genes

MDPI

Opinion

Whole Genome Sequencing Applied to Pathogen Source Tracking in Food Industry: Key Considerations for Robust Bioinformatics Data Analysis and Reliable Results Interpretation

Caroline Barretto *, Cristian Rincón, Anne-Catherine Portmann and Catherine Ngom-Bru

Institute of Food Safety and Analytical Sciences, Nestlé Research, 1000 Lausanne 26, Switzerland; cristian.rincon@rd.nestle.com (C.R.); annecatherine.portmann@gmail.com (A.-C.P.); catherine.ngombru@rdls.nestle.com (C.N.-B.)

* Correspondence: caroline.barretto@rdls.nestle.com

Abstract: Whole genome sequencing (WGS) has arisen as a powerful tool to perform pathogen source tracking in the food industry thanks to several developments in recent years. However, the cost associated to this technology and the degree of expertise required to accurately process and understand the data has limited its adoption at a wider scale. Additionally, the time needed to obtain actionable information is often seen as an impairment for the application and use of the information generated via WGS. Ongoing work towards standardization of wet lab including sequencing protocols, following guidelines from the regulatory authorities and international standardization efforts make the technology more and more accessible. However, data analysis and results interpretation guidelines are still subject to initiatives coming from distinct groups and institutions. There are multiple bioinformatics software and pipelines developed to handle such information. Nevertheless, little consensus exists on a standard way to process the data and interpret the results. Here, we want to present the constraints we face in an industrial setting and the steps we consider necessary to obtain high quality data, reproducible results and a robust interpretation of the obtained information. All of this, in a time frame allowing for data-driven actions supporting factories and their needs.

Keywords: whole genome sequencing; food industry; bioinformatics; workflow; data analysis; metadata; food safety; data quality

Citation: Barretto, C.; Rincón, C.; Portmann, A.-C.; Ngom-Bru, C. Whole Genome Sequencing Applied to Pathogen Source Tracking in Food Industry: Key Considerations for Robust Bioinformatics Data Analysis and Reliable Results Interpretation. *Genes* **2021**, *12*, 275. https://doi.org/10.3390/genes12020275

Academic Editor: Kevin Vanneste

Received: 18 January 2021
Accepted: 8 February 2021
Published: 15 February 2021

1. Introduction

Advances in DNA sequencing technologies have transformed the capacity to investigate the dynamics of foodborne pathogens inhabiting diverse environments. Food industry [1,2] and food safety agencies [3–5] have benefited from these improvements for pathogen source tracking. Whole genome sequencing (WGS) has arisen as a tool with a scope that goes beyond academic research proving to be an asset in ensuring food safety. The technologies have gone through considerable improvements in recent years making them both faster and cheaper, allowing thus for relatively wide adoption and use for food safety and public health on a routine basis [6]. Multiple governmental entities and regulatory authorities have established networks coordinating the efforts of nationwide laboratories to use WGS for foodborne pathogen source tracking [3,7,8] and real-time outbreak detection and investigation [9,10].

Nowadays, WGS is widely considered as the approach offering the highest resolution and precision for routine real-time surveillance concerning foodborne pathogens and has been adopted by multiple regulatory agencies worldwide [6,11–13]. This has been accompanied by the development of several computational software and methods allowing the analysis of these data. Nevertheless, a wider adoption of the technology, particularly in food industry, has been lessened by several factors [14,15]. First, the computational

infrastructure required to process (and store) the data. Moreover, little consensus exists concerning the use of the multitude of software available to process sequencing data, and although general guidelines are being consolidated, to date, multiple institutions use different internally validated workflows [16,17]. Despite advancements in recent years, the speed in which useful information can be provided to the factories has also limited the adoption of WGS for foodborne pathogen surveillance in industry. Timewise, collecting the samples, sending them for identification and sequencing is a major limiting factor. Additionally, the defined workflow for genome analysis should produce interpretable results within a time frame that does not limit their operational applicability. Finally, the added value of WGS over traditionally used typing methods relies on an accurate interpretation of the results. An analysis conducted using WGS is a powerful tool for determining the relatedness of bacterial isolates in pathogen source tracking. However, by itself, it can only indicate that isolates recently arose from the same source. Linking the information obtained via WGS with the likely origin of the foodborne contaminant requires contextualization of the results using information about the sample (metadata), thus requiring appropriate expertise. Detection of a pathogen or its relatedness to others, might not be enough to identify the root cause of the contamination. WGS has the power to provide information exploitable by quality managers that can be used to develop data-driven strategies for food safety management.

Over the years, guidelines for the standardization of wet lab protocols have been issued, however, for data processing, there is still a large gap to close regarding standardization to ensure the analysis is complete and correct, for ultimately making an accurate recommendation. Deep understanding of the results is key in order to support factories in their root cause analysis investigations. Software will always provide a result, however without critical review and appropriate verifications, these results can be erroneous and lead to incorrect interpretations. All in all, the integration of microbiology, genomics and bioinformatics knowledge is essential to ensure the quality of the data, validate the analytical results and provide a reliable interpretation of the obtained information, integrating metadata. Several bioinformatics approaches have been tested and validated [18–20] and here, more than introducing a workflow, we present what we consider important as key considerations for robust bioinformatics data analysis and reliable results interpretation in the context of whole genome sequencing applied to pathogen source tracking in food industry.

2. Materials and Methods

Workflow

The data analysis workflow was split into stages and, for each of them, quality control metrics were defined to ensure high quality of the obtained data. Based on selection criteria, several open-source software were identified for each step, tested and benchmarked.

The considered criteria for software selection were to be open-source and well documented, Linux-based, actively maintained, locally installable (i.e., not running in external or cloud servers), compliant with internally defined IT security regulations, adapted to be launched on large datasets and fast enough to allow the whole workflow to be run within 24 h on a dedicated server. Importantly, these tools can be used for a large range of pathogens. Here we present our experiences notably with *Listeria monocytogenes* and *Salmonella enterica*.

For each stage, the best performing software (for our specific case), fulfilling the above criteria, was streamlined into a workflow. This selection was based on accuracy and reproducibility of the results during the benchmarking. For each stage, several metrics from the output data were considered as quality checks. When an isolate failed one of these quality checks, it was tagged as of low quality to be further examined. As bioinformatics software are rapidly changing and developing, the intention of this work was not to promote a certain software or benchmark its performance but rather highlight the criteria

to be considered to ensure high quality data for downstream analyses. Table 1 summarizes the stages and metrics considered.

Table 1. List of main stages, benchmarked software and parameters considered to ensure high-quality data.

Stage	Evaluated Software	Parameters Considered to Ensure High-Quality Data	Examples of QC Evaluation
Raw reads quality control (QC)	FASTQC [21]	- Per base sequence quality - GC content - Average genome coverage	GC content deviating from expected indicates a possible contamination or sample mislabeling.
Isolate identification	SalmID [22] Sixess [23] KmerID [24] Kraken [25]	- Predicted genus and species of the isolate - Percentage of reads attributed to correct species - Percentage of unclassified reads	A relatively high number of unclassified reads has been associated with plasmid/phage presence, or with a contamination.
Read quality trim and removal	Trimmomatic [26]	- Number of discarded reads - Read length distribution after trimming	A large number of discarded reads was related to poor quality of the sequencing run.
Genome assembly	SPAdes [27] Skesa [28]	Assessed with QUAST [29] - Number of contigs - Genome size - N50	High number of contigs, or deviation of the expected genome size is an indication of low sequenced genome quality [30].
Sequence typing	mlst [31] MLST-CGE [32] stringMLST [33]	- 7 genes MLST composition	Lack of predicted MLST points to low assembly quality.
Salmonella serovar prediction *Bacillus* clade prediction	Sistr [34] SeqSero2 [35] BTyper [36]	Serovar/clade prediction [37]	Lack of predicted serovar points to low assembly quality.
First grouping (cg/wgMLST)	chewBBACA [38]	Number of uncalled loci in the genome	A high number of loci from the profile not found in the genome indicates low assembly quality, contamination, or misidentification of the species.
SNP calling	CFSAN SNP pipeline [16]	- Percentage of reads mapped to the reference - Number of SNP missing positions - Differences between raw and preserved matrices.	A large number of missing positions suggest an inappropriate choice of the reference.
Mobile Genetic Elements (MGE) identification such as phages or plasmids	Phigaro [39] ProphET [40] MOB-Suite [41]	- MGE type - MGE position in the genome	Eventually, SNP analyses are run after MGE removal/masking to confirm relatedness.

We also added, among others, data integrity verification with the md5sum software when data are transferred between servers and cross-referencing isolate identification with

metadata. Importantly, the criteria mentioned here are fit-for-purpose for Illumina data. Long read data (Pacific Biosciences or Oxford Nanopore) will require adjustments.

3. Metadata

Metadata, understood as the collection of information related to each sample and its processing, is key for results interpretation. These data (factory name, internal isolate number, sampling time, material type, country, physical description of the sample, sampling point, isolate taxonomy, DNA extraction kit, sequencing chemistry, sequencing platform, operator, etc.) need to be managed and curated to ensure its quality and utility for results interpretation. Thus, data quality checks and data management practices apply to metadata collection and storage. This kind of information is often handled via a laboratory information management system (LIMS). A LIMS frequently includes sample tracking, stock management and data exchange interfaces. As for any other database, metadata management requires resources, quality monitoring, data backup and recovery and security compliance. The more data is acquired, the more accurate the data interpretation will be. Different stakeholders in the food industry have a key role to collect and manage this metadata from the person who routinely takes the swabs/samples, to the laboratory that does the initial diagnostic and then to the WGS lab. Therefore, controls are needed to operate the flow of data and the data management should be considered as an integral and key component of any WGS analysis workflow.

4. Results

A stepwise approach is recommended to fully control the analysis and can be described in key components (Figure 1).

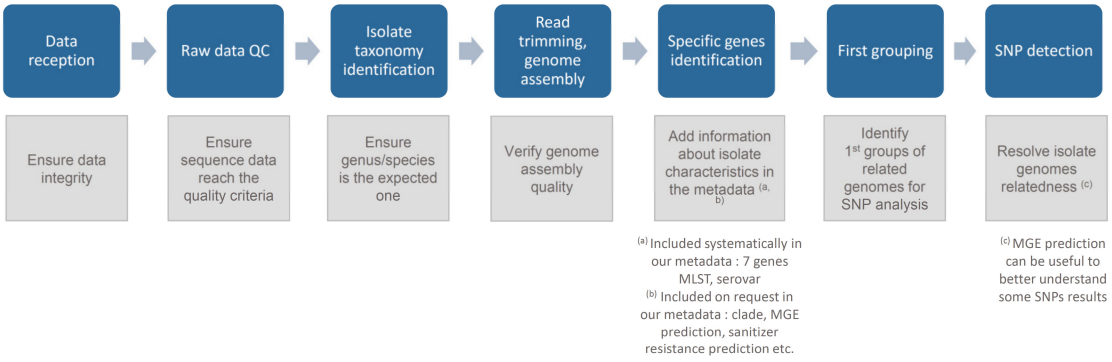

Figure 1. Key components of our pathogen source tracking whole genome sequencing (WGS) workflow: Overview and main goal of each step.

The software included in our workflow are listed in Table 2. The rationale for the choice of each software implemented in our workflow was a combination of: (i) the possibility of the software to produce in their output the parameters that are used as quality control (QC) evaluation (mentioned in Table 1); (ii) features mentioned in material and methods, and performance such as accuracy and precision.

Several of these conditions may be less relevant in another setting (e.g., cloud computing is a valid option when there is no sensitive data). Thus, the choice of tools is context specific.

Initial data integrity verification is important since data corruption may occur during transfer from the sequencing server to the analysis server, due to, for example, network interruptions. Once a sequencing run is ready to be analyzed, key quality verifications are necessary to make sure the information of each sample can be used appropriately.

The quality verification of the fastq files is done to ensure that sequencing reads have the appropriate size and nucleotide quality for each position. At this stage, the estimated genome average coverage and the GC content are also evaluated. Having

an appropriate genome average coverage ensures not only accurate Single Nucleotide Polymorphism (SNP)/allele calling, but also high-quality genome assemblies. A GC content not concordant with the expected genus can also be a good indicator of a possible contamination and should be examined. Including a negative control, meaning a sample that is supposed to be DNA free, could be a good practice in the sequencing laboratory. A high number of reads in the negative sample indicates contamination of the lab supplies with foreign DNA and therefore a contamination of the sequenced samples. Adding a negative control in the sequencing run would have no impact on the sequencing costs and should not preempt other isolates sequencing coverage since this sample is supposed to be DNA free. Similarly, a positive control should be included whenever possible since it can help for troubleshooting a failed run. For example, a suboptimal result in terms of reads throughput (including the positive control) points to issues with the library preparation or the sequencing chemistry quality. This can make a difference on the ability to maintain turnaround time of the sequencing and reduce consumables used in troubleshooting.

Table 2. List of currently implemented software in our workflow.

Stage	Software
Data integrity	md5sum
Raw reads quality control (QC)	FastQC
Isolate identification	Kraken
Read quality trim and removal	Trimmomatic
Genome assembly	SKESA
Sequence typing	mlst
Bacillus clade prediction	btyper
Salmonella serovar prediction	SeqSero2
First grouping (cg/wgMLST)	chewBBACA
SNP calling	cfsan_snp_pipeline
Mobile Genetic Elements (MGE) identification such as phages or plasmids	MOB-Suite Phigaro ProphET

After assembling the sequencing reads into a genome, a low number of contigs is one of the indications of high quality of the assembly. The overall genome size should enter within the expected size for the predicted genus [30]. Furthermore, for better reproducibility, a software producing an identical assembly for the same input when ran multiple times is advantageous [28]. As a laboratory often handles multiple sample types, it is not exceptional that a sample mix or a contamination occurs when extracting DNA or preparing sequencing libraries. It is thus essential to accurately assign a taxonomy to the sequenced isolate and to have an indication of potential contamination from other samples.

The additional information that can be obtained from whole genomes such as the serovar prediction for *Salmonella*, the seven genes Multi-Locus Sequence Type (MLST) prediction for a large number of organisms, mobile genetic element detection such as plasmids or phages, is highly valuable for a reliable data interpretation.

Before starting SNP analysis, a key step is to make sure that only closely related genomes are included in the analysis. The variant calling for SNP analyses relies in comparisons to a reference genome and the choice and quality of this reference impacts the obtained information [42]. Comparing genetically distant genomes, implies having portions of the genomes with low read mapping and potentially generating SNP hotspots (several SNPs in close proximity). As both, low mapping and SNP dense regions, are generally excluded from the analyses, SNPs at those positions are excluded too, leading to artificially low differences and to a false interpretation of the results. At this point, having the possibility to quickly compare large datasets in a short time performing a first grouping with, for instance, core genome/whole genome MLST (cg/wgMLST), enables to focus on groups of isolates that are related to then be further separated by SNP analyses. If during the allele call, a locus is not found in a genome, it is zeroed and removed from the analysis.

When this occurs for a large number of loci, similarity among the compared strains will be artefactually increased. Hence, it is important to also have a critical eye on the allele call quality and to further analyze the groups of related genomes with a SNP approach.

Even after implementing quality checks, a workflow may need several rounds of analyses to be completed. In some cases, SNPs hotspots are identified and frequently these SNPs are present on mobile genetic elements (MGE), such as phages and plasmids. As these variants are not phylogenetically relevant [6,43], they are often filtered out in the final analyses. An MGE identifier pipeline should be coupled with SNP analyses to verify whether these hotspots are due to the presence of MGEs and not because of low quality of the reference genome assembly or low mapping of reads versus the reference genome.

Finally, a crucial step for results interpretation is to contextualize the findings using the metadata. Its utility greatly depends on the implementation of a controlled vocabulary to ensure consistency among contributors to the database (e.g., factory, laboratory and sequencing facility). During the process of standardization, it is key to consider the requirements of bioinformaticians (or whomever must interpret the results of genomic analysis), microbiologists and quality managers to define the information to be captured. In recent years, several publications have addressed the minimum information required to process sequencing data [44,45]. Reliability of metadata is essential for its proper use in the interpretation of a WGS analysis. The importance for standardized, high-quality metadata, called for implementation of practices as the development of food safety specific ontologies [46,47] and, more recently, ontologies relevant to bioinformatics analyses in food safety [48]. These procedures aim to make the metadata accessible, exchangeable, and minable as well as to ease the information control and flow.

5. Discussion

Rigorous environmental monitoring in food factories aims at verifying microbiological food safety control measures and detection of a potential pathogen contamination before it reaches the product. WGS has proven to be of high sensitivity and is routinely used by authorities for pathogen source tracking in outbreak events. Similarly, WGS can be used in root cause investigation in case of factory pathogen contamination. Data produced with WGS by itself cannot provide an answer and needs to be interpreted and contextualized accordingly to the biological question asked. Additional discriminatory power of WGS analyses requires also a higher degree of expertise to understand and interpret the results. In order to obtain robust results, high quality data and analysis are required. Establishing several filters and metrics evaluation for raw data, genome assemblies and typing analyses avoid the inclusion of low-quality data that can impact the results and the subsequent interpretation. This not only aligns with the aim to obtain reliable results from the first attempt but also contributes to standardization and automation of such workflows. Capitalizing on hands-on experience to establish metrics that allow for high quality data, can also contribute to optimization of resources use (particularly time) and the obtention of interpretable information. A strong understanding of the bacterial genomics and the bioinformatics analysis is important to correctly interpret the results.

5.1. Turnaround Time

A long analysis turnaround-time (TAT) greatly limits the usability of the information generated via WGS. Implementing quality checks at several stages of the analyses not only increases the value of the analysis but also optimizes time use by having a "first time right" analysis. Every additional examination due to *a posteriori* corrections for low quality data increases TAT. Integrating quality checks in our workflow has allowed to reduce the bioinformatics analysis TAT by more than 80% (reaching in average ~1 day) and to provide timely results to factories.

5.2. cg/wgMLST versus SNP

It is widely recognized that cg/wgMLST and SNP analyses are highly concordant [49–53] for pathogen source tracking. However, contrary to cg/wgMLST, SNP typing allows for identification of mutations in the non-coding regions of the genome. This allows for the use of almost all the genetic information from a genome thus providing the theoretically highest level of precision available. This additional information may explain the cases in which the cg/wgMLST and SNP approaches were not concordant [50,54]. We have observed many cases in which cgMLST did not provide a clear-cut answer where SNP was able to fully resolve relatedness.

Combining additional information of the sample (metadata) cgMLST and SNP analyses allows formulation of data-driven actions for controlling the actual source of the contaminant and strategies to avoid the contaminant to ever reach the final product.

5.3. Standardization and Accreditation

International efforts are being carried out to harmonize protocols not only in the lab but also in the bioinformatics analysis. In an ISO working group, lab specialists, microbiologists, bioinformaticians and metadata experts are actively working on the international ISO 23418 standard (whole genome sequencing for typing and genomic characterization of foodborne bacteria) offering general guidelines (https://www.iso.org/standard/75509.html (accessed on 16 December 2020)).

Laboratory proficiency tests have been developed in Europe and in USA to monitor the WGS lab work, and are now running on a regular basis [55]. Proficiency tests for bioinformatics analysis for pathogen source tracking are less frequent, and more complicated to design due to lack of generalized guidelines. However, in recent years some have been developed and used [56,57]. This lack of consensus for data analysis highlights the importance of metrics evaluation and workflow validation [18–20] to ensure its quality.

Additionally, efforts towards standardization of the several steps in WGS bioinformatics analyses will greatly contribute to the results reproducibility and portability of the workflows. Open-source command-line software running in a high-performance computing infrastructure is often seen as the best way to perform these analyses. Standardization and data quality assurance, by optimizing resources use, could also enable the use of packaged workflows (e.g., Docker https://github.com/docker/docker-ce (accessed on 18 December 2020)) for smaller infrastructures or where cloud-computing is a possibility if IT security allows.

5.4. Internal Genome Database and Metadata Management

Having an established computing infrastructure allows also for maintaining an internal database of sequenced isolates. Although this requires data management, it offers multiple benefits. Firstly, it permits cross-referencing genomes from current and previous case studies to identify potential links between and within factories, raw materials or any information stored in the metadata. An example of this could be the case of a contaminant found in two, otherwise unrelated, facilities where the link is a shared raw material. These kinds of scenarios are by themselves a good justification for the use of an internal metadata database linked with WGS analysis. Furthermore, the sequencing data may be used for genome mining, definition of in-house allelic profiles (cg/wgMLST), antibiotic resistance surveys, mutation rates studies, among others.

Finally, having full genome information allows for other kinds of research that could ultimately benefit the food industry, for example, studying the effects of cleaning agents on microbial resistance to biocides [58–61]. Additionally, complete genome sequences can be used to predict antimicrobial resistance genes or virulence genes presence [62,63].

5.5. Analysis Reproducibility and Repeatability

As mentioned before, the fast advancements on sequencing technologies are often accompanied by developments of the bioinformatics software to handle these data. It is

important then to consider that multiple versions exist for the several software available and the impact it may have on the obtained results. For example, different versions of mapping and alignment processing software can produce different results when performing SNP calling, this depends among others on the dataset used (https://snp-pipeline.readthedocs.io/en/latest/reproducible.html (accessed on 10 December 2020)). From one software version to another, algorithm, parameters used and associated databases might change and then impact the result generated. For example, different allele databases can lead to different results when performing cgMLST analysis [64,65]. Additionally, handler misuse by lack of understanding of those possible changes might also lead to result differences. It is thus important to document and report the versions used to allow other users to reproduce the results and to be aware of the impact that a different version can have on their analyses. Likewise, if the software in the workflow uses a database, it is important to keep these databases updated when possible and to keep a log of the changes/updates.

5.6. Needed Expertise and Knowledge

The question that arises prior implementing WGS for pathogen contamination root cause investigation in food factories is: Which kind of expertise is required to use WGS and its results? As the end-to-end workflow encompasses sample taking, sample diagnostics, pathogen isolation, DNA extraction from the isolate, sequencing, bioinformatics analysis and results interpretation, the required know-how should cover these domains. The sequencing laboratory part (DNA extraction, library preparation and sequencing) is frequently outsourced, this does not imply blindly trusting the received data, on the contrary, there should be an additional check to ensure its quality as the specific lab practices are less well known. If ultimately a workflow gets integrated into a pipeline that can be run on a cloud-based server, scripting and programming would not be a must, but understanding the bioinformatics behind (sequence alignment, variant calling, k-mer based typing, etc.) remains an important part to ensure high quality results and to, eventually, perform troubleshooting when needed. As mentioned several times, data is not stand-alone, expertise in microbiology and food safety is necessary to link variant call data and metadata for a comprehensive and reliable interpretation in the factory context. With certain deployments in terms of wet-lab and bioinformatics workflow, it may not be relevant to have a person dedicated to each one of the steps (sequencing, data processing and results interpretation) but overall knowledge of each stage is indeed needed.

In many cases, cloud-based web interfaces (e.g., GalaxyTrakr available at https://www.galaxytrakr.org (accessed on 9 December 2020)) or windows-based commercial software have been developed and deployed. This represents a significative advantage as it enables data analysis without command-line or Linux-based infrastructure, permitting thus a wider adoption, when computing infrastructure or programming expertise is not readily available. It is, however, important to consider that such systems are not "fool-proof". A critical eye is still needed so the tool does not turn into a black box where the input and output are known but the procedure is not controlled. Often, all stages of the analysis use specific software, each one with multiple parameters and the deployed interfaces offer a higher or lesser control over those parameters that needs to be considered. Neglecting control over these parameters can interfere when comparing runs, even when using the same software and versions. Adding all those quality verifications at each stage should increase the level of confidence in the obtained results.

6. Conclusions

The intention is not to consider the described workflow as the new "gold standard" in WGS analyses for pathogen source tracking, but rather present our experiences as food industry and open the discussion about practices to be implemented when using this kind of data. We believe that sharing this stepwise approach with the community is a stage towards standardization of the analysis for robust bioinformatics data analysis and reliable results interpretation. Such a workflow can be implemented in commercial software

solutions or in open-source pipelines and would simplify the data analysis for users. Optimization of the workflow allows also for reducing "time-to-result" and indirectly costs. Once agreed upon guidelines will be further defined, more food companies may want to start using WGS in their surveillance programs. Ultimately, this work aims at promoting and facilitating the use of this, highly beneficial technology, to become a routine analysis among the food safety tools already available.

Author Contributions: Conceptualization, A.-C.P., C.N.-B. and C.B.; methodology, C.N.-B., A.-C.P., C.R. and C.B.; formal analysis, A.-C.P., C.R. and C.B.; validation, C.N.-B., A.-C.P., C.R. and C.B.; data curation, C.R. and C.B.; writing—original draft preparation, C.R. and C.B.; writing—review and editing C.N.-B., A.-C.P., C.R. and C.B. 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.

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

References

1. Rantsiou, K.; Kathariou, S.; Winkler, A.; Skandamis, P.; Saint-Cyr, M.J.; Rouzeau-Szynalski, K.; Amézquita, A. Next generation microbiological risk assessment: Opportunities of whole genome sequencing (WGS) for foodborne pathogen surveillance, source tracking and risk assessment. *Int. J. Food Microbiol.* **2018**, *287*, 3–9. [CrossRef]
2. Rouzeau-Szynalski, K.; Barretto, C.; Fournier, C.; Moine, D.; Gimonet, J.; Baert, L. Whole genome sequencing used in an industrial context reveals a *Salmonella* laboratory cross-contamination. *Int. J. Food Microbiol.* **2019**, *298*, 39–43. [CrossRef]
3. EFSA Panel on Biological Hazards (EFSA BIOHAZ Panel); Koutsoumanis, K.; Allende, A.; Alvarez-Ordóñez, A.; Bolton, D.; Bover-Cid, S.; Chemaly, M.; Davies, R.; De Cesare, A.; Hilbert, F.; et al. Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. *EFSA J.* **2019**, *17*, e05898. [CrossRef] [PubMed]
4. Hurley, D.; Luque-Sastre, L.; Parker, C.T.; Huynh, S.; Eshwar, A.K.; Van Nguyen, S.; Andrews, N.; Moura, A.; Fox, E.M.; Jordan, K.; et al. Whole-genome sequencing-based characterization of 100 *Listeria monocytogenes* isolates collected from food processing environments over a four-year period. *mSphere* **2019**, *4*, 1–14. [CrossRef]
5. Wang, S.; Weller, D.L.; Falardeau, J.; Strawn, L.K.; Mardones, F.O.; Adell, A.; Switt, A.I.M. Food safety trends: From globalization of whole genome sequencing to application of new tools to prevent foodborne diseases. *Trends Food Sci. Technol.* **2016**, *57*, 188–198. [CrossRef]
6. Brown, E.; Dessai, U.; McGarry, S.; Gerner-Smidt, P. Use of whole-genome sequencing for food safety and public health in the United States. *Foodborne Pathog. Dis.* **2019**, *16*, 441–450. [CrossRef]
7. Allard, M.W.; Strain, E.; Melka, D.; Bunning, K.; Musser, S.M.; Brown, E.W.; Timme, R. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *J. Clin. Microbiol.* **2016**, *54*, 1975–1983. [CrossRef]
8. Alegbeleye, O.O.; Sant'Ana, A.S. Pathogen subtyping tools for risk assessment and management of produce-borne outbreaks. *Curr. Opin. Food Sci.* **2020**, *32*, 83–89. [CrossRef]
9. Jackson, B.R.; Tarr, C.; Strain, E.; Jackson, K.A.; Conrad, A.; Carleton, H.; Katz, L.S.; Stroika, S.; Gould, L.H.; Mody, R.K.; et al. Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clin. Infect. Dis.* **2016**, *63*, 380–386. [CrossRef] [PubMed]
10. Yoshimura, D.; Kajitani, R.; Gotoh, Y.; Katahira, K.; Okuno, M.; Ogura, Y.; Hayashi, T.; Itoh, T. Evaluation of SNP calling methods for closely related bacterial isolates and a novel high-accuracy pipeline: BactSNP. *Microb. Genom.* **2019**, *5*, e000261. [CrossRef]
11. Kwong, J.C.; Stafford, R.; Strain, E.; Stinear, T.; Seemann, T.; Howden, B.P. Sharing is caring: International sharing of data enhances genomic surveillance of *Listeria monocytogenes*. *Clin. Infect. Dis.* **2016**, *63*, 846–848. [CrossRef] [PubMed]
12. Thompson, C.K.; Wang, Q.; Bag, S.K.; Franklin, N.; Shadbolt, C.T.; Howard, P.; Fearnley, E.J.; Quinn, H.E.; Sintchenko, V.; Hope, K.G. Epidemiology and whole genome sequencing of an ongoing point-source *Salmonella* Agona outbreak associated with sushi consumption in western Sydney, Australia 2015. *Epidemiol. Infect.* **2017**, *145*, 2062–2071. [CrossRef] [PubMed]
13. Van Walle, I.; Guerra, B.; Borges, V.; Carriço, J.A.; Cochrane, G.; Dallman, T.; Franz, E.; Karpíšková, R.; Litrup, E.; Mistou, M.-Y.; et al. EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human, animal, food, feed and food/feed environmental samples in the joint ECDC-EFSA molecular typing database. *EFSA Support. Publ.* **2019**, *EN-1337*, 1–92. [CrossRef]
14. Ferguson, B. Adoption of WGS: What is Going On? *Food Safety Magazine*. 2020. Available online: https://www.foodsafetymagazine.com/magazine-archive1/augustseptember-2020/adoption-of-wgs-what-is-going-on/ (accessed on 16 December 2020).

15. Klijn, A.D.; Akins-Lewenthal, B.; Jagadeesan, L.; Baert, A.; Winkler, C.B.; Amézquita, A. The Benefits and Barriers of Whole-Genome Sequencing for Pathogen Source Tracking: A Food Industry Perspective. *Food Safety Magazine*. 2020. Available online: https://www.foodsafetymagazine.com/magazine-archive1/junejuly-2020/the-benefits-and-barriers-of-whole-genome-sequencing-for-pathogen-source-tracking-a-food-industry-perspective/ (accessed on 14 December 2020).
16. Davis, S.; Pettengill, J.B.; Luo, Y.; Payne, J.; Shpuntoff, A.; Rand, H.; Strain, E. CFSAN SNP Pipeline: An automated method for constructing SNP matrices from next-generation sequence data. *PeerJ Comput. Sci.* **2015**, *1*, e20. [CrossRef]
17. Kwong, J.C.; Mercoulia, K.; Tomita, T.; Easton, M.; Li, H.Y.; Bulach, D.M.; Stinear, T.P.; Seemann, T.; Howden, B.P. Prospective Whole-Genome Sequencing Enhances National Surveillance of *Listeria monocytogenes*. *J. Clin. Microbiol.* **2016**, *54*, 333–342. [CrossRef] [PubMed]
18. Bogaerts, B.; Winand, R.; Fu, Q.; Van Braekel, J.; Ceyssens, P.-J.; Mattheus, W.; Bertrand, S.; De Keersmaecker, S.C.J.; Roosens, N.H.C.; Vanneste, K. Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: *Neisseria meningitidis* as a Proof-of-Concept. *Front. Microbiol.* **2019**, *10*, 362. [CrossRef] [PubMed]
19. Lepuschitz, S.; Weinmaier, T.; Mrazek, K.; Beisken, S.; Weinberger, J.; Posch, A.E. Analytical Performance Validation of Next-Generation Sequencing Based Clinical Microbiology Assays Using a K-mer Analysis Workflow. *Front. Microbiol.* **2020**, *11*, 1883. [CrossRef] [PubMed]
20. Portmann, A.-C.; Fournier, C.; Gimonet, J.; Ngom-Bru, C.; Barretto, C.; Baert, L. A Validation Approach of an End-to-End Whole Genome Sequencing Workflow for Source Tracking of *Listeria monocytogenes* and *Salmonella enterica*. *Front. Microbiol.* **2018**, *9*, 446. [CrossRef]
21. Andrews, S. Fastqc: A Quality Control Tool for High Throughput Sequence Data. 2010. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 17 December 2020).
22. Den Bakker, H.; van Heus, P.; Zhang, S. Rapid Confirmation of *Salmonella* Spp. and Subspp. From Sequence Data. 2018. Available online: https://github.com/hcdenbakker/SalmID (accessed on 17 December 2020).
23. Seeman, T. Rapid 16s rDNA from Isolate Fastq Files. 2017. Available online: https://github.com/tseemann/sixess (accessed on 17 December 2020).
24. Schaefer, U.; Gallop, S. K-Mer Based Isolate of Fastq Reads Against Reference Genomes. 2014. Available online: https://github.com/phe-bioinformatics/kmerid (accessed on 17 December 2020).
25. Wood, D.E.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, R46. [CrossRef]
26. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef]
27. 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.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef]
28. Souvorov, A.; Agarwala, R.; Lipman, D.J. SKESA: Strategic k-mer extension for scrupulous assemblies. *Genome Biol.* **2018**, *19*, 1–13. [CrossRef]
29. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
30. Timme, R.; Wolfgang, W.J.; Balkey, M.; Venkata, S.L.G.; Randolph, R.; Allard, M.W.; Strain, E. Optimizing open data to support one health: Best practices to ensure interoperability of genomic data from bacterial pathogens. *One Health Outlook* **2020**, *2*, 1–11. [CrossRef] [PubMed]
31. Seeman, T. Scan Contig Files Against Pubmlst Typing Schemes. 2018. Available online: https://github.com/tseemann/mlst (accessed on 16 December 2020).
32. Larsen, M.V.; Cosentino, S.; Rasmussen, S.; Friis, C.; Hasman, H.; Marvig, R.L.; Jelsbak, L.; Sicheritz-Pontén, T.; Ussery, D.W.; Aarestrup, F.M.; et al. Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* **2012**, *50*, 1355–1361. [CrossRef]
33. Gupta, A.; Jordan, I.K.; Rishishwar, L. stringMLST: A fast k-mer based tool for multilocus sequence typing. *Bioinformatics* **2017**, *33*, 119–121. [CrossRef] [PubMed]
34. Yoshida, C.E.; Kruczkiewicz, P.; Laing, C.R.; Lingohr, E.J.; Gannon, V.P.J.; Nash, J.H.E.; Taboada, E.N. The *Salmonella in silico* typing resource (SISTR): An open web-accessible tool for rapidly typing and subtyping draft *Salmonella* genome assemblies. *PLoS ONE* **2016**, *11*, e0147101. [CrossRef]
35. Zhang, S.; Bakker, H.C.D.; Li, S.; Chen, J.; Dinsmore, B.A.; Lane, C.; Lauer, A.C.; Fields, P.I.; Deng, X. SeqSero2: Rapid and improved *Salmonella* serotype determination using whole-genome sequencing data. *Appl. Environ. Microbiol.* **2019**. [CrossRef] [PubMed]
36. Carroll, L.M.; Kovac, J.; Miller, R.A.; Wiedmann, M. Rapid, high-throughput identification of anthrax-causing and emetic *Bacillus cereus* group genome assemblies via BTyper, a computational tool for virulence-based classification of *Bacillus cereus* group isolates by using nucleotide sequencing data. *Appl. Environ. Microbiol.* **2017**. [CrossRef] [PubMed]

37. Diep, B.; Barretto, C.; Portmann, A.-C.; Fournier, C.; Karczmarek, A.; Voets, G.; Li, S.; Deng, X.; Klijn, A. *Salmonella* serotyping; Comparison of the traditional method to a microarray-based method and an *in silico* platform using whole genome sequencing data. *Front. Microbiol.* **2019**, *10*, 2554. [CrossRef]
38. Silva, M.; Machado, M.P.; Silva, D.N.; Rossi, M.; Moran-Gilad, J.; Santos, S.; Ramirez, M.; Carriço, J.A. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. *Microb. Genom.* **2018**, *4*, e000166. [CrossRef]
39. Starikova, E.V.; Tikhonova, P.O.; Prianichnikov, N.A.; Rands, C.M.; Zdobnov, E.M.; Ilina, E.N.; Govorun, V.M. Phigaro: High-throughput prophage sequence annotation. *Bioinformatics* **2020**, *36*, 3882–3884. [CrossRef] [PubMed]
40. Reis-Cunha, J.L.; Bartholomeu, D.C.; Manson, A.L.; Earl, A.M.; Cerqueira, G.C. ProphET, prophage estimation tool: A stand-alone prophage sequence prediction tool with self-updating reference database. *PLoS ONE* **2019**, *14*, e0223364. [CrossRef] [PubMed]
41. Robertson, J.; Nash, J.H.E. MOB-suite: Software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb. Genom.* **2018**, *4*, e000206. [CrossRef] [PubMed]
42. Pightling, A.W.; Petronella, N.; Pagotto, F. Choice of reference sequence and assembler for alignment of *Listeria monocytogenes* short-read sequence data greatly influences rates of error in SNP analyses. *PLoS ONE* **2014**, *9*, e104579. [CrossRef] [PubMed]
43. Li, S.; Zhang, S.; Baert, L.; Jagadeesan, B.; Ngom-Bru, C.; Griswold, T.; Katz, L.S.; Carleton, H.A.; Deng, X. Implications of mobile genetic elements for *Salmonella enterica* single-nucleotide polymorphism subtyping and source tracking investigations. *Appl. Environ. Microbiol.* **2019**, *85*, 1–12. [CrossRef]
44. Bowers, R.M.; The Genome Standards Consortium; Kyrpides, N.C.; Stepanauskas, R.; Harmon-Smith, M.; Doud, D.; Reddy, T.B.K.; Schulz, F.; Jarett, J.; Rivers, A.R.; et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **2017**, *35*, 725–731. [CrossRef]
45. Yilmaz, P.; Kottmann, R.; Field, D.; Knight, R.; Cole, J.R.; Amaralzettler, L.A.; Gilbert, J.A.; Karsch-Mizrachi, I.; Johnston, A.; Cochrane, G.; et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* **2011**, *29*, 415–420. [CrossRef] [PubMed]
46. EFSA. The food classification and description system FoodEx 2 (revision 2). *EFSA Support. Publ.* **2015**, *12*. [CrossRef]
47. Ireland, J.D.; Møller, A. LanguaL Food Description: A Learning Process. *Eur. J. Clin. Nutr.* **2010**, *64*, S44–S48. [CrossRef]
48. Lambert, D.; Pightling, A.; Griffiths, E.; Van Domselaar, G.; Evans, P.; Berthelet, S.; Craig, D.; Chandry, P.S.; Stones, R.; Brinkman, F.; et al. Baseline practices for the application of genomic data supporting regulatory food safety. *J. AOAC Int.* **2017**, *100*, 721–731. [CrossRef] [PubMed]
49. Coipan, C.E.; Dallman, T.J.; Brown, D.; Hartman, H.; Van Der Voort, M.; Berg, R.R.V.D.; Palm, D.; Kotila, S.; Van Wijk, T.; Franz, E. Concordance of SNP- and allele-based typing workflows in the context of a large-scale international *Salmonella* Enteritidis outbreak investigation. *Microb. Genom.* **2020**, *6*, e000318. [CrossRef]
50. Gona, F.; Comandatore, F.; Battaglia, S.; Piazza, A.; Trovato, A.; Lorenzin, G.; Cichero, P.; Biancardi, A.; Nizzero, P.; Moro, M.; et al. Comparison of core-genome MLST, coreSNP and PFGE methods for *Klebsiella pneumoniae* cluster analysis. *Microb. Genom.* **2020**, *6*, mgen000347. [CrossRef]
51. Blanc, D.S.; Magalhães, B.; Koenig, I.; Senn, L.; Grandbastien, B. Comparison of Whole Genome (wg-) and Core Genome (cg-) MLST (BioNumericsTM) Versus SNP Variant Calling for Epidemiological Investigation of *Pseudomonas aeruginosa*. *Front. Microbiol.* **2020**, *11*, 1729. [CrossRef]
52. Henri, C.; Leekitcharoenphon, P.; Carleton, H.A.; Radomski, N.; Kaas, R.S.; Mariet, J.-F.; Felten, A.; Aarestrup, F.M.; Smidt, P.G.; Roussel, S.; et al. An assessment of different genomic approaches for inferring phylogeny of *Listeria monocytogenes*. *Front. Microbiol.* **2017**, *8*, 2351. [CrossRef]
53. Pearce, M.E.; Alikhan, N.-F.; Dallman, T.J.; Zhou, Z.; Grant, K.; Maiden, M.C. Comparative analysis of core genome MLST and SNP typing within a European *Salmonella* serovar Enteritidis outbreak. *Int. J. Food Microbiol.* **2018**, *274*, 1–11. [CrossRef]
54. Tsang, A.K.L.; Lee, H.H.; Yiu, S.-M.; Lau, S.K.P.; Woo, P.C.Y. Failure of phylogeny inferred from multilocus sequence typing to represent bacterial phylogeny. *Sci. Rep.* **2017**, *7*, 4536. [CrossRef]
55. Timme, R.E.; Rand, H.; Leon, M.S.; Hoffmann, M.; Strain, E.; Allard, M.; Roberson, D.; Baugher, J.D. GenomeTrakr proficiency testing for foodborne pathogen surveillance: An exercise from 2015. *Microb. Genom.* **2018**, *4*, e000185. [CrossRef] [PubMed]
56. ECDC. *Proficiency Test for Listeria monocytogenes Whole Genome Assembly*; ECDC: Solna, Sweden, 2019. [CrossRef]
57. Lau, K.A.; da Silva, A.G.; Ballard, S.A.; Theis, T.; Gray, J.; Rawlinson, W.D. Proficiency Testing for bacterial whole genome sequencing in assuring the quality of microbiology diagnostics in clinical and public health laboratories. *bioRxiv* **2020**. [CrossRef]
58. Pruden, A.; Larsson, D.G.J.; Amézquita, A.; Collignon, P.; Brandt, K.K.; Graham, D.W.; Lazorchak, J.M.; Suzuki, S.; Silley, P.; Snape, J.R.; et al. Management options for reducing the release of antibiotics and antibiotic resistance genes to the environment. *Environ. Health Perspect.* **2013**, *121*, 878–885. [CrossRef]
59. Nijsingh, N.; Munthe, C.; Larsson, D.G.J. Managing pollution from antibiotics manufacturing: Charting actors, incentives and disincentives. *Environ. Health* **2019**, *18*, 1–18. [CrossRef]
60. Fahimipour, A.K.; Ben Mamaar, S.; McFarland, A.G.; Blaustein, R.A.; Chen, J.; Glawe, A.J.; Kline, J.; Green, J.L.; Halden, R.U.; Wymelenberg, K.V.D.; et al. Antimicrobial chemicals associate with microbial function and antibiotic resistance indoors. *mSystems* **2018**, *3*, e00200-18. [CrossRef]
61. Kampf, G. Biocidal agents used for disinfection can enhance antibiotic resistance in gram-negative species. *Antibiotics* **2018**, *7*, 110. [CrossRef]

62. Ellington, M.; Ekelund, O.; Aarestrup, F.; Canton, R.; Doumith, M.; Giske, C.; Grundman, H.; Hasman, H.; Holden, M.; Hopkins, K.; et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: Report from the EUCAST Subcommittee. *Clin. Microbiol. Infect.* **2017**, *23*, 2–22. [CrossRef] [PubMed]
63. Oniciuc, E.A.; Likotrafiti, E.; Alvarez-Molina, A.; Prieto, M.; Santos, J.A.; Alvarez-Ordóñez, A. The present and future of whole genome sequencing (WGS) and whole metagenome sequencing (WMS) for surveillance of antimicrobial resistant microorganisms and antimicrobial resistance genes across the food chain. *Genes* **2018**, *9*, 268. [CrossRef] [PubMed]
64. Pietzka, A.; Allerberger, F.; Murer, A.; Lennkh, A.; Stöger, A.; Rosel, A.C.; Huhulescu, S.; Maritschnik, S.; Springer, B.; Lepuschitz, S.; et al. Whole genome sequencing based surveillance of *L. monocytogenes* for early detection and investigations of listeriosis outbreaks. *Front. Public Health* **2019**, *7*, 139. [CrossRef] [PubMed]
65. Higgins, P.G.; Prior, K.; Harmsen, D.; Seifert, H. Development and evaluation of a core genome multilocus typing scheme for whole-genome sequence-based typing of *Acinetobacter baumannii*. *PLoS ONE* **2017**, *12*, e0179228. [CrossRef] [PubMed]