Trends in Food Science & Technology 73 (2018) 67–75

Contents lists available at ScienceDirect

Trends in Food Science & Technology

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

Review

Whole genome sequencing as a typing tool for foodborne pathogens like *Listeria monocytogenes* – The way towards global harmonisation and data exchange

Stefanie Lüth[a,*], Sylvia Kleta[a], Sascha Al Dahouk[a,b]

[a] German Federal Institute for Risk Assessment (BfR), National Reference Laboratory for Listeria monocytogenes, Department of Biological Safety, Diedersdorfer Weg 1, 12277 Berlin, Germany

[b] RWTH Aachen University Hospital, Pauwelsstraße 30, 52074 Aachen, Germany

ABSTRACT

Background: Various molecular typing methods are used for the surveillance of foodborne pathogens and outbreak investigations, differing widely in information content and discriminatory power. Presently, not least because of the rapid technological development, the focus is shifting to whole genome sequencing (WGS) as an analytical tool. As a result of globalisation of food trade, a comprehensive understanding of the association between the occurrence of human infections and causative pathogens has to be established to monitor and to prevent their spread. In this respect, the accuracy of WGS clearly supersedes that of previous tools.

Scope and approach: Our review describes the status quo of WGS in surveillance and outbreak investigations of foodborne pathogens through the example of *Listeria monocytogenes*. It highlights the value of WGS in trace-back of infections to food sources and provides an overview of methods used for data generation (wet lab) and analysis (dry lab). Altogether, progress but also challenges in the worldwide practical application of WGS for bacterial typing are described.

Key findings and conclusions: The current status of WGS differs widely between countries and even laboratory sites. A consensus has to be found concerning methods, quality measures, thresholds for data generation and analysis as well as rules for data sharing. International harmonisation is going to be indispensable on the way to data exchangeability which will finally support global control of foodborne pathogens.

1. Introduction

The gram positive bacterium *Listeria monocytogenes* (*L. monocytogenes*) is the causative agent of the infectious disease listeriosis in humans. Although it is widely distributed in the environment, transmission of *L. monocytogenes* to humans mainly occurs via consumption of contaminated food, especially pre-packaged ready-to-eat products. Its ability to form biofilms paired with its resilience to salt, low temperatures and acidic environments enables survival or growth even in preserved and chilled food products rendering *L. monocytogenes* a serious foodborne pathogen (Swaminathan & Gerner-Smidt, 2007).

In 2015, a total of 2,206 human listeriosis cases were reported in the EU, corresponding to an incidence rate of 0.46 per 100,000 population (EFSA & ECDC, 2016). In spite of the low incidence of listeriosis, hospitalisation rates above 90% and mortalities of 20–30% make the disease a serious public health concern. Infection of otherwise healthy adults is rare, mostly leading to non-invasive, mild listeriosis (febrile gastroenteritis) or even absence of symptoms (Aureli et al., 2000). However, cases may accumulate in risk groups, including elderly, pregnant or immunocompromised patients. Then listeriosis can be an invasive disease associated with septicaemia or focal complications such as encephalitis and meningitis (Vázquez-Boland et al., 2001).

Globalisation of food trade, changing consumption habits towards highly processed foods and demographic changes with increase of susceptible populations have augmented risk of foodborne illnesses (Wang et al., 2016). As a result from the high case fatality rate, listeriosis is a notifiable disease in the vast majority of EU member states and associated countries. Occurrence of disease can either be sporadic or outbreak-related. Since foodborne outbreaks are of public health relevance and also cause tremendous economic losses e.g. due to product recall, internationally cross-linked surveillance of *L. monocytogenes* in humans and food is of crucial importance to identify clusters, trace the sources of infections and control outbreaks.

In order to identify epidemiologically linked isolates and thus be

* Corresponding author.
E-mail address: stefanie.lueth@bfr.bund.de (S. Lüth).

https://doi.org/10.1016/j.tifs.2018.01.008
Received 22 August 2017; Received in revised form 4 December 2017; Accepted 18 January 2018
Available online 06 February 2018

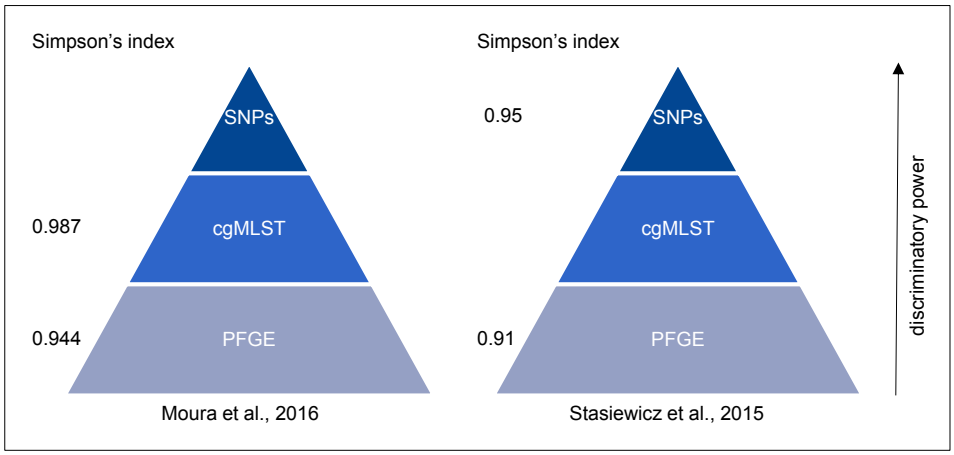

Fig. 1. Comparison of Simpson's indices of WGS-based cgMLST and SNP analysis, and PFGE.
Discriminatory power of WGS-based typing methods exceeds that of PFGE typing as shown by comparison of Simpson's indices. Studies were conducted analysing 100 *L. monocytogenes* isolates (Moura et al., 2016) and 188 *L. monocytogenes* isolates (Stasiewicz et al., 2015), respectively.

able to determine outbreak strains, differentiation of *L. monocytogenes* beyond the species level is indispensable. Consequently, several molecular typing methods have been established that are able to characterise, discriminate and index subtypes of microorganisms (ECDC, 2013). The first subtyping method applied to *L. monocytogenes* was serotyping (Seeliger & Höhne, 1979). Allocation to serotypes is still the initial step to roughly classify *L. monocytogenes* strains, although nowadays implemented by identification of serotype-specific genes through multiplex PCR assay rather than by agglutination (Doumith, Buchrieser, Glaser, Jacquet, & Martin, 2004). In the meantime, a great variety of typing methods have been developed, differing widely according to their discriminatory power, reproducibility, hands-on time and cost. In general, two different typing approaches exist, based either on phenotypic or on genotypic characteristics. Typing and identification methods based on the phenotype include e.g. classical serotyping with antisera, phage typing, multi-locus enzyme electrophoresis or esterase typing. A major weakness of these phenotypic methods is that some captured properties may vary in response to external stress or in dependence of the bacterial growth phase (Liu, 2006). To overcome these difficulties, several DNA-based methods have been established (Brosch, Buchrieser, & Rocourt, 1991). These genotypic methods aim to investigate DNA fragment length polymorphisms of restriction fragments (e.g. by ribotyping, pulsed-field gel electrophoresis), amplified DNA fragments (e.g. by random amplification of polymorphic DNA, repetitive element-PCR) or both (e.g. by amplified fragment length polymorphism), or polymorphisms that are directly found in the nucleotide sequence (e.g. by multi-locus sequence typing, multiple-locus variable number tandem repeat analysis, DNA microarray, whole genome sequencing). Among these techniques, due to its high discriminatory power, a good database and high level of standardisation, pulsed-field gel electrophoresis (PFGE) is often considered as gold standard for subtyping of *L. monocytogenes*. PFGE relies on the analysis of fragments generated by DNA digest with rarely cutting restriction enzymes, in the case of *L. monocytogenes* with *Asc*I and *Apa*I, and their separation on an agarose gel applying a periodically changing electrical field. Comparison of restriction patterns provides information on the relatedness of different isolates. Harmonisation of experimental procedures and data analysis through the European Union Reference Laboratory for L. *monocytogenes* (EURL *Lm*) and the PulseNet International Network USA as well as set-up of centralised databases like the EURL *Lm* database emphasise the fundamental importance of PFGE for routine surveillance of *L. monocytogenes* (Félix, Danan, Van Walle, et al., 2014; Félix et al., 2013; Graves & Swaminathan, 2001). However, PFGE is a rather time-consuming and labour-intensive technique. Furthermore, discriminatory power of PFGE profiles is limited as only nucleotide changes in the restriction enzyme recognition sites are detected. Consequently, relatedness of strains may be over- or underestimated and epidemiologically unrelated isolates may be assigned to one 'pseudo'-cluster whereas even highly related strains fall into distinct clusters.

Thanks to recent developments in next generation sequencing technologies (NGS), whole-genome sequencing (WGS) as a typing tool for *L. monocytogenes* and other foodborne pathogens is gaining in importance (Deng, den Bakker & Hendriksen, 2016; Gilchrist, Turner, Riley, Petri, & Hewlett, 2015). Sequencing of entire bacterial genomes provides an unparalleled depth of information. Base-by-base comparisons of entire genomes are possible as well as retrieval of additional information such as virulence or antimicrobial resistance markers. As opposed to traditional molecular methods like PFGE, WGS provides comprehensive insight into evolution of bacterial strains.

Comparing Simpson's Indices of molecular typing methods, discriminatory power of WGS-based typing clearly exceeds that of PFGE. The Simpson's Index is used to quantify the probability that two unrelated strains are assigned to different typing groups (Hunter & Gaston, 1988). Although the Simpson's Index only yields study-specific values as it depends on the number of identified types and of the isolates that fall into these types, it is a valuable tool for a relative, quantitative evaluation of the discriminatory power of typing methods (Fig. 1). In a study including 188 *L. monocytogenes* isolates from 30 retail delis in three U.S. states over 2 years, Simpson's Index for WGS-based single nucleotide polymorphism (SNP)-based subtyping was 0.95, compared to 0.91 for PFGE (Stasiewicz, Oliver, Wiedmann, & den Bakker, 2015). In a comparison between WGS-based core genome multi-locus sequence typing (cgMLST) and PFGE, Simpson's Indices of 0.987 and 0.944 were calculated based on the analysis of 100 isolates (Moura et al., 2016). In both cases unprecedented precision of outbreak investigations became possible using WGS.

Our reviews aims to demonstrate the value of WGS especially compared to PFGE in the field of foodborne pathogen surveillance through introduction into methodical aspects and presentation of application examples. We intend to provide an application-oriented overview on different approaches towards global data exchangeability and challenges involved, considering differences between the EU, the USA but also international initiatives.

2. Outbreak investigations using WGS for typing

In the USA, WGS was set up for *L. monocytogenes* outbreak investigations by the Centers for Disease Control and Prevention (CDC) in cooperation with the Food and Drug Administration (FDA), the National Institutes of Health (NIH) and the United States Department of Agriculture (USDA) in September 2013 with the long-term aim to completely supersede the pool of other typing techniques including PFGE (Jackson et al., 2016). Several independent studies have provided evidence of the usefulness and value of WGS in outbreak investigations compared to other molecular typing methods (see below).

In an outbreak related to contaminated ice cream, listeriosis cases first accumulated in a single hospital in Kansas from 2014 to 2015 (CDC., 2015a). Although isolates from two of the affected patients shared a common PFGE pattern, patterns for three other patients did not, suggesting independence of cases. However, this assumption was rejected when cgMLST based on WGS data identified four of the five isolates as highly related and thus allowed their link to the outbreak. SNP-based analysis of the WGS data confirmed this attribution by proving a difference of 1–19 SNPs between clinical isolates (Chen et al., 2017). In another outbreak in 2015, WGS analysis did not only allow cluster identification that was not possible via PFGE typing, but also enabled retrospective inclusion of a previously unsolved cluster from 2013 into the outbreak and trace-back to contaminated soft cheese (CDC., 2015c). Overall, routine use of WGS for typing of *L. monocytogenes* isolates from clinical and food samples in the USA has proven undoubtedly successful. The resolution of WGS exceeds the discriminatory power of PFGE and provides more precise and reliable data. As a result, smaller outbreaks can be recognised that would have otherwise been considered as sporadic cases. Furthermore, retrospective analysis allows grouping of individual sporadic cases over a longer period of time to one single outbreak and enables to link outbreak strains to a common source. Consecutive regulatory steps like product recalls or controlled sanitation of production plants can then prevent further listeriosis cases. Since its implementation for routine surveillance in the USA in 2013, WGS typing helped solving a variety of food-related listeriosis outbreaks and also to identify uncommon sources in a listeriosis outbreak, 2014–2015, linked to pre-packaged caramel apples (CDC., 2015b). Overall, identification of more outbreaks with fewer cases per outbreak becomes possible using WGS-based typing (CDC., 2016).

In the EU and associated countries, as opposed to the USA, WGS has not yet entered the status of a comprehensive routine typing method for *L. monocytogenes*. Between 2013 and 2015, twelve to fifteen outbreaks of listeriosis per year have been reported in the EU (EFSA & ECDC, 2016). The vast majority of them were resolved using traditional typing techniques like PFGE in combination with epidemiological evidence. However, to date, several exemplary studies investigating foodborne outbreaks by NGS techniques have been performed and published (Gillesberg Lassen et al., 2016; Kleta et al., 2017; Kvistholm Jensen et al., 2016; Ruppitsch, Prager, et al., 2015; Schmid et al., 2014).

In 2014, within an international collaboration of public health institutes and food authorities, a cluster of seven human listeriosis cases in Germany and Austria that emerged between April 2011 and July 2013 could be identified (Schmid et al., 2014). Initially, the respective outbreak strains were typed with PFGE and fluorescent amplified fragment length polymorphism (fAFLP) where they appeared indistinguishable and were assigned to one cluster. CgMLST based on WGS data, however, was capable to distinguish a cluster of four outbreak strains ($\leq$6 allelic differences) isolated in 2012–2013 from the other three strains isolated in 2011 ($\geq$48 allelic differences) that could subsequently be excluded from the outbreak. In addition, the four confirmed outbreak cases could be traced back to two different Austrian food products, an unaged soft cheese (food isolates differing $\leq$19 alleles from the human cluster) and a deli-meat (food isolates differing $\leq$8 alleles from the human cluster). However, no final attribution could be made because thresholds for strain differentiation have not yet been defined. Recently, a genetic distance of $\leq$10 alleles between human and food isolates has been proposed for unambiguous source attribution which would exclude the soft cheese as a possible source (Ruppitsch, Pietzka, et al., 2015).

During a long-term outbreak of listeriosis in Southern Germany from 2012 to 2015, WGS and cgMLST were used to confirm clustering of human isolates with an unusual PFGE pattern into one outbreak group (Ruppitsch, Prager, et al., 2015). Although six food-related isolates from Austria and Germany showed PFGE patterns identical to the human isolates, WGS revealed their belonging to independent cluster types. Thus, faulty source attribution could be averted. Later, the human cases could be traced back to a contaminated batch of smoked pork belly (Kleta et al., 2017). This observation underlines the importance of WGS for successful and reliable trace-back of listeriosis cases to a food source. Nonetheless, although PFGE cannot keep pace with the discriminatory power of WGS data based analysis, it might still be a suitable alternative in countries or regions where NGS is not established because of economic reasons.

Among European countries, Denmark has officially initiated nationwide real-time WGS typing of human *L. monocytogenes* isolates for routine surveillance in September 2013 (Kvistholm Jensen et al., 2016). In addition, interviews exploring consumption habits of listeriosis patients have been implemented and added to the typing data since June 2014. So far, these combined databases have proven successful in two different outbreak investigations. In 2014, 41 listeriosis cases in Denmark were assigned to one outbreak cluster through WGS-based SNP analysis with genetic differences between the isolates not exceeding 3 SNPs (Kvistholm Jensen et al., 2016). In cooperation with the Danish Veterinary and Food Administration (DVFA) who provided data on routinely collected food samples, human strains could be traced back to a common source, a ready-to-eat delicatessen meat which was subsequently recalled from the national market. In this way, the outbreak could be terminated. Similarly, a total of twenty listeriosis cases notified between 2013 and 2015 in Denmark could be assigned to two distinct outbreaks, each comprising ten cases (Gillesberg Lassen et al., 2016). Both clusters could be traced back to smoked salmon or smoked halibut and trout. Again, WGS typing of human clinical isolates and routinely collected food isolates allowed reliable source attribution, enabled to impose legal measures and thereby saved lives.

3. Wet lab standardisation of NGS methods for foodborne pathogen typing

International collaboration on the control of foodborne pathogens like *L. monocytogenes* has more than ever become indispensable to guarantee food safety. For that purpose, harmonisation and standardisation of WGS-based typing methods across countries and sectors (human, animal, food) need to be established to ensure comparability of typing results and to allow data exchange. To date, a variety of protocols have been developed for typing of *L. monocytogenes* and other foodborne pathogens using NGS technologies. The current challenge lies in identification of differences and definition of generally valid quality metrics to produce consistent results.

Four main factors allow a statement on the quality of sequencing results: coverage or sequencing depth, evenness of coverage, read length and read quality (Loman, Misra, et al., 2012). Coverage describes the average number of times a genome has been sequenced. It is equal to the product of read length and number of reads divided by the haploid genome length (Lander & Waterman, 1988). Bases are usually sequenced multiple times to increase the probability that all genomic regions are covered and to compensate for possible sequencing errors in order to increase confidence in sequencing results. High coverage but also evenness of coverage is essential to be able to consider a sequencing run successful. Read quality can be assessed through Phred Quality or Q-scores. This score gives the logarithmic probability of an incorrect base (Richterich, 1998). For example Q30 represents the probability of one incorrect base in 1000. Its determination is based on the comparison of measurement parameters during base detection (e.g intensity profile, signal-to-noise ratio) with empirically determined reference parameters that are linked to known quality scores. It is to note that Q-scores are hence specific for a platform and even for new hardware, software or chemistry within a platform and are dependent on algorithms used to predict them. One major step towards global harmonisation would be the definition of general quality metrics.

Sequencing technologies have massively evolved during the last decades and are still in a process of continuous development and

Table 1
Overview of sequencers most frequently used for WGS of bacterial pathogens.

Sequencer	Provider	Scale	Technology	Data collection	Read length	Run time[a]
Second Generation Sequencers						
MiSeq	Illumina	benchtop	sequencing by synthesis	optical signal	1 × 36 bp - 2 × 300 bp	4 h - 56 h[b]
Ion Torrent PGM	Life Technologies	benchtop	semiconductor sequencing	pH change	200 - 400 bp	2 h - 7 h
HiSeq 2500	Illumina	production-scale	sequencing by synthesis	optical signal	1 × 36 bp - 2 × 250 bp	7 h (rapid run mode) - 11d [b]
Third Generation Sequencer						
PacBio RS	Pacific Biosciences	benchtop	single molecule real-time sequencing	fluorescence pulse	> 20 kb	0.5 h - 10 h

[a] Manufacturer specifications; depending on run mode, kit and read length.
[b] Includes time for cluster generation.

improvement, accompanied by a substantial cost reduction. Following the 'First Generation' Sanger sequencing (Sanger, Nicklen, & Coulson, 1977), introduction of massively parallel NGS or 'Second Generation' sequencing (SGS) in 2005 (Margulies et al., 2005) revolutionised genomics. Independent of the platform, SGS methods share three main steps to obtain raw sequence data: isolation of DNA, preparation of a sequencing library and sequencing. Methods used for the individual steps, however, differ a lot between platforms and laboratories. Preparation of the library involves fragmentation of DNA and tagging with specific adaptor sequences. These templates are then amplified during sequencing. Two different technologies are used, emulsion PCR and enrichment like in the 454 GS Junior (Roche) or the Ion Torrent PGM (Life Technologies) or solid-phase bridge amplification like in Illumina's MiSeq (Loman, Constantinidou, et al., 2012). One advantage of the MiSeq is that no pre-amplification step is needed which shortens the hands-on time. Although these three benchtop sequencing platforms all rely on the principle of sequencing-by-synthesis, differences lie in the details of sequencing chemistry and sequence reading (Table 1). SGS technologies in general are characterised by a high accuracy and throughput but short read lengths. As a result, single reads can often not be assembled to an entirely closed genome but rather yield a 'draft' genome with unfilled gaps between the reads (Loman, Misra, et al., 2012). Still, this is often sufficient for the purpose of comparative genomics of different highly related strains by mapping the reads to a reference genome (Loman, Constantinidou, et al., 2012; Ronholm, Nasheri, Petronella, & Pagotto, 2016). This so-called reference guided assembly is a valuable tool in phylogenetic and epidemiological investigations.

As an alternative to SGS, in 2011, the first single-molecule, real-time, long-read sequencer, PacBio RS II (Pacific Biosciences) has been put on the market. In this 'Third Generation' sequencer (TGS), the amplification step is omitted and sequencing is based directly on a single DNA molecule. It thereby yields much longer reads than SGS for which fragmented DNA is used (Table 1). High error rates, lower throughput and higher costs per base are disadvantages of this platform (Rhoads & Au, 2015). Nevertheless, this approach is useful in *de novo* assembly of genomes as the long reads help to close gaps between shorter reads. Advantages of second and third generation sequencing can be combined in a complementary approach called 'hybrid sequencing'. Through combination of the high accuracy of SGS and the long reads produced by TGS, a reliable, closed reference genome can be created which can subsequently be used for example for reference guided assembly.

Currently for sequencing of bacterial genomes almost exclusively SGS is used, with a main focus on Illumina sequencers (Schürch & Schaik, 2017). Although general accuracy of SGS systems is high through redundancy of reads, different sequencing technologies exhibit different error characteristics (Junemann et al., 2013; Loman, Misra, et al., 2012). Among benchtop sequencers, the Illumina MiSeq revealed the lowest error rate (rate of < 0.001 indels per 100 bases). It also had the highest throughput per run (1.6 Giga bases of data per run and 60 Mb per hour) and the shortest hands-on time as the amplification step is performed on the sequencer (Junemann et al., 2013; Loman, Misra, et al., 2012). However, selection of the most suitable sequencer heavily depends on the application and specific needs. In clinical context and outbreak investigations, especially high throughput and user convenience are needed at a reasonable price. Besides the technical facts, also subjective preferences play an important role. Exemplary studies comparing different sequencers have shown that results from a single laboratory are neither significantly affected by the sequencing machine nor by the sequencing chemistry (Harris et al., 2013; Kaas, Leekitcharoenphon, Aarestrup, & Lund, 2014). However, detailed and extensive analysis and inter-laboratory evaluation of sequencing practices in use remains to be performed in order to assess minor differences and to establish robustness of results and global comparability.

Global Microbial Identifier (GMI) is an international initiative with the aim of real-time aggregating, sharing, mining and using microbial WGS data (Rindom, 2013). Currently, more than 200 experts from 43 countries are involved. Inclusion of intergovernmental organisations like the World Health Organization (WHO) and the World Organization for Animal Health (OIE) as well as a collaboration with the EU project COMPARE, a multidisciplinary research network establishing a globally linked data information sharing platform system for the control of emerging infectious diseases and foodborne outbreaks (Skiby, 2015), are expected to support international crosstalk and to strengthen the impact of the initiative. One main objective of GMI is the development and realisation of inter-laboratory proficiency testing (PT) to identify steps where quality assurance, control measures or methodological unification are essential to produce standardised high quality sequencing results. In 2014, a pilot PT with only six participants was performed to gain first experience in documentation and practical procedures for this kind of study (Hendriksen et al., 2016). Furthermore, for an optimal adjustment of testing conditions and focus areas prior to a large-scale study, requirements for a general PT among GMI members were interrogated by a survey. Of the 42 respondents, 31% were from the USA, 8.9% and 2.2% from Canada and Australia, respectively, and 51.2% from EU and associated countries (Moran-Gilad et al., 2015). The three most accessible sequencing platforms were MiSeq (23.7%), Ion Torrent PGM (15%) and HiSeq (10.5%), two benchtop and one production-scale sequencer (Table 1). While the benchtop solutions were mostly internally accessible, accessibility to HiSeq was predominantly external. Enquiry of sequencing priorities revealed that foodborne pathogens were the most frequently sequenced pathogens (75%) with high resolution outbreak analysis being the leading application. Among the priority pathogens, *L. monocytogenes* was on the fourth place behind *Escherichia coli, Salmonella* and *Campylobacter* spp. Although the majority of survey respondents agreed that quality filtering and criteria would be important, values specified varied that much that no conclusion could be drawn. For example especially coverage was mentioned as an important quality criterion by 90.9% of

Table 2
Overview of assembly algorithms most frequently used for WGS data.

Assembler	Algorithm	Availability
De novo assembly		
Velvet	De Bruijn Graph	free
Newbler	Overlap/Layout/Consensus	commercial
CLC Genomic	De Bruijn Graph	commercial
SOAPdenovo	De Bruijn Graph	free
Reference-based mapping		
BWA	FM-index	free
Bowtie 2	FM-index	free

respondents, but values ranged from 11-30x (21.6%) over 31-60x (51.3%) to over 60x (18.9%). Apart from the quality criteria, also laboratory methods reported to be used for sample preparation were highly diverse e.g. for DNA or library preparation.

PT based on survey results was performed in 2016 by GMI supported by the U.S. FDA. *Campylobacter coli* and *C. jejuni*, *L. monocytogenes* and *Klebsiella pneumoniae* were selected for analysis. The PT was designed to address three topics: DNA preparation and sequencing procedures, sequencing output, and variant calling of WGS data and cluster analysis. With submission deadline being the 13 January 2017, results remain to be published. Identification of differences and their impact on the sequencing and also analysis results represent an important step towards global harmonisation which would then open the door for international exchange of standardised WGS data.

In general, there are two possible approaches towards global harmonisation of WGS for bacterial typing. The first one is to validate whether different ways are able to yield equivalent results. GMI, for instance, focuses on comparable results among different laboratory practices, sequencing platforms and quality criteria, thereby considering already established local standards. However, this is an organisational and analytical challenge. The second approach is hence the setup of a standard protocol. An exemplary multi-center ring trial was successful in showing that accuracy and reproducibility of NGS based bacterial typing (in this case of *Staphylococcus aureus*) is very high if prescribed methods are applied (Mellmann et al., 2017).

PulseNet USA has developed a Standard Operating Procedure (SOP), PNL32, as a standardised laboratory protocol for WGS of bacterial organisms on the Illumina MiSeq benchtop sequencer (PulseNet, 2015, 2016). PulseNet has been established as a collaboration of CDC, state and local health departments in the USA for real-time comparison of human bacterial pathogens in order to define disease clusters. First initiated for comparison of PFGE profiles, its transition to WGS data is in full progress. With respect to isolates from food and the environment, FDA has launched the GenomeTrakr network. It collaborates with CDC allowing public health authorities to share data from patient and food isolates while investigating foodborne outbreaks and thus ameliorates food safety in the USA. The PulseNet protocol provides standardised and highly detailed methods for DNA isolation and quality control, library preparation and run setup for the sequencer. For example, it stipulates a quality check prior to library preparation where DNA-concentration should at least be 10 ng/µl and meet a 260/280 ratio between 1.75 and 2.05 measured by a Thermo Scientific™ NanoDrop™ spectrophotometer. Furthermore, quality benchmarks for sequence raw data have been specified, more precisely Q-scores and coverage. Q-score has to be Q30 for > 75% of the bases when using a 500 cycle kit and > 85% of bases for a 300 cycle kit and coverage for *L. monocytogenes* needs to be ≥ 20x before upload to PulseNet Central.

Although feasible for laboratories that newly establish WGS, for laboratories that already use WGS, implementation of a standardised protocol could suffer from low compliance as transition would need to be accompanied by investment and change of workflows. Furthermore, as a result of the constant evolution of NGS, continuous adaptation of the standardised procedure is necessary. Newly evolving techniques have to be validated before their inclusion into the SOP which could lead to a delay in their use. For an approach leaving methodical details to individual laboratories and in return defining quality parameters and general thresholds, the use of new techniques would not be a problem as long as final sequence data meet the set criteria. However, such universal quality criteria are very hard to define and are going to require further in-depth analyses and validations until applicable to WGS typing of pathogens in the field of public health and food safety.

4. Dry lab standardisation of NGS methods for foodborne pathogen typing

Through extensive development in NGS technologies, massive amounts of sequence information can be produced within a relatively short period of time. However, bioinformatics tools for the analysis and interpretation of big data struggle to keep pace. In the present time, the bottleneck for integration of NGS-based genome analysis into routine use in disease surveillance is shifting from sequencing to the bio-computational analysis and data storage (Wyres et al., 2014). Currently, no stand-alone tool is able to meet all requirements for a reliable, straightforward and automated analysis of the sequence reads.

For sequencing of bacterial genomes mainly SGS is applied, producing overlapping, short reads. As a result from the high coverage, a high accuracy is achieved. This makes the technology more feasible for variant analyses like SNP detection than the more inaccurate TGS technologies. Still, reads generated by SGS are significantly shorter than those produced by TGS. There are two strategies to deal with this problem: *de novo* assembly of sequence reads to reconstruct a genome or reference-based mapping where single reads are aligned to an already existing, closely related reference genome. For both methods, a variety of different bioinformatics solutions and programs exists (for examples see Table 2).

Algorithms for *de novo* assembly can be grouped into three main categories, all based on graphs: Overlap/Layout/Consensus (OLC)-methods using overlap graphs, de Bruijn Graph (DBG)-methods using k-mer graphs or greedy graph algorithms. Graphs are abstract structures of nodes connected by edges which are used to present relationships. In an overlap graph, the graph represents the sequencing reads (nodes) and their overlaps (edges) of varying length whereas k-mer graphs use subsequences and overlaps of fixed length of k nucleotides (Miller, Koren, & Sutton, 2010). Greedy graphs make use of either the one or the other. Differences lie in the details of graph construction and resulting definition of contiguous sequences of concatenated reads named contigs. Choice of assembler depends on the properties of the sequence data to be used; some assemblers are even specific for a certain sequencing platform. When using DBG assemblers for example, k-mer length has to be adjusted to the read-length while finding a trade-off between sensitivity of smaller and the specificity of longer k-mers (Compeau, Pevzner, & Tesler, 2011). Commercial as well as open-source solutions are available for *de novo* assembly. In a survey among 42 GMI members in 2014, 75% declared using Velvet, freely available software based on DBG and one of the most popular assemblers (Moran-Gilad et al., 2015). Other common assembly software according to the survey was Newbler (46.9%), CLC Genomics (46.9%) and SOAPdenovo (25%). The commercial software Newbler was implemented specifically for 454 GS sequencing platforms and uses OLC whereas the other two programs CLC Genomics (commercial) and SOAPdenovo (freeware), are based on DBG algorithms. It is to note that this survey represents only a relatively small number of software solutions. In practice, a variety of other tools is used as well. Because of the speed of development of software and algorithms, it is difficult to provide a comprehensive and up-to-date ranking. Also changes in the prevalence of use of different sequencing systems influence the popularity of assembly software. For most researchers, commercial software is not a feasible solution. Instead, open-

source software is far more popular. Besides Velvet, SPAdes is a frequently used freeware assembler. In a comparison of seven different assemblers using the quality assessment tool QUAST, it was able to reach the highest amount of mapped genes, the largest N50 value (a measure for the weighted median contig size) and the highest number of complete genes (Gurevich, Saveliev, Vyahhi, & Tesler, 2013). Another example for a freeware assembler is IDBA which in the same study showed its strength in the longest and lowest number of contigs.

Generally, it is difficult to define quality metrics for an assembly as usually the correct answer is not known and assembly errors are very hard to differentiate from biologically relevant SNPs or other genetic changes (Nagarajan & Pop, 2013). Commonly used parameters are hence limited to the assessment of contiguity of the assembly unless an already closed reference genome exists for comparison. Contiguity measures include total size and number of contigs and the N50 value. As a part of the Genome Assembly Gold-standard Evaluations (GAGE) study, eight popular assemblers were compared (Salzberg et al., 2012). One key finding of the study was that the degree of contiguity of an assembly heavily depended on the assembler and the genome to be assembled. Furthermore, quality of the raw data had a considerable impact on the overall quality of the assembly. In general, the correctness of an assembly was found to vary, albeit independently of statistics on contiguity. Consistent with these findings, in another extensive comparison of assemblers, the Assemblathon, significant differences between assemblies from different assembly strategies were revealed (Bradnam et al., 2013; Earl et al., 2011). As a result, it becomes clear that no generic answer can be given to the question of which assembler is the most appropriate for a given dataset. It is rather a case-by-case decision, highly depending on specific requirements.

For detection of variants, in most cases, reference-based mapping as a less computationally intensive method is used. First of all, single sequence reads need to be aligned to a closely related reference genome. There are three categories of alignment algorithms for that purpose: algorithms based on hash tables, based on suffix/prefix tries (i.e. suffix tree, enhanced suffix array or FM-index) or based on merge sorting (Li & Homer, 2010). The latter one is very rarely used, though. According to a survey among GMI members conducted in 2014, most used software solutions were the Burrow Wheeler Aligner (BWA) and Bowtie 2 with 66.7 and 53.3% respectively (Moran-Gilad et al., 2015). Both of them use the FM-index for alignment, the most memory saving and thus most common implementation (Li & Homer, 2010). Other software like e.g. Novoalign, SMALT, MAQ or SHRiMP was only used by 10% of users (Moran-Gilad et al., 2015).

De novo assembly as well as reference-based alignment is only the first step in a series of analytical steps needed for variant detection and in the end clustering of sequences into phylogenetic groups. In general, variants can be called on the basis of SNP detection or gene-by-gene comparison (i.e. cgMLST or wgMLST). Computing efforts in this context are mainly influenced by the decision whether read mapping (less computationally expensive) or a *de novo* assembly (more computationally expensive) is used as starting point for analysis. Although SNP detection provides the highest accuracy, it is often more complex to evaluate. WgMLST and cgMLST serve as a less burdensome alternative providing similar discriminatory power by putting the focus on allelic changes regardless of the number of SNPs involved (Fig. 2). As an extension of the classical MLST which is limited to the analysis of only few housekeeping genes (e.g. 7 for *L. monocytogenes*) (Maiden et al., 1998; Salcedo, Arreaza, Alcala, De La Fuente, & Vazquez, 2003), in cgMLST most genes of the core genome and in wgMLST even the entirety of genes are taken into account. The core genome is defined as the set of genes present in all strains of the same bacterial species whereas the whole genome also comprises accessory genes. One key advantage over SNP detection is that the nomenclature scheme from classical MLST can simply be extended, facilitating consistent classification according to a standardised subtype nomenclature. For L. *monocytogenes*, a core genome scheme has been proposed by Ruppitsch and colleagues in 2015 (Ruppitsch, Pietzka, et al., 2015). Alternatively, a bioinformatics pipeline was designed for cgMLST of *L. monocytogenes*, taking raw sequence reads as input and calculating a core genome profile by comparing it to an expandable database to compile a phylogeny (Pightling, Petronella, & Pagotto, 2015b).

For both, SNP- and allele-based variation detection, different steps and programs have to be combined to receive an informative result. Furthermore, there might be the need for repetitive or consecutive steps within one piece of software. For example a reference-based assembly has to be followed by analysis through a SNP caller to identify mutations. Then, for creation of a phylogenetic tree, another program is usually needed. Likewise, the general analytical process for cgMLST and wgMLST includes assembly, annotation of genes and comparison to a reference. Instead of a user-specific, stepwise analysis, often involving combinations of available programs and custom scripts, the establishment of a generalised and standardised analysis pipeline would be helpful in order to generate a universally valid output.

For isolates of *L. monocytogenes*, different combinations of assembly tools and SNP callers were tested with the result that they varied heavily in the number of true and false positively called SNPs and in accuracy (Pightling, Petronella, & Pagotto, 2015a). Altogether, no general statement on the influence of different parameters could be made as different combinations behaved differently and sometimes even opposed to one another. For example for some combinations, higher coverage led to more true positive identifications of variant sites but in some cases also produced more false positive hits. Besides, read quality trimming and filtering impacted the quality of results either positively or even negatively depending on the software combinations used. This underlines the drastic effect of the variety of methodical implementations on the analysis of NGS data.

Nevertheless, in a proficiency testing for the dry lab part of WGS in 2015 by GMI with more than 40 participants, > 93% of the samples clustered correctly using various analytical approaches (Pettengill et al., 2015). However, number of variants and branch lengths differed considerably indicating that thresholds that led to clustering varied markedly. This shows how difficult it would be to standardise thresholds for different methods. Another study comparing typing capabilities of five different laboratories showed that if methods are prescribed in great detail, a unified output can be reached, underlining the usefulness of a standardised analytical approach (Mellmann et al., 2017). The easiest way for standardisation would probably be the implementation of an analytical pipeline as an aggregation of individually operating segments. This could simplify the transfer of intermediate results between single analytical steps to the level of a single input and output. Current pipelines link individual software pieces for an overarching analysis. Interoperability of components needs to be assured by compatible data formats of program in- and outputs. Furthermore, record-keeping should be included for transparency and reproducibility of

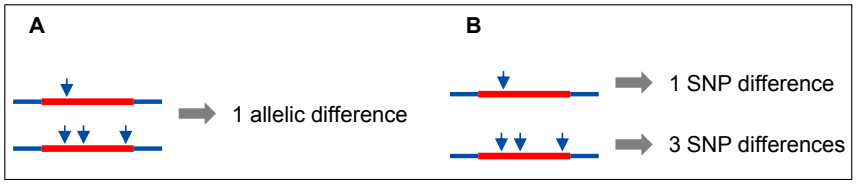

Fig. 2. Detection of differences in cgMLST (A) and SNP analysis (B). In cgMLST analysis, only allelic changes are considered, no matter how many nucleotide changes are found within one allele. In contrast to that, in SNP analysis, every single difference between nucleotide sequences is taken into account.

analysis which is not only crucial for standardisation but ultimately also to secure legal evidence of data (Wyres et al., 2014). In the ideal case, a pipeline should be oriented to the needs of non-specialist, non-bioinformatician users enabling an intuitive and automated analysis.

The American Center for Food Safety and Applied Nutrition (CFSAN), a branch of the FDA, has developed a pipeline for construction of SNP matrices from NGS data (Davis et al., 2015). It combines the following steps: mapping of reads to a reference genome with Bowtie 2, processing of the mapping files with SAMtools, identification of variant sites using VarScan and finally production of a SNP matrix using a custom Python script. The steps are run automatically, converting the input FASTQ data from sequencing reads into a SNP matrix in FASTA format. Still, some points are left to the hands of the user. For example no quality filtering is included. The reason is most likely the number and complexity of possible steps for that purpose and the considerable variations between platforms that make it almost impossible to find a general solution (Edwards & Holt, 2013). Furthermore, the created SNP matrix only serves as an input for the construction of a phylogeny. As a result, despite the robustness and accurateness of the pipeline within the scope of its implementation, it is still no complete solution. Besides the CFSAN pipeline, numerous other, often custom and facility-specific analytical approaches are used and their detailed description would go beyond the scope of this article.

Standardisation for the dry lab analytical part of WGS appears to be even more difficult to achieve than for the wet lab. The easiest way would probably be the establishment of a universal analytical pipeline, in the ideal case directly linked to the sequencer itself and outputting desired results. Although this would hamper flexibility, it would likely be the least complicated and least labour-intensive solution to produce consistent and globally interchangeable data indispensable for molecular surveillance of bacterial infections.

5. Metadata and databases for global sharing of WGS data

On the way to a global sharing of bacterial WGS data, several obstacles remain to be overcome. Besides standardisation of data generation and analysis, international structures and standards for data sharing need to be established.

There are three main databases for the storage of WGS data which together form the International Sequence Database Collaboration (INSDC): the National Center for Biotechnology (NCBI), the European Nucleotide Archive (ENA) and the DNA Data Bank of Japan (DDBJ). GMI as well as the FDA network GenomeTrakr's database use the publicly accessible data layer of parts of the INSDC. GenomeTrakr for example submits data to the NCBI under a single BioProject (Jackson et al., 2016). In Europe, EFSA and ECDC have established a joint database for foodborne pathogens of human and non-human origin. Although not yet used for collection of WGS data but for management of PFGE typing data, it is built to be extended. The ECDC-EFSA database is independent of such public data layers as from the INSDC. It is physically hosted at the TESSy database, the ECDC's database for human strains.

One fundamental issue is the question on the metadata that can be made available together with the WGS data in an open-access database. Metadata is described as “information that is held as a description of stored data” (Dictionary.com), such as isolate specific details like isolation date or source. It is indisputable that such additional information about a bacterial strain provided together with the sequence data greatly increases its utility (Allard et al., 2016). However, national legislation and data protection acts may restrict data sharing and thus limit free data exchange. As a result, there is no consensus about the level of metadata that should be made publicly available. Although different initiatives started collecting bacterial typing data, different concepts of metadata-linkage are proposed. GenomeTrakr includes a minimum set of metadata fields that need to be filled when submitting sequence data to the database. They comprise the collector of the isolate (i.e. the submitting lab), its taxonomic name, sample date and site, the isolation source and sequencing parameters (Allard et al., 2016). GMI expands this list by information on pathogen-associated attributes like specific host or host disease (GMI., 2013). Compared to the GenomeTrakr or the GMI database, for the ECDC-EFSA database, a more restrictive approach is pursued according to EU legislation on data protection. Additional information is restricted to the source of the sample (food, animal, feed, human), typing data and date of sampling whereas for example information related to the origin (country) is considered as potentially sensitive (Rizzi et al., 2017). Identification of the submitting laboratory is also not considered admissible. Furthermore, differentiated access rights for different user groups and stakeholders are incorporated for sensitive data as a compromise between data accessibility and protection. It is to note that the current EU wide Data Protection Basic Regulation including the Data Protection Directive 95/46/EG is about to be changed. However, key principles will remain valid only some aspects have been changed or added. The amendment will become active in May 2018.

Apart from strict legal regulations, there are further reservations regarding free publication of genomic data and metadata (Aarestrup, 2012). As a result of concerns about the ultimate use and possibly the fear of unauthorised application, researchers may not be willing to share their data in public databases before publication. For governments and institutions, competing interests in trade or tourism could be a problem as especially data on foodborne pathogens, the detection of food contamination or even related outbreaks can have far-reaching consequences. Also patenting and intellectual property issues might arise from a free information exchange. Another major challenge is to reconcile protection of confidential patient information and the patients' privacy rights and provision of information needed for epidemiological investigations. These reservations have to be considered when developing a legal framework for public information accessibility. For the EURL *Lm* DB as part of the joint EFSA-ECDC database, compliance with a memorandum of understanding is a prerequisite for participation (Félix, Danan, Makela, et al., 2014). Among others, it regulates data ownership and publication.

An example how well a thought-out and user-oriented database system can work, as among others shown by an almost exponential growth of available genomes, is the Pathosystems Resource Integration Center (PATRIC) (Gillespie et al., 2011; Wattam et al., 2017). PATRIC represents a database coupled with an analysis resource center. Initially designed for the integration of research data and metadata for various pathogens, it now aims to also adapt to the needs of clinical application. Genomic information is linked to metadata including information on organism, isolate, host, sequence, phenotype and project. However, not all fields need to be filled. Although all entries of the database and also information on metadata are publicly available, privacy of data can be maintained by analysing own sequences in a private space without disclosing the information. Still, comparison with public database entries remains possible. On the one hand, this one-way data exchange guarantees protection of possibly sensitive data while allowing reconciliation with already existing data. On the other hand, if all new sequences remain private, progress and timeliness will probably be obstructed. Hence, although technical basis is provided, again a compromise and consensus on the side of the user needs to be found in order to allow for efficient global data sharing.

6. Conclusion

During the last years, WGS has proven its value in the surveillance of *L. monocytogenes* and related outbreak investigations, enabling fast and precise identification of coherent clusters of infection cases, their trace back to food-sources and ultimately elimination of the infection root. So far, several initiatives have been launched to promote WGS based subtyping of *L. monocytogenes* and other foodborne pathogens. However, transition to an international standard remains to be

established.

The One Health strategy makes information exchange between public health, food safety and veterinary authorities indispensable. Hence, a globally accessible sequence database of foodborne pathogens linked with a minimum set of metadata would bring major benefits. The separation of laboratory data and epidemiological or clinical data as well as restricted access to these databases could be an approach to meet data protection criteria. However, more detailed information should be available on demand to effectively protect public health. If an agreement on data format and quality parameters of raw sequence data could be made, direct upload into a centralised analysis pipeline linked to the central database might help to yield standardised and thus comparable sequence information. Still, necessary IT infrastructure has to be established to cope with the problem of data transfer.

The benefits of global data sharing are clear. It helps to provide a comprehensive picture of the appearance and spread of pathogens associated with public health concerns and economic losses around the world. Global data accessibility and exchange is resource-saving as financial burden and workload can be reduced by preventing unnecessary duplication. In addition, it gives the opportunity for a global view and thus improved scientific quality and effective risk management.

Funding

This work was conducted within the project MolTypList which is supported by a grant of the Federal Ministry of Health (GE 2016 03 26), a project in the framework of the German Research Platform for Zoonoses.

Conflicts of interest

None.

References

Aarestrup, F. M. (November 2012). Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerging Infectious Disease Journal-CDC, 18*(11).

Allard, M. W., Strain, E., Melka, D., Bunning, K., Musser, S. M., Brown, E. W., et al. (2016). Practical value of food pathogen traceability through building a whole-genome sequencing network and database. *Journal of Clinical Microbiology, 54*(8), 1975–1983. http://dx.doi.org/10.1128/JCM.00081-16.

Aureli, P., Fiorucci, G. C., Caroli, D., Marchiaro, G., Novara, O., Leone, L., et al. (2000). An outbreak of febrile gastroenteritis associated with corn contaminated by Listeria monocytogenes. *New England Journal of Medicine, 342*(17), 1236–1241. http://dx.doi.org/10.1056/NEJM200004273421702.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., ... Korf, I. F. (2013). Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience, 2*(1), 10. http://dx.doi.org/10.1186/2047-217x-2-10.

Brosch, R., Buchrieser, C., & Rocourt, J. (1991). Subtyping of Listeria monocytogenes serovar 4b by use of low-frequency-cleavage restriction endonucleases and pulsed-field gel electrophoresis. *Research in Microbiology, 142*(6), 667–675. https://doi.org/10.1016/0923-2508(91)90080-T.

CDC (2015a). *Multistate outbreak of listeriosis linked to blue bell creameries products (final update)*. Retrieved from https://www.cdc.gov/listeria/outbreaks/ice-cream-03-15/.

CDC (2015b). *Multistate outbreak of listeriosis linked to commercially produced, prepackaged caramel apples made from bidart bros. Apples (final update)*. Retrieved from https://www.cdc.gov/listeria/outbreaks/caramel-apples-12-14/.

CDC (2015c). *Multistate outbreak of listeriosis linked to soft cheeses distributed by Karoun Dairies, Inc. (final update)*. Retrieved from https://www.cdc.gov/listeria/outbreaks/soft-cheeses-09-15/.

CDC (2016). *The Listeria whole genome sequencing project.* Retrieved from https://www.cdc.gov/listeria/surveillance/whole-genome-sequencing.html.

Chen, Y., Luo, Y., Curry, P., Timme, R., Melka, D., Doyle, M., ... Strain, E. A. (2017). Assessing the genome level diversity of Listeria monocytogenes from contaminated ice cream and environmental samples linked to a listeriosis outbreak in the United States. *PLoS One, 12*(2), e0171389. http://dx.doi.org/10.1371/journal.pone.0171389.

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology, 29*(11), 987–991. http://www.nature.com/nbt/journal/v29/n11/abs/nbt.2023.html#supplementary-information.

Davis, S., Pettengill, J. B., Luo, Y., Payne, J., Shpuntoff, A., Rand, H., et al. (2015). CFSAN SNP Pipeline: An automated method for constructing SNP matrices from next-generation sequence data. *The Perl Journal, 2015*(1), http://dx.doi.org/10.7717/peerj-cs.20.

Deng, X., den Bakker, H. C., & Hendriksen, R. S. (2016). Genomic epidemiology: Whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annual Review of Food Technology, 7*, 353–374. http://dx.doi.org/10.1146/annurev-food-041715-033259.

Dictionary.com. Collins English Dictionary - Complete & Unabridged 10th Edition. Retrieved from http://www.dictionary.com/browse/metadata.

Doumith, M., Buchrieser, C., Glaser, P., Jacquet, C., & Martin, P. (2004). Differentiation of the major Listeria monocytogenes serovars by multiplex PCR. *Journal of Clinical Microbiology, 42*(8), 3819–3822.

Earl, D., Bradnam, K., John, J. S., Darling, A., Lin, D., Fass, J., ... Diekhans, M. (2011). Assemblathon 1: A competitive assessment of de novo short read assembly methods. *Genome Research, 21*(12), 2224–2241.

ECDC (2013). *ECDC roadmap for integration of molecular typing into European level surveillance and epidemic preparedness, Version 1.2.* Retrieved from http://ecdc.europa.eu/en/publications/Publications/molecular-typing-EU-surveillance-epidemic-preparedness-2013.pdf.

Edwards, D. J., & Holt, K. E. (2013). Beginner's guide to comparative bacterial genome analysis using next-generation sequence data. *Microbial Informatics and Experimentation, 3*. http://dx.doi.org/10.1186/2042-5783-3-2 2–2.

EFSA, & ECDC (2016). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2015. *EFSA Journal, 14*(12), e04634. http://dx.doi.org/10.2903/j.efsa.2016.4634 n/a.

Félix, B., Danan, C., Makela, P., Van Walle, I., Lailler, R., Texier, T., ... Roussel, S. (2014a). *Development of a European molecular typing database for food, environmental and veterinary Listeria monocytogenes strains* (euroreference).

Félix, B., Danan, C., Van Walle, I., Lailler, R., Texier, T., Lombard, B., ... Roussel, S. (2014b). Building a molecular Listeria monocytogenes database to centralize and share PFGE typing data from food, environmental and animal strains throughout Europe. *Journal of Microbiological Methods, 104*, 1–8. http://dx.doi.org/10.1016/j.mimet.2014.06.001.

Félix, B., Niskanen, T., Vingadassalon, N., Dao, T. T., Assere, A., Lombard, B., ... Roussel, S. (2013). Pulsed-field gel electrophoresis proficiency testing trials: Toward european harmonization of the typing of food and clinical strains of Listeria monocytogenes. *Foodbourne Pathogens & Disease, 10*(10), 873–881. http://dx.doi.org/10.1089/fpd.2013.1494.

Gilchrist, C. A., Turner, S. D., Riley, M. F., Petri, W. A., & Hewlett, E. L. (2015). Whole-genome sequencing in outbreak analysis. *Clinical Microbiology Reviews, 28*(3), 541–563.

Gillesberg Lassen, S., Ethelberg, S., Bjorkman, J. T., Jensen, T., Sorensen, G., Kvistholm Jensen, A., ... Molbak, K. (2016). Two listeria outbreaks caused by smoked fish consumption-using whole-genome sequencing for outbreak investigations. *Clinical Microbiology and Infections, 22*(7), 620–624. http://dx.doi.org/10.1016/j.cmi.2016.04.017.

Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., ... Mao, C. (2011). PATRIC: The comprehensive bacterial bioinformatics resource with a focus on human pathogenic species. *Infection and Immunity, 79*(11), 4286–4298.

GMI (2013). *Report on the 6th meeting 11-12 september 2013, UC davis*. Sacramento, California. Retrieved from www.globalmicrobialidentifier.org/-/media/Sites/gmi/News-and-events/2013/6th-meeting-2013-report.ashx?la=da.

Graves, L. M., & Swaminathan, B. (2001). PulseNet standardized protocol for subtyping Listeria monocytogenes by macrorestriction and pulsed-field gel electrophoresis. *International Journal of Food Microbiology, 65*(1–2), 55–62. https://doi.org/10.1016/S0168-1605(00)00501-8.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics, 29*(8), 1072–1075.

Harris, S. R., Török, M., Cartwright, E. J., Quail, M. A., Peacock, S. J., & Parkhill, J. (2013). Read and assembly metrics inconsequential for clinical utility of whole-genome sequencing in mapping outbreaks. *Nature Biotechnology, 31*(7), 592–594.

Hendriksen, R. S., Pedersen, S. K., Larsen, M. V., Pedersen, J. N., Lukjancenko, O., Kaas, R. S., ... Sintchenko, V. (2016). *The proficiency test (pilot) report of the global microbial identifier (GMI) initiative, year 2014. Report.*

Hunter, P. R., & Gaston, M. A. (1988). Numerical index of the discriminatory ability of typing systems: An application of Simpson's index of diversity. *Journal of Clinical Microbiology, 26*(11), 2465–2466.

Jackson, B. R., Tarr, C., Strain, E., Jackson, K. A., Conrad, A., Carleton, H., ... Gerner-Smidt, P. (2016). Implementation of nationwide real-time whole-genome sequencing to enhance listeriosis outbreak detection and investigation. *Clinical Infectious Diseases, 63*(3), 380–386. http://dx.doi.org/10.1093/cid/ciw242.

Junemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., ... Harmsen, D. (2013). Updating benchtop sequencing performance comparison. *Nature Biotechnology, 31*(4), 294–296. http://dx.doi.org/10.1038/nbt.2522.

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M., & Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS One, 9*(8), e104984.

Kleta, S., Hammerl, J., Dieckmann, R., Malorny, B., Borowiak, M., Halbedel, S., ... Al Dahouk, S. (2017). Molecular tracing to find source of protracted invasive listeriosis outbreak, Southern Germany, 2012–2016. *Emerging Infectious Disease Journal, 23*(10), 1680. http://dx.doi.org/10.3201/eid2310.161623.

Kvistholm Jensen, A., Nielsen, E. M., Bjorkman, J. T., Jensen, T., Muller, L., Persson, S., ... Ethelberg, S. (2016). Whole-genome sequencing used to investigate a nationwide outbreak of listeriosis caused by ready-to-eat delicatessen meat, Denmark, 2014. *Clinical Infectious Diseases, 63*(1), 64–70. http://dx.doi.org/10.1093/cid/ciw192.

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics, 2*(3), 231–239.

Li, H., & Homer, N. (2010). A survey of sequence alignment algorithms for next-

generation sequencing. *Briefings in Bioinformatics, 11*(5), 473–483.
Liu, D. (2006). Identification, subtyping and virulence determination of Listeria monocytogenes, an important foodborne pathogen. *Journal of Medical Microbiology, 55*(6), 645–659.
Loman, N. J., Constantinidou, C., Chan, J. Z., Halachev, M., Sergeant, M., Penn, C. W., ... Pallen, M. J. (2012a). High-throughput bacterial genome sequencing: An embarrassment of choice, a world of opportunity. *Nature Reviews Microbiology, 10*(9), 599–606.
Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., et al. (2012b). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology, 30*(5), 434–439. http://dx.doi.org/10.1038/nbt.2198.
Maiden, M. C. J., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., ... Spratt, B. G. (1998). Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences, 95*(6), 3140–3145.
Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., ... Chen, Z. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376–380.
Mellmann, A., Andersen, P. S., Bletz, S., Friedrich, A. W., Kohl, T. A., Lilje, B., ... Harmsen, D. (2017). High interlaboratory reproducibility and accuracy of next-generation sequencing-based bacterial genotyping in a ring-trial. *Journal of Clinical Microbiology, JCM* 02242–02216.
Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315–327. http://dx.doi.org/10.1016/j.ygeno.2010.03.001.
Moran-Gilad, J., Sintchenko, V., Pedersen, S. K., Wolfgang, W. J., Pettengill, J., Strain, E., et al. (2015). Proficiency testing for bacterial whole genome sequencing: An end-user survey of current capabilities, requirements and priorities. *BMC Infectious Diseases, 15*(1), http://dx.doi.org/10.1186/s12879-015-0902-3.
Moura, A., Criscuolo, A., Pouseele, H., Maury, M. M., Leclercq, A., Tarr, C., ... Enouf, V. (2016). Whole genome-based population biology and epidemiological surveillance of Listeria monocytogenes. *Nature Microbiology, 2*, 16185.
Nagarajan, N., & Pop, M. (2013). Sequence assembly demystified. *Nature Reviews Genetics, 14*(3), 157–167.
Pettengill, J., Underwood, A., Lukjancenko, O., Strain, E., Karlsmose Pedersen, S., & Hendriksen, R. S. (2015). *2015 GMI PT dry-lab analyses and report*. Retrieved from http://www.globalmicrobialidentifier.org/Workgroups/GMI-Proficiency-Test-Reports.
Pightling, A. W., Petronella, N., & Pagotto, F. (2015a). Choice of reference-guided sequence assembler and SNP caller for analysis of Listeria monocytogenes short-read sequence data greatly influences rates of error. *BMC Research Notes, 8*, 748. http://dx.doi.org/10.1186/s13104-015-1689-4.
Pightling, A. W., Petronella, N., & Pagotto, F. (2015b). The Listeria monocytogenes core-genome sequence typer (LmCGST): A bioinformatic pipeline for molecular characterization with next-generation sequence data. *BMC Microbiology, 15*, 224. http://dx.doi.org/10.1186/s12866-015-0526-1.
PulseNet, C. (2015). *Pulsenet standard operating procedure for illumina miseq data quality control.* Retrieved from https://www.cdc.gov/pulsenet/pdf/pnq07_illumina-miseq-data-qc-508-v1.pdf.
PulseNet, C. (2016). *Laboratory standard operating procedure for pulsenet nextera xt library prep and run setup for the illumina miseq.* Retrieved from https://www.cdc.gov/pulsenet/pdf/pnl32-miseq-nextera-xt.pdf.
Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics, 13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002.
Richterich, P. (1998). Estimation of errors in "raw" DNA sequences: A validation study. *Genome Research, 8*(3), 251–259.
Rindom, S. (2013). About GMI - vision and objectives. Retrieved from http://www.globalmicrobialidentifier.org/about-gmi/vision-and-objectives.
Rizzi, V., Felicio, T. D. S., Felix, B., Gossner, C. M., Jacobs, W., Johansson, K., ... Mooijman, K. (2017). *The ECDC-EFSA molecular typing database for European Union public health protection* (Euroreference).
Ronholm, J., Nasheri, N., Petronella, N., & Pagotto, F. (2016). Navigating microbiological food safety in the era of whole-genome sequencing. *Clinical Microbiology Reviews, 29*(4), 837–857. http://dx.doi.org/10.1128/CMR.00056-16.
Ruppitsch, W., Pietzka, A., Prior, K., Bletz, S., Fernandez, H. L., Allerberger, F., ... Mellmann, A. (2015a). Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of Listeria monocytogenes. *Journal of Clinical Microbiology, 53*(9), 2869–2876. http://dx.doi.org/10.1128/JCM.01193-15.
Ruppitsch, W., Prager, R., Halbedel, S., Hyden, P., Pietzka, A., Huhulescu, S., ... Wilking, H. (2015b). Ongoing outbreak of invasive listeriosis, Germany, 2012 to 2015. *Euro Surveillance, 20*(50), http://dx.doi.org/10.2807/1560-7917.es.2015.20.50.30094.
Salcedo, C., Arreaza, L., Alcala, B., De La Fuente, L., & Vazquez, J. (2003). Development of a multilocus sequence typing method for analysis of Listeria monocytogenes clones. *Journal of Clinical Microbiology, 41*(2), 757–762.
Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., ... Yorke, J. A. (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research, 22*(3), 557–567. http://dx.doi.org/10.1101/gr.131383.111.
Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences, 74*(12), 5463–5467.
Schmid, D., Allerberger, F., Huhulescu, S., Pietzka, A., Amar, C., Kleta, S., ... Mellmann, A. (2014). Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-2013. *Clinical Microbiology and Infections, 20*(5), 431–436. http://dx.doi.org/10.1111/1469-0691.12638.
Schürch, A. C., & Schaik, W. (2017). Challenges and opportunities for whole-genome sequencing–based surveillance of antibiotic resistance. *Annals of the New York Academy of Sciences, 1388*(1), 108–120.
Seeliger, H. P. R., & Höhne, K. (1979). Chapter II serotyping of Listeria monocytogenes and related species. In T. Bergan, & J. R. Norris (Vol. Eds.), *Methods in microbiology: Vol. 13*, (pp. 31–49). Academic Press.
Skiby, J. E. (2015). About COMPARE. Retrieved from http://www.compare-europe.eu/about.
Stasiewicz, M. J., Oliver, H. F., Wiedmann, M., & den Bakker, H. C. (2015). Whole-genome sequencing allows for improved identification of persistent Listeria monocytogenes in food-associated environments. *Applied and Environmental Microbiology, 81*(17), 6024–6037. http://dx.doi.org/10.1128/AEM.01049-15.
Swaminathan, B., & Gerner-Smidt, P. (2007). The epidemiology of human listeriosis. *Microbes and Infection, 9*(10), 1236–1243. https://doi.org/10.1016/j.micinf.2007.05.011.
Vázquez-Boland, J. A., Kuhn, M., Berche, P., Chakraborty, T., Domínguez-Bernal, G., Goebel, W., ... Kreft, J. (2001). Listeria pathogenesis and molecular virulence determinants. *Clinical Microbiology Reviews, 14*(3), 584–640.
Wang, S., Weller, D., Falardeau, J., Strawn, L. K., Mardones, F. O., Adell, A. D., et al. (2016). Food safety trends: From globalization of whole genome sequencing to application of new tools to prevent foodborne diseases. *Trends in Food Science & Technology, 57*(Part A), 188–198. https://doi.org/10.1016/j.tifs.2016.09.016.
Wattam, A. R., Davis, J. J., Assaf, R., Boisvert, S., Brettin, T., Bun, C., ... Stevens, R. L. (2017). Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. *Nucleic Acids Research, 45*(D1), D535–d542. http://dx.doi.org/10.1093/nar/gkw1017.
Wyres, K. L., Conway, T. C., Garg, S., Queiroz, C., Reumann, M., Holt, K., et al. (2014). WGS analysis and interpretation in clinical and public health microbiology laboratories: What are the requirements and how do existing tools compare? *Pathogens, 3*(2), 437–458.