11]. Metagenomics obviates the need for the isolation and cultivation of microorganisms in vitro prior to DNA extraction. The focal point lies in the examination of the comprehensive genetic makeup of microorganisms existing in various natural specimens, such as food or the surrounding environment. Metagenomics encompasses scrutiny of the combined genome of microorganisms inhabiting a specified milieu, thus facilitating a comparative assessment of the microbiome through the utilization of NGS. The sequencing process can be directed towards a specific gene of interest, such as the 16S rRNA in bacteria, 18S in eukaryotes, or intergenic regions/internal transcribed spacers in fungi, or it can entail an untargeted approach known as "shotgun" sequencing, encompassing the sequencing of all genomes found within the sample, referred to as whole-metagenome sequencing (WMS) [12–14].

Sequencing can be executed through either of two methodologies: shotgun sequencing or targeted sequencing. Within the framework of shotgun sequencing, the DNA molecules undergo a random fragmentation process, resulting in small DNA pieces that are subsequently sequenced comprehensively. This particular approach is frequently applied in research endeavors pertaining to the characterization of microbial populations within metagenomic undertakings, aiming to discern microorganisms at a granular strain level. Conversely, targeted sequencing is exclusively concerned with the sequencing of a predetermined genomic region.

Shotgun sequencing, in comparison to targeted sequencing, provides additional insights into the functionality of the microbiome, its plasticity, and ongoing biological processes, as well as sequence variations and evolutionary variability. It also enables the identification of organisms at a higher taxonomic resolution [15], meaning at the lowest

possible difference between organisms (taxonomic level). Access to complete genomes, rather than being confined to a solitary 16S/18S gene, is the reason for this phenomenon. The efficacy of sequencing the 16S gene persists in investigations of microbiota abundant in uncharacterized microorganisms. Research that focuses on sequencing for a specific microorganism or marker gene does not fall under the category of metagenomics, as it does not encompass the entirety of the genetic material in the specimen. Apart from the consistent reduction in sequencing expenses over time, one challenge of whole-metagenome sequencing (WMS) relates to the hardware prerequisites and intricate data interpretation. A strategy to address the high costs associated with WMS involves a two-stage approach. Initially, cost-effective targeted sequencing (16S rRNA) is carried out as a preliminary assessment, followed by untargeted sequencing on a chosen subset of specimens [13].

Metagenomic shotgun sequencing comprises both wet and dry phases. The wet phase denotes a laboratory procedure encompassing two steps: (i) acquiring and preserving samples and (ii) conducting sequencing. Conversely, the dry phase pertains to the computational handling of data derived from sequencing. Optimization of each phase in the analytical process is imperative, tailored to the specific material under investigation and the research goals.

3. Sampling and Storage

The most favorable approach for collecting samples entails striking a balance and the volume of samples needed, the frequency of repetitions, and the practical and financial viability of performing analyses [16]. The quantity of samples and repetitions significantly influences the precision, defined as the level of concordance between outcomes; accuracy, representing the level of adherence to the true condition; and the reproducibility of findings. The strategy for determining sampling sites with respect to the number of repetitions differs depending on the origin of the samples. Within the food sector, difficulties emerge concerning the optimal sample selection and the determination of repetition quantities. The microbiota of raw materials experiences rapid modifications throughout processing, even within brief time frames. Variations in the microbiota's composition may be either straightforward to anticipate (e.g., milk after pasteurization) or unforeseeable, linked to other technological elements (e.g., frequency of equipment sanitation) [17]. The detailed procedure for sampling and DNA isolation from raw milk and cheese was developed by C. Barcenilla [18]. The uncertainty of temporal changes adds complexity to determining the optimal number of repetitions. Furthermore, the intricacy of the food production process presents additional hurdles. For instance, the production of cheese entails multiple phases carried out using distinct apparatus. Comprehending the production procedure is crucial for establishing the necessary quantity of sample collection points. Acquiring control samples in an environment with a variable microbiota composition, influenced by various external factors, can prove challenging. In such instances, it is advisable to substitute cross-sectional research, which compares the microbiota at a single time point, with prolonged investigations involving the analysis of samples from the same setting over an extensive duration. Prolonged studies do not depend on individual outcomes, which might deviate from the standard, and enable the removal of samples impacted by unfavorable alterations. If potential confounding variables cannot be ruled out, they should be factored into the comparative assessment [19]. A crucial aspect of collecting metagenomic samples is the preparation of detailed and standardized metadata. These are essential for comparative studies and result generation. Adopting a reporting standard enhances the quality, accessibility, and usefulness of information that can be stored in data repositories. The Genomic Standards Consortium (GSC) has proposed standards for the minimal information about genomic sequences (MIGS) and metagenomic sequences (MIMS). Additionally, these standards have been further detailed to include an environmental package comprising a set of measurements and observations describing the habitats from which the samples were collected. The environmental package includes sampling information such as geographical data on the location (country, region, latitude, and lon-

gitude), date of collection, environment, and material type [20]. It is advisable to gather a wide array of parameters, particularly the attributes specific to a given environment, in order to enhance the probability of establishing correlations between outcomes and a particular environmental factor [19]. A summary of sample fermented products along with the most commonly used DNA isolation kits and the required sample quantity is presented in Table 1.

Table 1. Fermented products and methods of DNA isolation.

Product	Sampling Detail	DNA Isolation	References
Cheese	1 g sample	PowerFood Microbial DNA Isolation Kit	[21–23]
Kombucha	50 mL sample	PureLink Microbiome DNA Purification Kit	[24–26]
Kefir	3–5 g sample	DNeasy PowerSoil Kit	[27,28]
Yogurt	10–20 mL sample	QIAamp Fast DNA Stool Mini Kit	[29,30]
Kimchi	3 mL sample	QIAamp Fast DNA Stool Mini Kit	[31–33]
Sauerkraut	1.2 mL sample	FastDNA™ S PIN kit for Soil	[34,35]

4. Sequencing

Among the second-generation NGS technologies, the 454/Roche and Illumina/Solexa platforms are commonly utilized in metagenomic research. Key features of these technologies include the simultaneous production of millions of brief reads, decreased sequencing duration, reduced expenses in contrast to first-generation sequencing, and the capacity to acquire immediate outcomes. The advent of the third generation of NGS technologies, specifically long-read technologies such as PacBio and Oxford Nanopore, carries substantial implications for metagenomic investigations, especially in the process of genome assembly [36]. Long reads, abundant in valuable data, aid in de novo assembly and alignment with a reference genome (associating the sequenced genetic material with a recognized reference genome) [37]. These technologies facilitate the production of sequences measuring 10 kbp in length, achieving an accuracy rate of 85–87% for PacBio [38] and 88–94% for Nanopore [39]. This approach proves to be cost-efficient as it eliminates the need for extensive sample preparation procedures, thereby expediting the generation of outcomes in non-specialized laboratory settings. Nevertheless, a notable drawback of third-generation NGS lies in the inadequacy of bioinformatics resources tailored for the interpretation of lengthy sequences. Existing tools are predominantly optimized for the comparative assessment of precise data derived from short genetic sequences. Table 2 presents a summary of sample fermented products, target sequences, and the most commonly used sequencing technologies.

Table 2. Fermented products and methods of sequencing.

Product	Target	Sequencing Platform	References
Cheese	16S rRNA V3-V4 16S rRNAV4 MGS	Illumina MiSeq Illumina MiSeq Illumina HiSeq	[40–43]
Kombucha	16S rRNA V1-V9 MGS MGS 16S rRNA V1-V9	Oxford Nanopore Technologies MinION Illumina HiSeq Illumina Novaseq 6000 Illumina NextSeq 500	[44–47]

Table 2. *Cont.*

Product	Target	Sequencing Platform	References
Kefir	MGS 16S rRNA V3-V4	Illumina HiSeq Illumina MiSeq	[48–51]
Yogurt	MGS 16S rRNA V2, V4, V6, V7, V8, V9 16S rRNA V2-4-8, V3-7-9	Illumina HiSeq Ion GeneStudio S5 Ion Torrent PGM	[52–55]
Kimchi	16S rRNA V3-V4 16S rRNA V1-V3	Illumina MiSeq Roche 454 GS-FLX Plus	[56,57]
Sauerkraut	16S rRNA V3-V4	Illumina MiSeq Illumina NovaSeq	[58–61]

5. Bioinformatic Processing

Working with sequences obtained using HTS involves a few manipulations of raw data performed by various programs in order to generate the desired final results. These procedures can be divided into 3 parts: (i) first-level analysis; (ii) second-level analysis, and (iii) integrating results with metadata. Figures 2 and 3 present the subsequent stages of first- and second-level analyses.

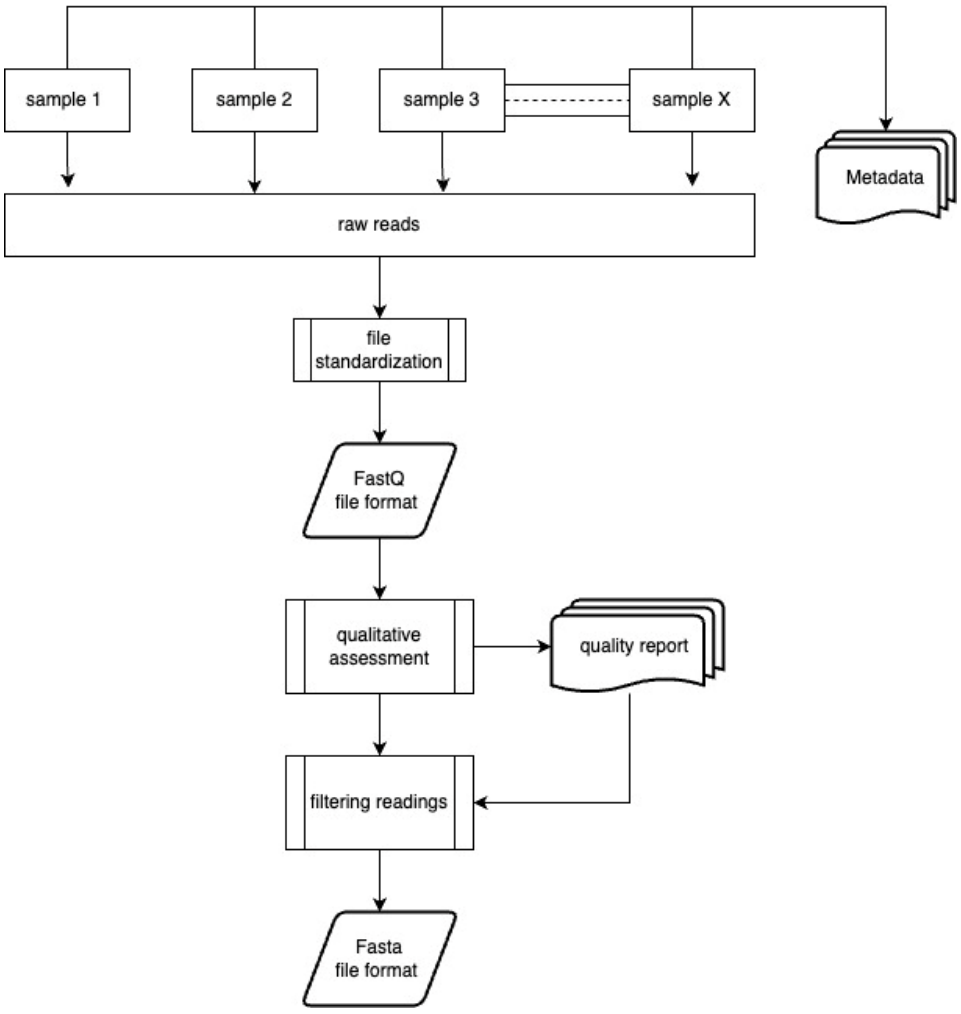

Figure 2. First-level analysis.

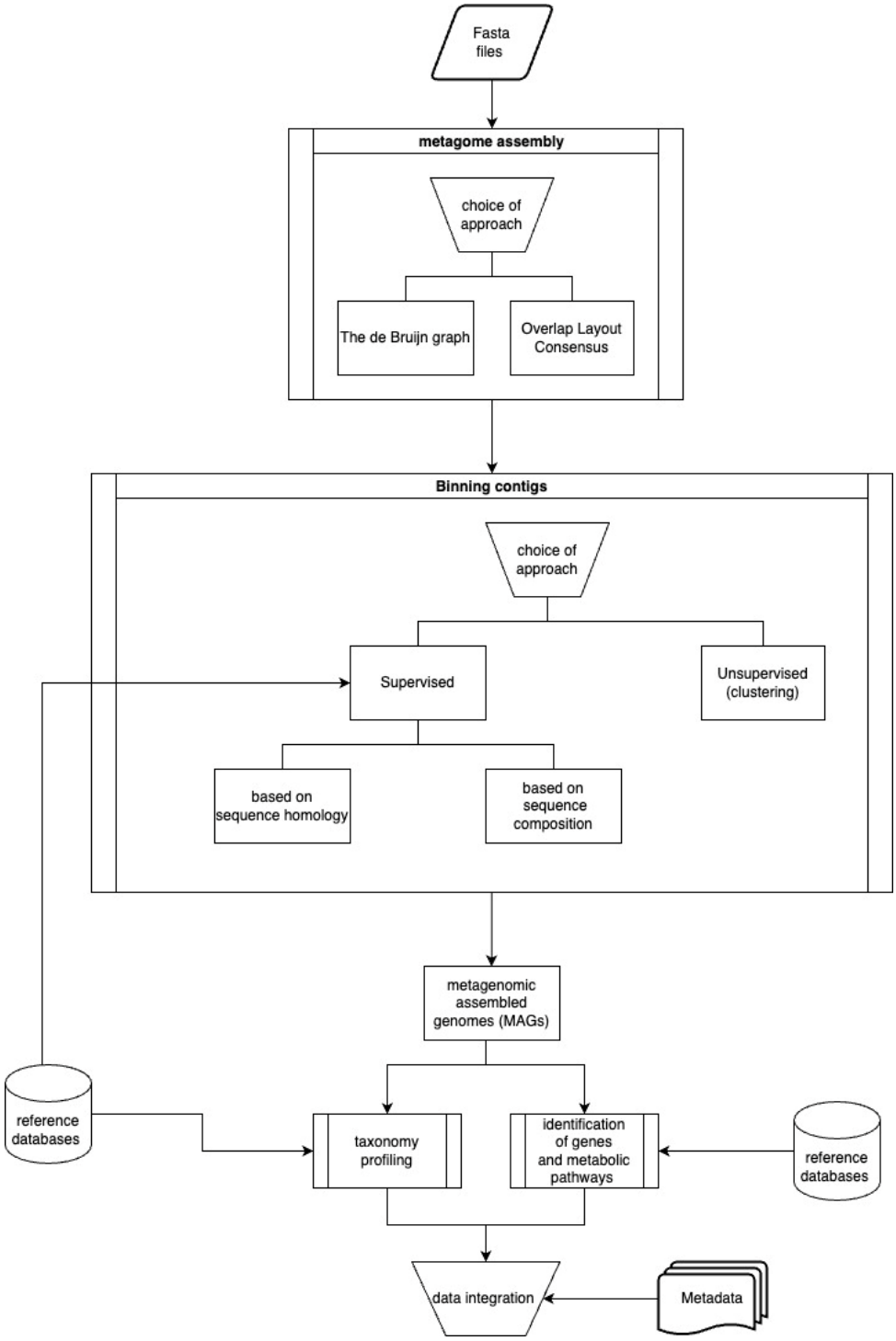

Figure 3. Second-level analysis and integrating results with metadata.

5.1. Quality Assessment and Filtration of Readings

The objective of the initial processing stage is to enhance the overall quality of sequence files prior to their utilization in analytical procedures. Enhancing quality entails the filtration and elimination of low-quality sequences, along with the exclusion of any adapters that may be present. Quality evaluation is conducted at the individual base level utilizing the PHRED scale score, which indicates the likelihood of incorrect base assignment. Illumina-generated data typically exhibit high quality (Q30–40) at the onset of the

read, gradually declining towards the read's conclusion. Bases with a quality score below Q15–Q20 towards the end of the read are deemed insufficiently accurate for interpretation.

Depending on the employed sequencing methodologies, data are presented in various file formats. Assessment of the quality of reads can be conducted utilizing files in the FastQ formatting. Within the FastQ configuration, sequences and quality outcomes are delineated through individual ASCII characters. Each sequence comprises of four lines arranged one above the other. The initial line commences with the "@" symbol, followed by the sequence identifier (e.g., information on flow cell ID, read pairing). The succeeding line depicts the nucleotide sequence. The subsequent line initiates with the "+" symbol, succeeded by the identical sequence identifier as in the first line, denoting the conclusion of the sequence. The final line illustrates the quality of the sequence, with a solitary character encoding the quality (PHRED score) of a specific base within the sequence.

Files in the FASTQ format (Illumina, 454/Roche) can undergo qualitative evaluation through the utilization of the FASTQC software from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 13 July 2024). FASTQC is responsible for generating a total of ten visual representations aimed at assessing the quality of the file. The outcomes are stored in HTML format and are amenable to visualization through a web browser. Subsequent to this, the process of filtering results based on suitable quality parameters can be executed by employing trimming software such as SeqTrim [62], EA-Tools [63], and Trimmomatic [64].

Data acquired through the utilization of Nanopore technology are documented in the FAST5 file configuration, which encompasses raw signal information. The aforementioned data have the potential to be transformed into the FASTQ format by employing suitable software applications such as Guppy [39], Fast-Bonito [65], Causalcall [66], and NanoOK [67]. Guppy provides two distinct analysis models: rapid and high-precision. The mean quality of sequences produced by Nanopore falls within the range of Q7 to Q14, with the quality exhibiting fluctuations throughout the reading process. All numerical values in this context are highly significant.

PacBio sequences are archived in the Binary Alignment Map (BAM) format, a format that does not include annotations on sequence quality. The transformation to FASTQ format can be executed through software tools such as SAMTOOLS [68] or the SMRT portal [69]. Furthermore, the SMRT portal not only conducts demultiplexing (the inverse operation of multiplexing, which involves segregating signal components) but also eliminates hairpin adapter sequences from the reads and sieves out reads characterized by superior quality.

5.2. Contig Assembly De Novo

The process of de novo assembly involves the grouping of reads into contigs. Multiple techniques exist for determining the makeup of a multi-species microbial population from a set of sequence reads. When it comes to the methodology of metagenome assembly, the process is akin to piecing together individual whole genomes [68]. Different computational strategies are utilized for reconstructing the composition of microbial communities from a set of sequence reads, with the selection of approach being dependent on the objectives of the study. There are two primary methods for assembling contigs: (i) the utilization of a de Bruijn graph (DBG), and (ii) the alignment of overlapping OLC (overlap/layout/consensus) reads.

The prevalent approach utilized is based on the de Bruijn graph. This graph is formulated by disintegrating each read into overlapping fragments of a consistent length, known as k-mers. These k-mers establish the vertices and edges of the de Bruijn graph. Subsequently, the software traces a route within the graph, consequently ascertaining the accurate genomic sequence. The emergence of branches in the graph, which complicates the identification of the correct sequence, is attributed to sequencing inaccuracies, fluctuations in coverage, the existence of repetitive sequences, and various other structural variations [69].

A less frequently utilized approach entails the superimposition of reads, where the identification of overlapping sequences is achieved through the comparison of each read with all other reads. The overlapping reads are then categorized into contigs. Subsequently, a continuous sequence is established by choosing the most probable nucleotides from the overlapping contigs. However, a limitation of this method lies in the need to compare each read with every other read within the dataset, and in high-throughput sequencing (HTS) methodologies, the number of reads can reach millions. Figures 4 and 5 illustrate the techniques utilized in contig assembly.

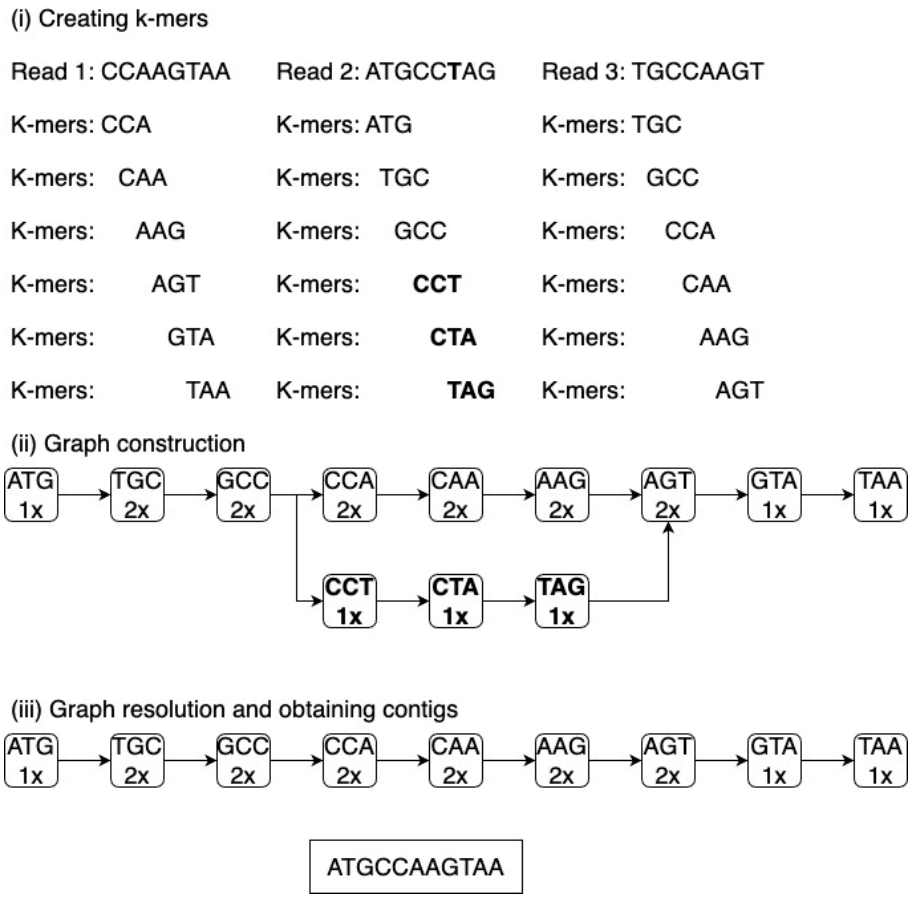

Figure 4. The de Bruijn graph. De Bruijn graph (**i**) reads are decomposed into k-mers by sliding a window of length k along the read. (**ii**) k-mers become nodes in the graph, and edges connect overlapping k-mers. Polymorphisms (highlighted in bold) create branches in the graph. The count of each k-mer's occurrences is indicated (number above the k-mer). (**iii**) Contigs are built by traversing the graph between branching nodes. Algorithms may interpret branches differently; the example shown ignores low-coverage paths.

Assembling metagenomes poses a greater challenge compared to assembling individual genomes due to the absence of uniform sequence coverage throughout the genome. The coverage of each genome present in the metagenome is contingent upon the prevalence of microorganisms within the environment under study. Genomes that are poorly represented may experience fragmentation in cases where sequencing depth is insufficient [70]. The reduction in k-mer length utilized for graph construction may facilitate the reconstruction of less prevalent genomes, albeit with the drawback of heightened repeats that impede precise genome assembly [71]. An additional complexity emerges from the examination of closely related genomes exhibiting variations in individual genes or nucleotides, resulting in graph branching. The occurrence of branching within the graph contributes to the fragmented reconstruction of genomes.

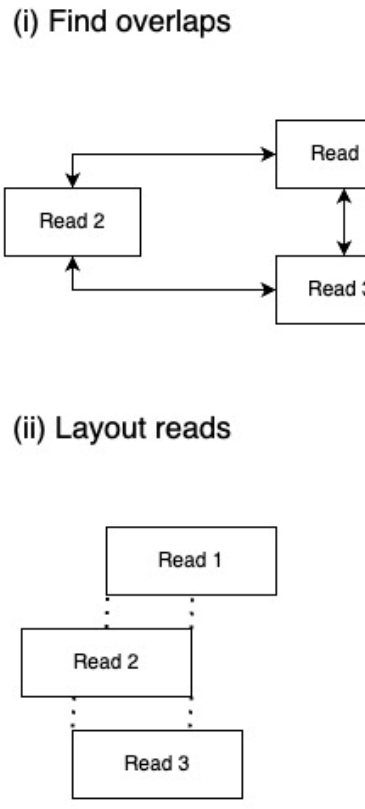

Figure 5. Overlap, layout, and consensus. Overlapping reads; (**i**) searching for overlapping read fragments. (**ii**) Layout reads determine the assembly of reads into contigs by considering coverage (dashed lines indicate overlapping fragments). (**iii**) The most probable nucleotides are selected for sequence construction.

Diverse strategies are being devised to tackle the challenges linked to metagenome assembly. Meta-IDBA [72] and RAMPART [73] employ varied k-mer lengths, eliminating the necessity of selecting a singular averaged k-mer length, thereby facilitating the assembly of genomes with diverse abundances. Additionally, RAMPART produces a concise overview for each assembly. In metagenome reconstruction, Meta-IDBA also accounts for irregular sequencing depth [74]. MetaSPAdes software allows hybrid metagenome assembly by utilizing short and long sequences acquired from different technologies [75]. Samples containing intricate microbiota compositions often encompass numerous closely related strains with varied abundances. Augmented sequencing depth enables their identification but demands substantial computational resources and time, which may prove inadequate. The MEGAHIT tool confronts this challenge by employing streamlined data structures to diminish memory requirements and expedite the analysis process when assembling intricate metagenomes [76]. A decentralized metagenome assembly framework harnesses

Ray software [77] to allocate memory load across individual machines. Genovo diminishes analysis duration via advanced learning methodologies [78].

The initial categorization of reads into potential taxonomic clusters by comparing them to known genomes, and subsequently excluding them from the read collection, can optimize the process of constructing intricate metagenomes. MEGAHIT differentiates reads into well-defined and ambiguous categories based on their coverage relative to reference genomes. Even reads with limited coverage could be integrated into the metagenome assembly if they complement well-established contigs [76]. MetaCRAM [79] leverages Kraken [80], a k-mer-centric tool for taxonomic classification, to assign reads to reference genomes initially and eliminate any familiar sequences from the dataset before assembly. VICUNA software [81] facilitates the elimination of non-target reads through multiple sequence alignment (MSA) of the sequences.

Paired-end sequencing reads, encompassing both brief and extended fragments, offer significant advantages in the process of constructing individual genomes from scratch, yielding insights into the interconnections among divergent contigs and facilitating the formation of scaffolds. The utility of paired-end reads is less evident when dealing with metagenomic datasets. Several software applications designed for metagenome assembly adopt a methodology akin to that employed for individual genome assembly, involving the creation of scaffolds. Utilizing paired-end reads, tools such as MEGAHIT [78], BIGMAC [82], SPAdes [77], PRICE [83], and Omega [84] aim to identify and eliminate chimeric contigs resulting from the erroneous merging of distinct genomes, thereby enhancing the quality of the resultant assemblies. A comparative analysis of contig assembly software is presented in Table 3.

Table 3. Programs assembling contigs.

Program	Method (dBG—de Bruijn Graph; OLC—Overlap Layout Consensus)	Characteristic Feature	Publications
Genovo	OLC	It employs deep learning; randomly selects contigs for matching reads	[78]
IDBA-UD	dBG	It breaks down the graph locally at each depth.	[72,74]
MEGAHIT	dBG	K-mers split based on identification with reference genomes.	[76]
Omega	OLC	Scaffolding using long reads; unmatched contigs are grouped based on coverage.	[85]
Price	Hybrid	Identical reads are assembled first, followed by less similar ones.	[86]
Ray	dBG	Distributed program connected to the network; profiles the microbiome based on unique labeled k-mers.	[77]
SPAdes	dBG	The metaSPAdes extension utilizes stream processing to resolve the graph.	[75]

5.3. Contig Clustering

Complex metagenomes, characterized by extensive fragmentation resulting in thousands of contigs, pose challenges in determining the number of genomes present in the dataset and the assignment of contigs to specific genomes. The process of contig clustering aims to categorize these fragments into distinct groups representing individual species. This clustering endeavor leads to the reconstruction of components of intricate metagenomic genomes, referred to as metagenomic assembled genomes (MAGs). Subsequently, contigs belonging to each cluster are stored in separate files formatted in FASTA. FASTA-formatted data comprises sequences presented in a single line of text, with descriptions on

subsequent lines initiated by the ">" symbol. Ideally, each file corresponds exclusively to a single genome.

Currently employed clustering methods can be divided into two categories: (i) supervised and (ii) unsupervised. Both methods assess the similarity between contigs and sets, subsequently transforming these similarities into assignments.

The method of supervised learning makes use of reference databases to categorize contigs into distinct taxonomic categories. Moreover, a distinction can be made within this method between sequence homology-based strategies and sequence structure-based strategies. Various tools that utilize sequence homology, such as PhymmBL [82] and BLAST [83], rely on vast and inclusive databases. Software programs such as PhyloPythiaS+ [84], EnSVMB [87], and Kraken [80] utilize k-mers to either compare sequences or generate patterns, ultimately reducing the time taken for analysis. The k-mer approach mandates the development of specific reference databases, while pattern generation necessitates the use of training files. The examination of metagenomes which encompass a multitude of genomes, lacking any reference databases, presents a significant challenge. The absence of reference genomes hinders the creation of training files. The substantial diversity of species within the files requires the generation of a larger quantity of patterns, consequently extending the time required for analysis.

Unsupervised clustering techniques aim to identify intrinsic variations within the examined dataset. The genomes of diverse organisms exhibit distinct arrangements of nitrogenous bases, which are manifested by discrepancies in the occurrence of k-mers [88]. The use of tetramers is considered to be most effective for clustering metagenomic information [89]. The process of grouping contigs is employed in various automated tools such as SCIMM [90] and MegaWatt [91], which rely on parameters related to species distribution and DNA sequence representation. Additionally, more advanced automated software such as MetaBAT2 [92], GroopM [93], and CONCOCT [94], as well as semi-automated algorithms that involve human evaluation, utilize contig clustering for analysis [95].

The hybrid method BMCCR (Binning Metagenomic Contigs using unsupervised Clustering and Reference databases) introduces a novel approach that integrates the benefits of two distinct methods, as outlined in a previous study [96].

5.4. Quality Assessments of MAGs

The quality of MAGs depends on the genome size of the species, its abundance in the environment, and the sequencing depth. Parameters determining MAG quality include completeness and the degree of genome contamination. CheckM [97] utilizes a set of marker genes, along with information about their positions in reference genomes, and then utilizes information about the co-localization of these genes in the studied genomes. The program initially places individual MAGs in the reference tree to adapt the set of marker genes to the specific lineage.

The optimal scenario entails acquiring a solitary contig within a document of comparable length to the taxonomically aligned genome. Nevertheless, the attainment of such a scenario is practically unattainable, hence, it becomes imperative to employ criteria that evaluate the assembly's quality. One such criterion is N50, while another is L50. The quality of a MAG is deemed superior when N50 exhibits a higher value and L50 displays a lower value for the genomes being analyzed. The elucidation of parameters N50 and L50 is depicted in Figure 6.

Contigs sorted by length in descending order with a total length of 350 kbp. The N50 parameter for this sequence is the length of the contig located at the midpoint of the sequence–70 kbp. The L50 parameter is the position of the contig located at the midpoint of the sequence–3.

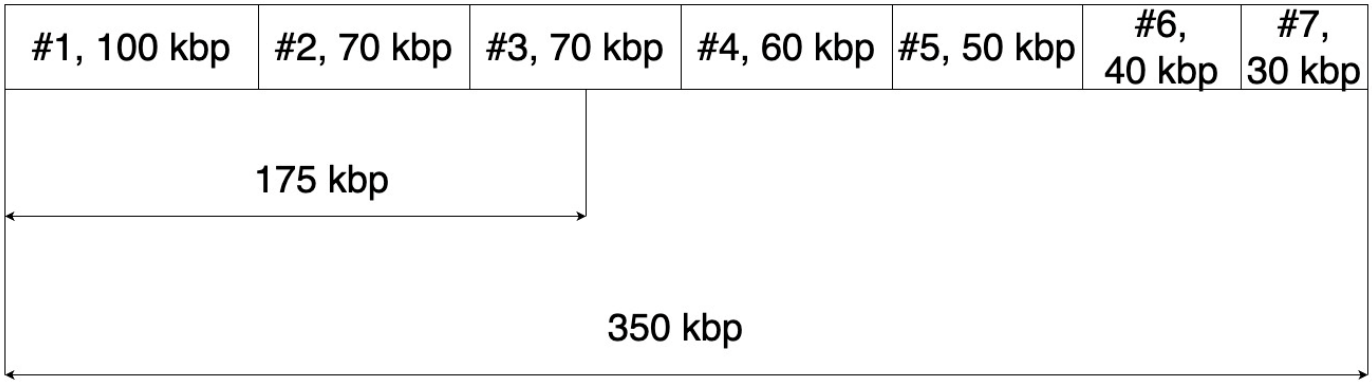

Figure 6. Construction of a MAG.

5.5. Defining and Analyzing the Pangenome

The pangenome denotes the entire array of genes discovered in diverse strains of a certain species, as ascertained through the utilization of comparative genomics examinations. This pangenome is structured hierarchically, comprising a genomic core, an accessory genome, and single genes. The genomic core consists of genes present in all strains of the species analyzed and is primarily responsible for crucial functions necessary for the microorganism's survival [98], including those related to pathogenicity and virulence [99]. The accessory genome encompasses genes found in at least two strains but not in all [100], while single genes are those present in only one strain. Genes within the accessory genome or single genes can be acquired through processes such as horizontal gene transfer or mutations evolving in other genes. These genetic elements are crucial in facilitating adjustment to the surroundings, such as distinct biochemical pathways, harmfulness characteristics, and drug resistance [101]. The configuration of the pangenome can either be open or closed, depending on the likelihood of identifying novel gene families through the inclusion of genomes for comparative study. An open pangenome, when more genomes are added, tends to exhibit a growing diversity of gene families. Conversely, a closed pangenome remains stable in terms of the number of gene families present.

Pangenomic analysis entails a three-step procedure: (i) standardizing annotations, (ii) categorizing genes by orthology, and (iii) adjusting curves. Standardizing annotations in the initial stage aims to avoid misidentifying core genes as universal and incorrectly assigning universal genes to individual genes. This task typically utilizes genome annotation tools such as RAST [102] and Prokka [103]. At the second phase, a comprehensive table containing all orthologous genes is acquired. In this phase, software programs such as OrthoMCL [104] and Orthofinder [105] are utilized. This table enables the fitting of the specific curve that arises from permutations of all genomes at various positions during the third step. The alignment process incorporates Heaps' law and the power law, while the alignment of the curve of the common genome and individual genes is accomplished through an exponential regression distribution. Table 4 showcases instances of pre-existing bioinformatics solutions that are capable of carrying out all three phases of analysis.

Table 4. Programs designed for the analysis of pangenomes.

Software	Orthology Analysis	Pangenome Construction	References
BPGA	CD-HIT, OrthoMCL	Power-law regression	[106]
EDGAR 2.0	Score ratio values	Heaps' law	[107]
GET_HOMOLOGUES	COGtriangles, OrgoMCL	Plot_pancore_matrix.pl	[108]
PanWeb	PGAP	PGAP	[109]

Table 4. *Cont.*

Software	Orthology Analysis	Pangenome Construction	References
PGAP	MultiParanoid, Gene Family	Heaps' law	[110]
Roary	CD-HIT, BLAST, MCL	(Not mentioned)	[111]

5.6. Taxonomic Profiling

The process of taxonomic profiling entails the assignment of operational taxonomic units (OTUs) to individual contigs. The primary objective of profiling is to ascertain the species composition of a metagenome and to estimate the representation of each species. There are two predominant strategies utilized for taxonomic identification: (i) aligning sequences using databases such as BLAST [83] and (ii) employing k-mers, exemplified by Kraken [80]. Kraken makes use of databases that house sequences fragmented into k-mers to seek out distinct fragments based on taxonomic categories, ranging from the lowest common ancestor (LCA) to the target species. The software dissects the queried contig into k-mers and subsequently assigns them to the most likely position in the reference taxonomic tree. A notable result of profiling is the prevalence of each OTU in the sample. The prevalence of individual species can be expressed in relative or absolute counts of OTUs [112].

5.7. Construction of Phylogenetic Trees from Metagenomic Data

The procedure for constructing a phylogenetic tree is a multi-step process. It consists of the following: (i) orthology prediction (orthologous genes, i.e., genes whose relationships will reliably reflect species relationships), (ii) alignment, (iii) identification of outliers, (iv) site filtering, and (v) phylogenetic inference.

There are bioinformatics solutions that perform all these steps within their scope. These include: PhyloPhlAn [113], PhyloSift [114], ezTree [115], GToTree [116], and AMPHORA [117], which rely on the analysis of specific genes, their sets, or specific regions in the genome [118]. Additionally, there are algorithms responsible for specific stages of the analysis, such as algorithms for multiple-sequence alignment (MSA): MUSCLE [119], MAFFT [120], T-Coffee [121], OPAL [122], PASTA [123], and UPP [124], and phylogenetic reconstruction algorithms such as FastTree [125,126], RAxML [125,127,128], ASTRAL [129], ASTRID [130], and IQ-TREE [131]. Each algorithm performs analyses separately or sequentially, requiring the researcher to have substantial knowledge in identifying appropriate targets, parameters, and steps of computational phylogenetics. A detailed review on obtaining phylogenetic trees using a non-automated multi-step method has been presented by P. Kapli in his article [132].

The separate and human-monitored execution of these processes is not feasible, especially when a large quantity of genomes are collected and analyzed together. Efficient algorithms have been suggested, such as those utilizing a small number of representative marker genes, such as the multilocus sequence typing (MLST) method or core genes at the species level. Computational MLST can function rapidly by employing only five to ten loci for each species. An example of an MLST-based program is chewBBACA. chewBBACA is capable of constructing phylogenetic trees using whole-genome MLST (wgMLST) or core genome MLST (cgMLST) [133]. Nevertheless, this is achieved at the cost of significantly reduced accuracy in phylogenetic placement. Pangenome-based profiling, exemplified by Roary [111], excels in accurate phylogenetic modeling at the species level but cannot be applied broadly to higher clades. Phylogenies that isolate strains and incorporate thousands of reference genomes from various species—or at least those most closely related to new sequences—result in a more precise depiction of microbial population structures and traits, aiding in more precise taxonomy.

5.8. Determining Gene Functions and MAG Metabolic Profiles

For fragmented yet of high quality MAGs, it is possible to establish a metabolic profile. The process of genome annotation can be conducted using two primary methodologies. One strategy entails the identification of genes along with their respective functions; however, this approach is constrained by a vast reservoir of genes in databases that lack characterization. The alternative approach involves a translational search for proteins affiliated with specific functional categories. Databases such as UniProt [134] and KEGG [135] offer both annotation services and insights into the categorization of proteins into distinct functional clusters and metabolic pathways. The annotation outcomes are typically displayed graphically or in the format of a TSV file containing numerical data regarding the presence or absence of particular pathways. A notable limitation in metabolic profiling lies in the absence of scrutiny of accessory genes, thereby leading to the recognition and quantification of primary metabolic pathways within a comparable framework. Despite variations in microbiological and environmental compositions, samples demonstrate analogous functional attributes [136].

Moreover, the process of identifying genes can be further advanced through the utilization of specific software designed for identifying virulence genes, such as MetaPhinder [105], or genes associated with antibiotic resistance [106].

5.9. Integrating Metagenomic Data with Metadata

The obtained data on the microbiological characteristics of the metagenome are processed using statistical tools for interpretation and exploration of correlations with the collected metadata of the samples. R, a popular programming language, is commonly employed for statistical analysis. It encompasses a range of packages dedicated to both metagenomics and genomics, which can be adapted for metagenomic purposes. Detailed packages have been described by Calle in his review article [112].

6. Software for Comprehensive Metagenomic Analyses

There exist pre-configured software bundles with predetermined parameters for every phase of analysis. These are characterized by a higher level of user-friendliness attributed to a graphical user interface (GUI), thereby obviating the necessity of employing a text-based interface. In addition, these bundles encompass an exhaustive array of tools essential for multifaceted data analysis, visualization, and interpretation. The setup process of such bundles is streamlined through an integrated automated installer that is compatible with various platforms. Illustrative instances of such software bundles encompass Parallel-Meta Suite [82] and EPA-ng [137]. One drawback of these solutions pertains to the limited adaptability in adjusting parameters at specific analysis stages, a factor that could prove pivotal in achieving the desired outcomes.

Mothur [138] and Qiime2 [139] are widely used solutions in bioinformatics analysis, offering flexibility in parameter adjustments. Mothur, initiated in 2009 by Dr. Patrick Schloss at the University of Michigan, provides an integrated platform for ecological research through a command line interface, while Qiime2, designed with a plugin system, allows operation via API, graphical interface, and command line for decentralized use. Figure 7 illustrates a graph presenting the citation frequencies of Qiime2 and Mothur in scholarly articles from the last 5 years, based on data obtained from PubMed on 7 July 2024.

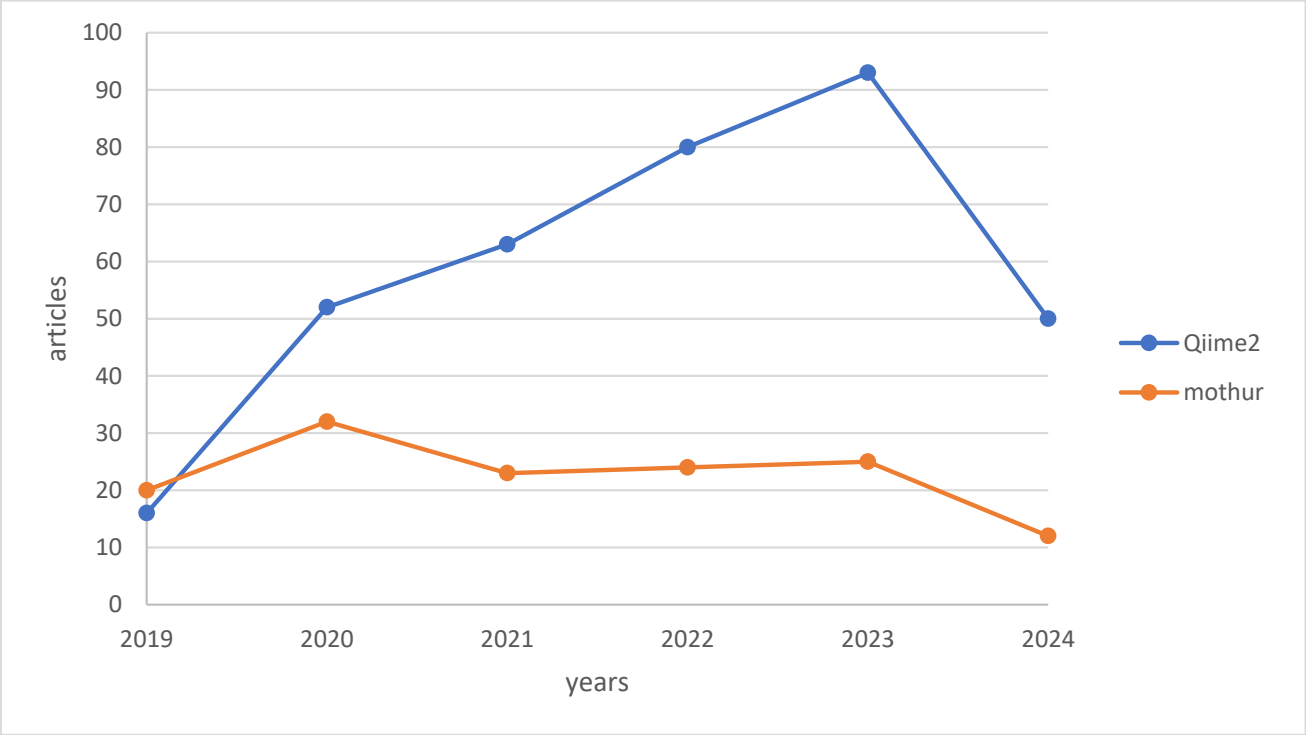

Figure 7. Comparison of citation counts: QIIME2 vs. Mothur.

7. Conclusions

"Foodomics" represents a novel research approach within the realm of food studies, demonstrating considerable potential for application in the realm of food production. This methodology facilitates a profound examination of food microbiota through the utilization of contemporary, cost-effective, and efficient DNA sequencing techniques. Through metagenomic scrutiny, a precise taxonomic classification can be achieved, pinpointing individual species and strains. Furthermore, this approach allows the identification of specific gene functions that could influence the production process and, consequently, the sensory attributes of the final product. The acquisition of dependable and replicable outcomes necessitates the consideration of various factors during the design phase of studies. Fundamental elements encompass the approach to sample collection, storage, and pre-sequencing preparation, alongside subsequent bioinformatic scrutiny. The formulation of optimal and consistent sampling techniques, coupled with the meticulous documentation of pertinent environmental variables specific to distinct production procedures, is imperative. The realm of bioinformatic solutions geared towards comprehensive microbiota analysis is in a perpetual state of evolution, being continually generated, refined, and enhanced. At present, a majority of existing programs lack a user-friendly graphical interface, thereby heightening initial operational complexities. In the forthcoming years, bioinformatic solutions featuring intuitive graphical interfaces will emerge. Software integrating deep learning methodologies will streamline the analysis timeframe and diminish hardware prerequisites. Nevertheless, systems leveraging deep learning necessitate a training phase, mandating substantial resources and time investments. The efficacy of bioinformatic tools is contingent upon the breadth of reference databases, underscoring the need for their continual expansion to encompass newly unearthed genes and proteins. The evolution of foodomics towards heightened accessibility and efficacy paves the way for its integration into the commercial sphere. The employment of foodomic technologies in production monitoring will aid in refining the production pipeline by pinpointing and eradicating avenues of entry for pathogenic microorganisms, while simultaneously overseeing the growth of beneficial microorganisms. In contrast to traditional microbiological methodolo-

gies, outcomes will be expedited (independent of microbial incubation periods) and will furnish a more precise depiction of the scrutinized metagenome composition.

8. Glossary

- ASCII (American Standard Code for Information Interchange) is a character encoding system utilized in computers and communication devices to symbolize textual characters, with each character being allocated a distinct numerical value represented as an integer within the range of 0–127.
- FASTA is a file format employed for the storage of DNA, RNA, and protein sequences.
- Deep learning is a division of artificial intelligence (AI) that concentrates on the development and training of neural networks capable of learning and executing tasks automatically, without the need for explicit programming.
- HTML (Hypertext Markup Language) is a markup language utilized for the construction of websites, serving as the foundational language for structuring and presenting content on the web.
- HTS (high-throughput screening) involves high-throughput techniques for screening vast quantities of substances, leveraging automation and miniaturization to analyze numerous substances simultaneously. Various detection methods are employed, such as chemical reactions, absorbance, fluorescence, and bioluminescence, to identify the substances being tested.
- Kbp (kilobase pair) is a unit of measurement in molecular biology equivalent to 1000 nucleobase pairs.
- A k-mer is a nucleotide sequence of length k in DNA or RNA, comprising any of the four nucleotides: adenine, guanine, cytosine, and thymine in DNA or uracil in RNA.
- Contig refers to a continuous series of nucleotides within the genome, generated by amalgamating DNA sequence reads.
- MAG (metagenome-assembled genome) denotes a genome reconstructed from the combined genetic material present in a sample containing taxonomically diverse organisms from a specific environment.
- NGS (next-generation sequencing) encompasses advanced sequencing methodologies that facilitate rapid and simultaneous reading of multiple DNA fragments.
- OTU (operational taxonomic unit) is a taxonomic grouping for nucleotide sequences based on their sequence similarity.
- PHRED is a computational tool that evaluates the quality of DNA sequences acquired during sequencing, providing a probability estimation of errors in reading specific nucleotides. The resultant quality assessment, known as the PHRED score, is expressed as a numerical value on a logarithmic scale (0, 20, 40, 60), where higher values indicate greater accuracy in reading.
- Scaffolds denote extended sequences comprising ordered and linked contigs, representing a segment of the genome not assigned to a particular chromosome.
- TSV (tab-separated values) is a text file format where values are delimited by the tab character (TAB), facilitating the storage and transmission of data in tabular form.
- WMS (whole-metagenome sequencing) encompasses complete sequencing of the metagenome, enabling the analysis of all genetic material within a metagenomic sample.

Author Contributions: Conceptualization, J.S. and A.W.; writing—original draft preparation, J.S.; writing—review and editing, J.S., M.P.-B., A.S. and A.W.; visualization, J.S. and A.S.; and supervision, M.P.-B. and A.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cocolin, L.; Alessandria, V.; Dolci, P.; Gorra, R.; Rantsiou, K. Culture independent methods to assess the diversity and dynamics of microbiota during food fermentation. *Int. J. Food Microbiol.* **2013**, *167*, 29–43. [CrossRef]
2. Pogačić, T.; Kelava, N.; Zamberlin, Š.; Dolenčić-Špehar, I.; Samaržija, D. Methods for culture-independent identification of lactic acid bacteria in dairy products. *Food Technol. Biotechnol.* **2010**, *48*, 3–11.
3. Capozzi, F.; Bordoni, A. Foodomics: A new comprehensive approach to food and nutrition. *Genes Nutr.* **2013**, *8*, 1–4. [CrossRef] [PubMed]
4. Cifuentes, A. Food analysis and foodomics. *J. Chromatogr. A* **2009**, *1216*, 7109. [CrossRef] [PubMed]
5. Alhoshy, M.; Shehata, A.I.; Habib, Y.J.; Abdel-Latif, H.M.; Wang, Y.; Zhang, Z. Nutrigenomics in crustaceans: Current status and future prospects. *Fish Shellfish. Immunol.* **2022**, *129*, 1–12. [CrossRef] [PubMed]
6. Marcum, J.A. Nutrigenetics/nutrigenomics, personalized nutrition, and precision healthcare. *Curr. Nutr. Rep.* **2020**, *9*, 338–345. [CrossRef]
7. Ordovas, J.M.; Ferguson, L.R.; Tai, E.S.; Mathers, J.C. Personalised nutrition and health. *BMJ* **2018**, *361*, bmj.k2173. [CrossRef]
8. Dong, Z.; Chen, Y. Transcriptomics: Advances and approaches. *Sci. China Life Sci.* **2013**, *56*, 960–967. [CrossRef]
9. Allard, M.W.; Bell, R.; Ferreira, C.M.; Gonzalez-Escalona, N.; Hoffmann, M.; Muruvanda, T.; Ottesen, A.; Ramachandran, P.; Reed, E.; Sharma, S. Genomics of foodborne pathogens for microbial food safety. *Curr. Opin. Biotechnol.* **2018**, *49*, 224–229. [CrossRef]
10. Van Eck, N.; Waltman, L. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics* **2010**, *84*, 523–538. [CrossRef]
11. Gilbert, J.A.; Hughes, M. Gene expression profiling: Metatranscriptomics. In *High-Throughput Next Generation Sequencing: Methods and Applications*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 195–205.
12. Dollive, S.; Peterfreund, G.L.; Sherrill-Mix, S.; Bittinger, K.; Sinha, R.; Hoffmann, C.; Nabel, C.S.; Hill, D.A.; Artis, D.; Bachman, M.A. A tool kit for quantifying eukaryotic rRNA gene sequences from human microbiome samples. *Genome Biol.* **2012**, *13*, 1–13. [CrossRef]
13. Tickle, T.L.; Segata, N.; Waldron, L.; Weingart, U.; Huttenhower, C. Two-stage microbial community experimental design. *ISME J.* **2013**, *7*, 2330–2339. [CrossRef] [PubMed]
14. Staniszewski, A.; Kordowska-Wiater, M. Probiotic Yeasts and How to Find Them—Polish Wines of Spontaneous Fermentation as Source for Potentially Probiotic Yeasts. *Foods* **2023**, *12*, 3392. [CrossRef] [PubMed]
15. Qin, J.; Li, R.; Raes, J.; Arumugam, M.; Burgdorf, K.S.; Manichanh, C.; Nielsen, T.; Pons, N.; Levenez, F.; Yamada, T. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **2010**, *464*, 59–65. [CrossRef] [PubMed]
16. Zinger, L.; Gobet, A.; Pommier, T. Two decades of describing the unseen majority of aquatic microbial diversity. *Mol. Ecol.* **2012**, *21*, 1878–1896. [CrossRef] [PubMed]
17. Kable, M.E.; Srisengfa, Y.; Xue, Z.; Coates, L.C.; Marco, M.L. Viable and total bacterial populations undergo equipment-and time-dependent shifts during milk processing. *Appl. Environ. Microbiol.* **2019**, *85*, e00270-19. [CrossRef]
18. Barcenilla, C.; Cobo-Díaz, J.F.; De Filippis, F.; Valentino, V.; Cabrera Rubio, R.; O'Neil, D.; Mahler de Sanchez, L.; Armanini, F.; Carlino, N.; Blanco-Míguez, A. Improved sampling and DNA extraction procedures for microbiome analysis in food-processing environments. *Nat. Protoc.* **2024**, *19*, 1–20. [CrossRef] [PubMed]
19. Knight, R.; Jansson, J.; Field, D.; Fierer, N.; Desai, N.; Fuhrman, J.A.; Hugenholtz, P.; Van Der Lelie, D.; Meyer, F.; Stevens, R. Unlocking the potential of metagenomics through replicated experimental design. *Nat. Biotechnol.* **2012**, *30*, 513–520. [CrossRef]
20. Yilmaz, P.; Kottmann, R.; Field, D.; Knight, R.; Cole, J.R.; Amaral-Zettler, L.; Gilbert, J.A.; Karsch-Mizrachi, I.; Johnston, A.; Cochrane, G. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **2011**, *29*, 415–420. [CrossRef]
21. Salazar, J.K.; Carstens, C.K.; Ramachandran, P.; Shazer, A.G.; Narula, S.S.; Reed, E.; Ottesen, A.; Schill, K.M. Metagenomics of pasteurized and unpasteurized gouda cheese using targeted 16S rDNA sequencing. *BMC Microbiol.* **2018**, *18*, 1–13. [CrossRef]
22. Soyuçok, A.; Yurt, M.N.Z.; Altunbas, O.; Ozalp, V.C.; Sudagidan, M. Metagenomic and chemical analysis of Tarhana during traditional fermentation process. *Food Biosci.* **2021**, *39*, 100824. [CrossRef]
23. Mancini, A.; Rodriguez, M.C.; Zago, M.; Cologna, N.; Goss, A.; Carafa, I.; Tuohy, K.; Merz, A.; Franciosi, E. Massive survey on bacterial–bacteriophages biodiversity and quality of natural whey starter cultures in Trentingrana cheese production. *Front. Microbiol.* **2021**, *12*, 678012. [CrossRef] [PubMed]
24. Kaashyap, M.; Cohen, M.; Mantri, N. Microbial diversity and characteristics of kombucha as revealed by metagenomic and physicochemical analysis. *Nutrients* **2021**, *13*, 4446. [CrossRef] [PubMed]
25. Fabricio, M.F.; Mann, M.B.; Kothe, C.I.; Frazzon, J.; Tischer, B.; Flôres, S.H.; Ayub, M.A.Z. Effect of freeze-dried kombucha culture on microbial composition and assessment of metabolic dynamics during fermentation. *Food Microbiol.* **2022**, *101*, 103889. [CrossRef] [PubMed]

26. Treviso, R.L.; Sant'Anna, V.; Fabricio, M.F.; Ayub, M.A.Z.; Brandelli, A.; Hickert, L.R. Time and temperature influence on physicochemical, microbiological, and sensory profiles of yerba mate kombucha. *J. Food Sci. Technol.* **2024**, 1–10. [CrossRef]
27. González-Orozco, B.D.; García-Cano, I.; Escobar-Zepeda, A.; Jiménez-Flores, R.; Álvarez, V.B. Metagenomic analysis and antibacterial activity of kefir microorganisms. *J. Food Sci.* **2023**, *88*, 2933–2949. [CrossRef] [PubMed]
28. Nejati, F.; Capitain, C.C.; Krause, J.L.; Kang, G.-U.; Riedel, R.; Chang, H.-D.; Kurreck, J.; Junne, S.; Weller, P.; Neubauer, P. Traditional Grain-Based vs. Commercial Milk Kefirs, How Different Are They? *Appl. Sci.* **2022**, *12*, 3838. [CrossRef]
29. Qu, T.; Wang, P.; Zhao, X.; Liang, L.; Ge, Y.; Chen, Y. Metagenomics reveals differences in the composition of bacterial antimicrobial resistance and antibiotic resistance genes in pasteurized yogurt and probiotic bacteria yogurt from China. *J. Dairy Sci.* **2024**, *107*, 3451–3467. [CrossRef] [PubMed]
30. Luzzi, G.; Brinks, E.; Fritsche, J.; Franz, C.M. Microbial composition of sweetness-enhanced yoghurt during fermentation and storage. *AMB Express* **2020**, *10*, 1–7. [CrossRef]
31. Kim, E.; Cho, E.-J.; Yang, S.-M.; Kim, M.-J.; Kim, H.-Y. Novel approaches for the identification of microbial communities in kimchi: MALDI-TOF MS analysis and high-throughput sequencing. *Food Microbiol.* **2021**, *94*, 103641. [CrossRef]
32. Hwang, H.; Lee, H.J.; Lee, M.-A.; Sohn, H.; Chang, Y.H.; Han, S.G.; Jeong, J.Y.; Lee, S.H.; Hong, S.W. Selection and characterization of Staphylococcus hominis subsp. hominis WiKim0113 isolated from kimchi as a starter culture for the production of natural pre-converted nitrite. *Food Sci. Anim. Resour.* **2020**, *40*, 512. [CrossRef]
33. Jeong, C.-H.; Sohn, H.; Hwang, H.; Lee, H.-J.; Kim, T.-W.; Kim, D.-S.; Kim, C.-S.; Han, S.-G.; Hong, S.-W. Comparison of the probiotic potential between Lactiplantibacillus plantarum isolated from kimchi and standard probiotic strains isolated from different sources. *Foods* **2021**, *10*, 2125. [CrossRef] [PubMed]
34. Tlais, A.Z.A.; Lemos Junior, W.J.F.; Filannino, P.; Campanaro, S.; Gobbetti, M.; Di Cagno, R. How microbiome composition correlates with biochemical changes during sauerkraut fermentation: A focus on neglected bacterial players and functionalities. *Microbiol. Spectr.* **2022**, *10*, e00168-22. [CrossRef]
35. Zhang, J.; Song, H.S.; Zhang, C.; Kim, Y.B.; Roh, S.W.; Liu, D. Culture-independent analysis of the bacterial community in Chinese fermented vegetables and genomic analysis of lactic acid bacteria. *Arch. Microbiol.* **2021**, *203*, 4693–4703. [CrossRef]
36. Frank, J.A.; Pan, Y.; Tooming-Klunderud, A.; Eijsink, V.G.; McHardy, A.C.; Nederbragt, A.J.; Pope, P.B. Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Sci. Rep.* **2016**, *6*, 25373. [CrossRef] [PubMed]
37. Nicholls, S.M.; Quick, J.C.; Tang, S.; Loman, N.J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *Gigascience* **2019**, *8*, giz043. [CrossRef]
38. Ardui, S.; Ameur, A.; Vermeesch, J.R.; Hestand, M.S. Single molecule real-time (SMRT) sequencing comes of age: Applications and utilities for medical diagnostics. *Nucleic Acids Res.* **2018**, *46*, 2159–2168. [CrossRef] [PubMed]
39. Wick, R.R.; Judd, L.M.; Holt, K.E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol.* **2019**, *20*, 1–10. [CrossRef]
40. Kothe, C.I.; Mohellibi, N.; Renault, P. Revealing the microbial heritage of traditional Brazilian cheeses through metagenomics. *Food Res. Int.* **2022**, *157*, 111265. [CrossRef]
41. Suárez, N.; Weckx, S.; Minahk, C.; Hebert, E.M.; Saavedra, L. Metagenomics-based approach for studying and selecting bioprotective strains from the bacterial community of artisanal cheeses. *Int. J. Food Microbiol.* **2020**, *335*, 108894. [CrossRef]
42. Kothe, C.I.; Bolotin, A.; Kraïem, B.-F.; Dridi, B.; Team, F.M.; Renault, P. Unraveling the world of halophilic and halotolerant bacteria in cheese by combining cultural, genomic and metagenomic approaches. *Int. J. Food Microbiol.* **2021**, *358*, 109312. [CrossRef]
43. Bellassi, P.; Rocchetti, G.; Nocetti, M.; Lucini, L.; Masoero, F.; Morelli, L. A combined metabolomic and metagenomic approach to discriminate raw milk for the production of hard cheese. *Foods* **2021**, *10*, 109. [CrossRef] [PubMed]
44. Pradhan, S.; Prabhakar, M.R.; Karthika Parvathy, K.; Dey, B.; Jayaraman, S.; Behera, B.; Paramasivan, B. Metagenomic and physicochemical analysis of Kombucha beverage produced from tea waste. *J. Food Sci. Technol.* **2023**, *60*, 1088–1096. [CrossRef] [PubMed]
45. Góes-Neto, A.; Kukharenko, O.; Orlovska, I.; Podolich, O.; Imchen, M.; Kumavath, R.; Kato, R.B.; de Carvalho, D.S.; Tiwari, S.; Brenig, B. Shotgun metagenomic analysis of kombucha mutualistic community exposed to mars-like environment outside the international space station. *Environ. Microbiol.* **2021**, *23*, 3727–3742. [CrossRef] [PubMed]
46. Yang, J.; Lagishetty, V.; Kurnia, P.; Henning, S.M.; Ahdoot, A.I.; Jacobs, J.P. Microbial and chemical profiles of commercial kombucha products. *Nutrients* **2022**, *14*, 670. [CrossRef] [PubMed]
47. Landis, E.A.; Fogarty, E.; Edwards, J.C.; Popa, O.; Eren, A.M.; Wolfe, B.E. Microbial diversity and interaction specificity in kombucha tea fermentations. *mSystems* **2022**, *7*, e00157-22. [CrossRef]
48. Liu, S.; Lu, S.-Y.; Qureshi, N.; Enshasy, H.A.E.; Skory, C.D. Antibacterial property and metagenomic analysis of milk kefir. *Probiotics Antimicrob. Proteins* **2022**, *14*, 1170–1183. [CrossRef] [PubMed]
49. Biçer, Y.; Telli, A.E.; Sönmez, G.; Turkal, G.; Telli, N.; Uçar, G. Comparison of commercial and traditional kefir microbiota using metagenomic analysis. *Int. J. Dairy Technol.* **2021**, *74*, 528–534. [CrossRef]
50. Walsh, L.H.; Coakley, M.; Walsh, A.M.; Crispie, F.; O'Toole, P.W.; Cotter, P.D. Analysis of the milk kefir pan-metagenome reveals four community types, core species, and associated metabolic pathways. *Iscience* **2023**, *26*, 108004. [CrossRef]

51. Aydin, S.; Erözden, A.A.; Tavşanlı, N.; Müdüroğlu, A.; Çalışkan, M.; Kara, İ. Anthocyanin Addition to Kefir: Metagenomic Analysis of Microbial Community Structure. *Curr. Microbiol.* **2022**, *79*, 327. [CrossRef]
52. Qiu, S.; Zeng, H.; Yang, Z.; Hung, W.L.; Wang, B.; Yang, A. Dynamic metagenome-scale metabolic modeling of a yogurt bacterial community. *Biotechnol. Bioeng.* **2023**, *120*, 2186–2198. [CrossRef] [PubMed]
53. Suh, S.H.; Kim, M.K. Microbial communities related to sensory characteristics of commercial drinkable yogurt products in Korea. *Innov. Food Sci. Emerg. Technol.* **2021**, *67*, 102565. [CrossRef]
54. Samelis, J.; Doulgeraki, A.I.; Bikouli, V.; Pappas, D.; Kakouri, A. Microbiological and metagenomic characterization of a retail delicatessen Galotyri-like fresh acid-curd cheese product. *Fermentation* **2021**, *7*, 67. [CrossRef]
55. Le Roy, C.I.; Kurilshikov, A.; Leeming, E.R.; Visconti, A.; Bowyer, R.C.; Menni, C.; Falchi, M.; Koutnikova, H.; Veiga, P.; Zhernakova, A. Yoghurt consumption is associated with changes in the composition of the human gut microbiome and metabolome. *BMC Microbiol.* **2022**, *22*, 39.
56. Oh, Y.-J.; Park, Y.-R.; Hong, J.; Lee, D.-Y. Metagenomic, Metabolomic, and Functional Evaluation of Kimchi Broth Treated with Light-Emitting Diodes (LEDs). *Metabolites* **2021**, *11*, 472. [CrossRef] [PubMed]
57. Park, D.H. Effects of carbon dioxide on metabolite production and bacterial communities during kimchi fermentation. *Biosci. Biotechnol. Biochem.* **2018**, *82*, 1234–1242. [CrossRef]
58. Gaudioso, G.; Weil, T.; Marzorati, G.; Solovyev, P.; Bontempo, L.; Franciosi, E.; Bertoldi, L.; Pedrolli, C.; Tuohy, K.M.; Fava, F. Microbial and metabolic characterization of organic artisanal sauerkraut fermentation and study of gut health-promoting properties of sauerkraut brine. *Front. Microbiol.* **2022**, *13*, 929738. [CrossRef] [PubMed]
59. Huang, W.; Peng, H.; Chen, J.; Yan, X.; Zhang, Y. Bacterial diversity analysis of Chaozhou Sauerkraut based on high-throughput sequencing of different production methods. *Fermentation* **2023**, *9*, 282. [CrossRef]
60. Zhang, S.; Zhang, Y.; Wu, L.; Zhang, L.; Wang, S. Characterization of microbiota of naturally fermented sauerkraut by high-throughput sequencing. *Food Sci. Biotechnol.* **2023**, *32*, 855–862. [CrossRef]
61. Thriene, K.; Hansen, S.S.; Binder, N.; Michels, K.B. Effects of fermented vegetable consumption on human gut microbiome diversity—A pilot study. *Fermentation* **2022**, *8*, 118. [CrossRef]
62. Falgueras, J.; Lara, A.J.; Fernández-Pozo, N.; Cantón, F.R.; Pérez-Trabado, G.; Claros, M.G. SeqTrim: A high-throughput pipeline for pre-processing any type of sequence read. *BMC Bioinform.* **2010**, *11*, 1–12. [CrossRef]
63. Aronesty, E. *Ea-Utils: Command-Line Tools for Processing Biological Sequencing Data*; Expression Analysis: Durham, NC, USA, 2011.
64. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
65. Xu, Z.; Mai, Y.; Liu, D.; He, W.; Lin, X.; Xu, C.; Zhang, L.; Meng, X.; Mafofo, J.; Zaher, W.A. Fast-bonito: A faster deep learning based basecaller for nanopore sequencing. *Artif. Intell. Life Sci.* **2021**, *1*, 100011. [CrossRef]
66. Zeng, J.; Cai, H.; Peng, H.; Wang, H.; Zhang, Y.; Akutsu, T. Causalcall: Nanopore basecalling using a temporal convolutional network. *Front. Genet.* **2020**, *10*, 1332. [CrossRef] [PubMed]
67. Leggett, R.M.; Heavens, D.; Caccamo, M.; Clark, M.D.; Davey, R.P. NanoOK: Multi-reference alignment analysis of nanopore sequencing data, quality and error profiles. *Bioinformatics* **2016**, *32*, 142–144. [CrossRef] [PubMed]
68. Simpson, J.T.; Pop, M. The theory and practice of genome sequence assembly. *Annu. Rev. Genom. Hum. Genet.* **2015**, *16*, 153–172. [CrossRef] [PubMed]
69. Schwartz, D.C.; Waterman, M.S. New generations: Sequencing machines and their computational challenges. *J. Comput. Sci. Technol.* **2010**, *25*, 3. [CrossRef]
70. Zaheer, R.; Noyes, N.; Ortega Polo, R.; Cook, S.R.; Marinier, E.; Van Domselaar, G.; Belk, K.E.; Morley, P.S.; McAllister, T.A. Impact of sequencing depth on the characterization of the microbiome and resistome. *Sci. Rep.* **2018**, *8*, 1–11. [CrossRef]
71. Howe, A.; Chain, P.S. Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Front. Microbiol.* **2015**, *6*, 678. [CrossRef]
72. Peng, Y.; Leung, H.C.; Yiu, S.-M.; Chin, F.Y. Meta-IDBA: A de Novo assembler for metagenomic data. *Bioinformatics* **2011**, *27*, i94–i101. [CrossRef]
73. Mapleson, D.; Drou, N.; Swarbreck, D. RAMPART: A workflow management system for de novo genome assembly. *Bioinformatics* **2015**, *31*, 1824–1826. [CrossRef] [PubMed]
74. Peng, Y.; Leung, H.C.; Yiu, S.-M.; Chin, F.Y. IDBA-UD: A de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **2012**, *28*, 1420–1428. [CrossRef]
75. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
76. Li, D.; Liu, C.-M.; Luo, R.; Sadakane, K.; Lam, T.-W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **2015**, *31*, 1674–1676. [CrossRef]
77. Boisvert, S.; Raymond, F.; Godzaridis, É.; Laviolette, F.; Corbeil, J. Ray Meta: Scalable de novo metagenome assembly and profiling. *Genome Biol.* **2012**, *13*, 1–13. [CrossRef] [PubMed]
78. Sato, K.; Sakakibara, Y. An extended genovo metagenomic assembler by incorporating paired-end information. *PeerJ* **2013**, *1*, e196.
79. Kim, M.; Zhang, X.; Ligo, J.G.; Farnoud, F.; Veeravalli, V.V.; Milenkovic, O. MetaCRAM: An integrated pipeline for metagenomic taxonomy identification and compression. *BMC Bioinform.* **2016**, *17*, 1–13. [CrossRef] [PubMed]

80. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, 1–12. [CrossRef]
81. Yang, X.; Charlebois, P.; Gnerre, S.; Coole, M.G.; Lennon, N.J.; Levin, J.Z.; Qu, J.; Ryan, E.M.; Zody, M.C.; Henn, M.R. De novo assembly of highly diverse viral populations. *BMC Genom.* **2012**, *13*, 1–13. [CrossRef]
82. Brady, A.; Salzberg, S. PhymmBL expanded: Confidence scores, custom databases, parallelization and more. *Nat. Methods* **2011**, *8*, 367. [CrossRef]
83. Cock, P.J.; Chilton, J.M.; Grüning, B.; Johnson, J.E.; Soranzo, N. NCBI BLAST+ integrated into Galaxy. *Gigascience* **2015**, *4*, s13742-015. [CrossRef]
84. Gregor, I.; Dröge, J.; Schirmer, M.; Quince, C.; McHardy, A.C. PhyloPythiaS+: A self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ* **2016**, *4*, e1603. [CrossRef]
85. Haider, B.; Ahn, T.-H.; Bushnell, B.; Chai, J.; Copeland, A.; Pan, C. Omega: An overlap-graph de novo assembler for metagenomics. *Bioinformatics* **2014**, *30*, 2717–2722. [CrossRef]
86. Ruby, J.G.; Bellare, P.; DeRisi, J.L. PRICE: Software for the targeted assembly of components of (Meta) genomic sequence data. *G3 Genes Genomes Genet.* **2013**, *3*, 865–880. [CrossRef] [PubMed]
87. Jiang, Y.; Wang, J.; Xia, D.; Yu, G. EnSVMB: Metagenomics fragments classification using ensemble SVM and BLAST. *Sci. Rep.* **2017**, *7*, 9440. [CrossRef] [PubMed]
88. Myers, E.W.; Sutton, G.G.; Delcher, A.L.; Dew, I.M.; Fasulo, D.P.; Flanigan, M.J.; Kravitz, S.A.; Mobarry, C.M.; Reinert, K.H.; Remington, K.A. A whole-genome assembly of Drosophila. *Science* **2000**, *287*, 2196–2204. [CrossRef] [PubMed]
89. Wang, Z.; Huang, P.; You, R.; Sun, F.; Zhu, S. MetaBinner: A high-performance and stand-alone ensemble binning method to recover individual genomes from complex microbial communities. *Genome Biol.* **2023**, *24*, 1. [CrossRef]
90. Kelley, D.R.; Salzberg, S.L. Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinform.* **2010**, *11*, 1–12. [CrossRef]
91. Strous, M.; Kraft, B.; Bisdorf, R.; Tegetmeyer, H.E. The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front. Microbiol.* **2012**, *3*, 410. [CrossRef]
92. Kang, D.D.; Li, F.; Kirton, E.; Thomas, A.; Egan, R.; An, H.; Wang, Z. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **2019**, *7*, e7359. [CrossRef]
93. Imelfort, M.; Parks, D.; Woodcroft, B.J.; Dennis, P.; Hugenholtz, P.; Tyson, G.W. GroopM: An automated tool for the recovery of population genomes from related metagenomes. *PeerJ* **2014**, *2*, e603. [CrossRef]
94. Alneberg, J.; Bjarnason, B.S.; de Bruijn, I.; Schirmer, M.; Quick, J.; Ijaz, U.Z.; Loman, N.J.; Andersson, A.F.; Quince, C. CONCOCT: Clustering contigs on coverage and composition. *arXiv* **2013**, arXiv:1312.4038.
95. Albertsen, M.; Hugenholtz, P.; Skarshewski, A.; Nielsen, K.L.; Tyson, G.W.; Nielsen, P.H. Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nat. Biotechnol.* **2013**, *31*, 533–538. [CrossRef] [PubMed]
96. Jiang, Z.; Li, X.; Guo, L. Binning Metagenomic Contigs Using Unsupervised Clustering and Reference Databases. In *Interdisciplinary Sciences: Computational Life Sciences*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 14, pp. 795–803.
97. Parks, D.H.; Imelfort, M.; Skennerton, C.T.; Hugenholtz, P.; Tyson, G.W. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **2015**, *25*, 1043–1055. [CrossRef] [PubMed]
98. Tettelin, H.; Riley, D.; Cattuto, C.; Medini, D. Comparative genomics: The bacterial pan-genome. *Curr. Opin. Microbiol.* **2008**, *11*, 472–477. [CrossRef] [PubMed]
99. Mosquera-Rendón, J.; Rada-Bravo, A.M.; Cárdenas-Brito, S.; Corredor, M.; Restrepo-Pineda, E.; Benítez-Páez, A. Pangenome-wide and molecular evolution analyses of the Pseudomonas aeruginosa species. *BMC Genom.* **2016**, *17*, 1–14. [CrossRef] [PubMed]
100. Lapierre, P.; Gogarten, J.P. Estimating the size of the bacterial pan-genome. *Trends Genet.* **2009**, *25*, 107–110. [CrossRef] [PubMed]
101. Jordan, I.K.; Makarova, K.S.; Spouge, J.L.; Wolf, Y.I.; Koonin, E.V. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* **2001**, *11*, 555–565. [CrossRef] [PubMed]
102. Overbeek, R.; Olson, R.; Pusch, G.D.; Olsen, G.J.; Davis, J.J.; Disz, T.; Edwards, R.A.; Gerdes, S.; Parrello, B.; Shukla, M. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Res.* **2014**, *42*, D206–D214. [CrossRef]
103. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef]
104. Tabari, E.; Su, Z. PorthoMCL: Parallel orthology prediction using MCL for the realm of massive genome availability. *Big Data Anal.* **2017**, *2*, 1–5. [CrossRef] [PubMed]
105. Emms, D.M.; Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **2019**, *20*, 1–14. [CrossRef] [PubMed]
106. Chaudhari, N.M.; Gupta, V.K.; Dutta, C. BPGA-an ultra-fast pan-genome analysis pipeline. *Sci. Rep.* **2016**, *6*, 24373. [CrossRef] [PubMed]
107. Yu, J.; Blom, J.; Glaeser, S.; Jaenicke, S.; Juhre, T.; Rupp, O.; Schwengers, O.; Spänig, S.; Goesmann, A. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. *J. Biotechnol.* **2017**, *261*, 2–9. [CrossRef] [PubMed]

108. Contreras-Moreira, B.; Vinuesa, P. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Appl. Environ. Microbiol.* **2013**, *79*, 7696–7701. [CrossRef] [PubMed]
109. Pantoja, Y.; Pinheiro, K.; Veras, A.; Araújo, F.; Lopes de Sousa, A.; Guimarães, L.C.; Silva, A.; Ramos, R.T. PanWeb: A web interface for pan-genomic analysis. *PLoS ONE* **2017**, *12*, e0178154. [CrossRef] [PubMed]
110. Zhao, Y.; Wu, J.; Yang, J.; Sun, S.; Xiao, J.; Yu, J. PGAP: Pan-genomes analysis pipeline. *Bioinformatics* **2012**, *28*, 416–418. [CrossRef] [PubMed]
111. Page, A.J.; Cummins, C.A.; Hunt, M.; Wong, V.K.; Reuter, S.; Holden, M.T.; Fookes, M.; Falush, D.; Keane, J.A.; Parkhill, J. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **2015**, *31*, 3691–3693. [CrossRef]
112. Calle, M.L. Statistical analysis of metagenomics data. *Genom. Inform.* **2019**, *17*, e6. [CrossRef]
113. Segata, N.; Börnigen, D.; Morgan, X.C.; Huttenhower, C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nat. Commun.* **2013**, *4*, 2304. [CrossRef]
114. Darling, A.E.; Jospin, G.; Lowe, E.; Matsen IV, F.A.; Bik, H.M.; Eisen, J.A. PhyloSift: Phylogenetic analysis of genomes and metagenomes. *PeerJ* **2014**, *2*, e243. [CrossRef]
115. Wu, Y.-W. ezTree: An automated pipeline for identifying phylogenetic marker genes and inferring evolutionary relationships among uncultivated prokaryotic draft genomes. *BMC Genom.* **2018**, *19*, 7–16. [CrossRef] [PubMed]
116. Lee, M.D. GToTree: A user-friendly workflow for phylogenomics. *Bioinformatics* **2019**, *35*, 4162–4164. [CrossRef] [PubMed]
117. Wu, M.; Eisen, J.A. A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.* **2008**, *9*, 1–11. [CrossRef]
118. Marçais, G.; Delcher, A.L.; Phillippy, A.M.; Coston, R.; Salzberg, S.L.; Zimin, A. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **2018**, *14*, e1005944. [CrossRef]
119. Edgar, R.C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797. [CrossRef]
120. Xie, R.; Zan, X.; Chu, L.; Su, Y.; Xu, P.; Liu, W. Study of the error correction capability of multiple sequence alignment algorithm (MAFFT) in DNA storage. *BMC Bioinform.* **2023**, *24*, 111. [CrossRef] [PubMed]
121. Garriga, E.; Di Tommaso, P.; Magis, C.; Erb, I.; Mansouri, L.; Baltzis, A.; Floden, E.; Notredame, C. Multiple sequence alignment computation using the t-coffee regressive algorithm implementation. In *Multiple Sequence Alignment: Methods and Protocols*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 89–97.
122. Wheeler, T.J.; Kececioglu, J.D. Multiple alignment by aligning alignments. *Bioinformatics* **2007**, *23*, i559–i568. [CrossRef] [PubMed]
123. Mirarab, S.; Nguyen, N.; Warnow, T. PASTA: Ultra-large multiple sequence alignment. In Proceedings of the Research in Computational Molecular Biology: 18th Annual International Conference, RECOMB 2014, Pittsburgh, PA, USA, 2–5 April 2014; Proceedings 18; pp. 177–191.
124. Nguyen, N.-P.; Mirarab, S.; Kumar, K.; Warnow, T. Ultra-large alignments using ensembles of hidden Markov models. In Proceedings of the Research in Computational Molecular Biology: 19th Annual International Conference, RECOMB 2015, Warsaw, Poland, 12–15 April 2015; Proceedings 19; pp. 259–260.
125. Liu, K.; Linder, C.R.; Warnow, T. RAxML and FastTree: Comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS ONE* **2011**, *6*, e27731. [CrossRef]
126. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **2009**, *26*, 1641–1650. [CrossRef]
127. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef]
128. Kozlov, A.M.; Darriba, D.; Flouri, T.; Morel, B.; Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **2019**, *35*, 4453–4455. [CrossRef] [PubMed]
129. Mirarab, S.; Warnow, T. ASTRAL-II: Coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* **2015**, *31*, i44–i52. [CrossRef] [PubMed]
130. Vachaspati, P.; Warnow, T. ASTRID: Accurate species trees from internode distances. *BMC Genom.* **2015**, *16*, 1–13. [CrossRef] [PubMed]
131. Nguyen, L.-T.; Schmidt, H.A.; Von Haeseler, A.; Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, *32*, 268–274. [CrossRef]
132. Kapli, P.; Yang, Z.; Telford, M.J. Phylogenetic tree building in the genomic age. *Nat. Rev. Genet.* **2020**, *21*, 428–444. [CrossRef]
133. Silva, M.; Machado, M.P.; Silva, D.N.; Rossi, M.; Moran-Gilad, J.; Santos, S.; Ramirez, M.; Carrico, J.A. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb. Genom.* **2018**, *4*, e000166. [CrossRef]
134. Consortium, U. UniProt: A hub for protein information. *Nucleic Acids Res.* **2015**, *43*, D204–D212. [CrossRef] [PubMed]
135. Kanehisa, M.; Sato, Y.; Morishima, K. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J. Mol. Biol.* **2016**, *428*, 726–731. [CrossRef]
136. Gevers, D.; Knight, R.; Petrosino, J.F.; Huang, K.; McGuire, A.L.; Birren, B.W.; Nelson, K.E.; White, O.; Methé, B.A.; Huttenhower, C. The Human Microbiome Project: A community resource for the healthy human microbiome. *PLoS Biol.* **2012**, *10*, e1001377. [CrossRef]
137. Barbera, P.; Kozlov, A.M.; Czech, L.; Morel, B.; Darriba, D.; Flouri, T.; Stamatakis, A. EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Syst. Biol.* **2019**, *68*, 365–369. [CrossRef] [PubMed]

138. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef] [PubMed]
139. Bolyen, E.; Rideout, J.R.; Dillon, M.R.; Bokulich, N.A.; Abnet, C.C.; Al-Ghalith, G.A.; Alexander, H.; Alm, E.J.; Arumugam, M.; Asnicar, F. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **2019**, *37*, 852–857. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.