PATHOGEN GENOMICS: EMPOWERING INFECTIOUS DISEASE SURVEILLANCE AND OUTBREAK INVESTIGATIONS

EDITED BY: Marc Jean Struelens and Vitali Sintchenko
PUBLISHED IN: Frontiers in Public Health

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88963-822-2
DOI 10.3389/978-2-88963-822-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: researchtopics@frontiersin.org

PATHOGEN GENOMICS: EMPOWERING INFECTIOUS DISEASE SURVEILLANCE AND OUTBREAK INVESTIGATIONS

Topic Editors:
Marc Jean Struelens, European Centre for Disease Prevention and Control (ECDC), Sweden
Vitali Sintchenko, University of Sydney, Australia

Citation: Struelens, M. J., Sintchenko, V., eds. (2020). Pathogen Genomics: Empowering Infectious Disease Surveillance and Outbreak Investigations. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88963-822-2

Table of Contents

frontiers in Public Health

EDITORIAL
published: 19 May 2020
doi: 10.3389/fpubh.2020.00179

Editorial: Pathogen Genomics: Empowering Infectious Disease Surveillance and Outbreak Investigations

Marc J. Struelens[1]* *and Vitali Sintchenko*[2]

[1] European Centre for Disease Prevention and Control (ECDC), Solna, Sweden, [2] Sydney Medical School and Marie Bashir Institute for Infectious Diseases and Biosecurity, University of Sydney, Sydney, NSW, Australia

Keywords: microbial genomics, whole genome sequencing, epidemiology, antimicrobial resistance (AMR), public health laboratory surveillance

Editorial on the Research Topic

Pathogen Genomics: Empowering Infectious Disease Surveillance and Outbreak Investigations

Comparative microbial genomics analysis by high-throughput whole-genome sequencing (WGS) offers exquisite resolution for epidemiological investigations of infectious disease. This approach has revolutionized outbreak detection and monitoring of transmission dynamics of infectious agents and antimicrobial resistance across humans, animals, and environment. The objective of this Research Topic was to assemble articles on genomic epidemiological approaches to identify sources and mechanisms of transmission of infection and antimicrobial resistance. Leading experts discuss advances in fine-tuning the WGS laboratory workflows and bioinformatics for analyzing viral, bacterial, and protozoan genomic data as well as best available WGS data sharing and vizualization tools for infectious disease surveillance and control. The Research Topic consists of 13 articles on public health applications of comparative genomics of key human and zoonotic pathogens, including Original research, Reviews, Systematic Review and Curriculum, instruction, and pedagogy reports.

The ongoing COVID-19 pandemic underlines the crucial need for open access to advanced viral infectious disease surveillance systems that integrate genomic and epidemiological data on epidemic pathogens in real-time. Theys et al. review cutting-edge phylodynamic analysis tools and visualization solutions for translating these data into information for disease control decisions by public health and health policy professionals. They discuss bioinformatics platform use for temporal and spatial visualization through examples for tracking viral disease dissemination across populations and monitoring viral evolution and adaptation.

Hendriksen et al. describe the value of bacterial genomics for monitoring the global threat of antimicrobial resistance by efficient and rapid identification of genetic determinants of drug resistance. They review the operational characteristics, functionalities, strengths, and limitations of bioinformatics tools and knowledge bases accessible online for these purposes. They offer their expert perspective for standardization of analytical pipelines and database validation for more robust genomic-enhanced antimicrobial resistance surveillance.

Ribeiro Duarte et al. report on their e-learning and crowdsourcing pedagogic experience to explore expectations and opinions as well as educate academic, public health, and food safety professionals on the potential of microbial metagenomics in surveillance of pathogens and antimicrobial resistance. They ran a blended training exercise followed by massive open online interactive course to disseminate metagenomics knowledge and skills to a global audience of research and surveillance experts and gather their opinions. A majority of participants expected a

OPEN ACCESS

Edited and reviewed by:
Nicola Petrosillo, Istituto Nazionale per le Malattie Infettive Lazzaro Spallanzani (IRCCS), Italy

***Correspondence:**
Marc J. Struelens
marc.struelens@ecdc.europa.eu

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *31 March 2020*
Accepted: *22 April 2020*
Published: *19 May 2020*

Citation:
Struelens MJ and Sintchenko V (2020) Editorial: Pathogen Genomics: Empowering Infectious Disease Surveillance and Outbreak Investigations. Front. Public Health 8:179. doi: 10.3389/fpubh.2020.00179

slow transition to metagenomics for surveillance and food safety risk assessment subject to further harmonization of experimental protocols and interpretation of results.

The value of next-generation sequencing technologies for surveillance and study was reviewed for several human pathogens. *Clostridioides (Clostridium) difficile* is an important enteric pathogen in the healthcare setting where it causes both sporadic and epidemic infections with substantial morbi-mortality. It is increasingly frequent as a community-acquired pathogen although infection sources remain elusive. In a mini-review, Janezic and Rupnik summarize the WGS-based typing schemes for *C. difficile* and compare the merits of single nucleodide variant typing and core genome multilocus sequence typing (cgMLST) methods for surveillance and cluster investigations. They highlight how WGS-based studies help elucidating the global population structure of *C. difficile,* mapping the intercontinental spread of epidemic lineages, resolving relapses, and reinfections in recurrent disease and evaluating the effect of disease prevention measures. Knight and Riley review the diverse ecological reservoirs of *C. difficile* from a One-Health perspective. Micro-evolutionary studies are revealing how the open pan-genome of several successful zoonotic lineages that have adapted to different ecological niches. This foster their global spread between food animals and farm environment to humans under the selective pressure of antimicrobial use in livestock and human medicine.

van der Werf and Ködmön report on a systematic review of three international tuberculosis outbreak investigations that were supported by WGS-based typing. WGS data analysis from different sequencing platforms used the SNV mapping approach with diverse bioinformatics tools. WGS analysis was helpful in supporting evidence from epidemiological data for delineating outbreak-related-cases and excluding unrelated ones. Further standardization of WGS methodology and data sharing procedures is desirable for tuberculosis control.

Morris et al. assess the state-of-the-art genome sequencing methods for *Cryptosporidium* species identification and genotyping in the public health setting. Technical hurdles relate to genomic DNA extraction, sequencing depth, and assembly. Biological complexity is challenging due to sexual recombination of the parasite and multiplicity of infection in humans and animals. While WGS is not yet feasible for routine genotyping, the increasing volume of *Cryptosporidium* genome data available from diverse hosts and geographical sources is helping design novel genotyping markers and better understand its population diversity and virulence variation.

Seth-Smith and Egli have examined current evidence on high-resolution typing of *Corynebacterium diphtheriae* using WGS for surveillance of this re-emerged pathogen in low incidence settings and describe international networks supporting this new approach to the control of diphtheria. The authors review the phylogeny of this diverse species and explain how the timing of disease can be inferred from WGS data. They argue that de-centralized sequencing strategies with redundancy in sequencing capacities, followed by data exchange, may be a valuable future option, especially as WGS becomes more available and portable.

Several contributions to this Research Topic have focused the attention on the added value of genomic surveillance for controlling foodborne diseases. Chattaway et al. describe the transformational impact of WGS on reference microbiology practice and public health laboratory surveillance for Salmonella. They document experience from Public Health England and review challenges of implementing WGS as a routine country-wide reference laboratory service. Their experience started in 2014 and led to the radical transformation of public health practice based on the integrated and cross-disciplinary analysis and decision-making. The authors explain how this transformation led to improved accuracy of results, reduced turn-around times of reports and better recognition and monitoring of smaller and geographically dispersed outbreaks of common Salmonella serovars and outbreaks of prolonged duration. They outline the PHE approaches to the bioinformatics pipelines, detection of antimicrobial resistance in foodborne bacteria and integrated analysis of data. The authors also remind us about the essentiality of inter-agency sharing and comparisons of microbiological, epidemiological, and food chain analyses for effective food safety and control.

These messages are reinforced and expanded by Gerner-Smidt et al. from the US Centers of Disease Control and Prevention who presented compelling evidence for the One Health approach to foodborne surveillance. They argue that such an approach takes public health surveillance to the next level as many foodborne outbreaks ultimately originate from animal or environmental sources. The authors illustrate the power of this approach in helping to successfully solve several persistent community outbreaks, including polyclonal listeriosis associated with contaminated ice cream, multidrug resistant Salmonella Heidelberg linked to contaminated chicken, an outbreak caused by six serotypes of Salmonella associated with consumption of an imported herbal supplement, and multidrug resistant Campylobacter linked to contact with puppies sold by a specific pet store chain. Such outbreaks are significantly more complex than typical point source outbreaks and might be difficult to solve using traditional epidemiological approaches. This paper presents examples of how WGS surveillance enables flexible outbreak case definitions and efficient epidemiological traceback.

An insightful report from the Austrian Agency for Health and Food Safety led by Pietzka et al. demonstrates the potential of genome sequencing in identifying sources of outbreaks of listeriosis. The authors utilized a core genome MLST scheme based on 1,701 target genes to type over six thousand isolates of *Listeria monocytogenes* from human and food associated sources. The typing results helped to identify a community outbreak in eastern Austria and trace back the source of the outbreak to one meat-processing company. The whole-genome sequence based typing yielded better accuracy and higher discriminatory power than pulsed-field gel electrophoresis as well as higher laboratory throughput at a lower cost. These findings are of particulars relevance to public health microbiologists as the growing proportion of elderly citizens drives up listeriosis notifications across many countries in the EU and Northern America.

Seth-Smith et al. aim to improve our understanding of whole genome sequencing protocols employed for laboratory

surveillance. The authors evaluated three popular library preparation protocols based on enzymatic fragmentation which are fast and require minimal amounts of genomic DNA. They provided in-depth analysis of WGS results obtained from libraries prepared by Nextera XT (Illumina), Nextera Flex (Illumina), and QIAseq FX (Qiagen) protocols using a set of 12 reference strains representing pathogenic bacteria with different DNA guanine-cytosine (GC) content. The results suggest that Nextera Flex and Qiaseq FX are less sensitive than Nextera XT to variable GC content. Interestingly, more alleles were detected in the cgMLST analysis with these two best library preparation protocols, producing better discrimination of closely related genomes. Furthermore, these protocols achieved a more complete representation of accessory genes and ensured the detection of every antibiotic resistance gene from short read data with coverage of 50 or higher.

We hope both public health professionals and clinicians will find this issue useful for their practice. The editorial team thanks external reviewers for their constructive criticism and hopes that this issue of Frontiers in Public Health will assist healthcare professionals and scientists involved in translational research and genomics-informed public health laboratory surveillance and will be of interest to everyone who is passionate about international efforts to control communicable diseases.

AUTHOR CONTRIBUTIONS

MS and VS jointly drafted and approved the final manuscript.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

SYSTEMATIC REVIEW
published: 17 April 2019
doi: 10.3389/fpubh.2019.00087

Whole-Genome Sequencing as Tool for Investigating International Tuberculosis Outbreaks: A Systematic Review

Marieke J. van der Werf and Csaba Ködmön*

European Centre for Disease Prevention and Control, Stockholm, Sweden

Background: Whole-genome sequencing (WGS) can support the investigation of tuberculosis (TB) outbreaks. The technique has been applied to estimate the timing and directionality of transmission and to exclude cases from an investigation. This review assesses how WGS was applied in international outbreak investigations and discusses the advantages and challenges of the application of WGS.

Methods: Databases were searched for reports on international TB outbreak investigations. Information was extracted on: Why was WGS applied?; How was WGS applied?; Organizational issues; WGS methodology; What was learned/what were the implications of the WGS investigation?; and challenges and lessons learned.

Results: Three studies reporting on international outbreak investigations were identified. Retrospective WGS sequencing was performed in all studies and prospective typing in two to study TB transmission. In one study, WGS data were produced centrally (i.e., in one laboratory) and analysis was done centrally. In two studies, WGS data production was done in a decentralized manner, and analysis was centralized in one laboratory. Three groups of professionals were involved in the international outbreak investigation: public health authorities, laboratory experts, and clinicians. The reported WGS methodology applied differed between the studies in some aspects, e.g., sequencing platform; quality measures, percentage of the reference genome covered, and the mean genomic coverage; analysis, use of a reference genome or *de novo* assembly; and software used for alignment and analysis. In all three studies, in-house scripts were used for variance calling, and the single nucleotide polymorphism (SNP) approach was used for analysis. All outbreak investigation reports stated that WGS refuted suspected transmission events and provided supporting evidence for epidemiological data. Several challenges were reported of which most were not related to WGS. The only challenge related to WGS was the timeframe of getting WGS data if WGS is not routinely performed.

Conclusions: WGS was considered a useful addition in international TB outbreak investigations. Further standardization of the WGS methodology and good structures for international collaboration and coordination are needed to take full advantage of this new technology. Whether the use of WGS results in earlier detection of cases and thus limits transmission still needs to be determined.

Keywords: whole-genome sequencing, outbreak, tuberculosis, multicountry, international, cluster, Europe

OPEN ACCESS

Edited by:
Vitali Sintchenko,
University of Sydney, Australia

Reviewed by:
Digby Warner,
University of Cape Town, South Africa
Zisis Kozlakidis,
International Agency for Research on Cancer (IARC), France

****Correspondence:***
Marieke J. van der Werf
marieke.vanderwerf@ecdc.europa.eu

Specialty section:
This article was submitted to Infectious Diseases-Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *06 February 2019*
Accepted: *01 April 2019*
Published: *17 April 2019*

Citation:
van der Werf MJ and Ködmön C (2019) Whole-Genome Sequencing as Tool for Investigating International Tuberculosis Outbreaks: A Systematic Review. Front. Public Health 7:87. doi: 10.3389/fpubh.2019.00087

INTRODUCTION

Rationale

Tuberculosis (TB) is an infectious disease caused by the *Mycobacterium tuberculosis* complex. It is airborne and transmitted through droplet aerosols containing the bacillus. Globally it is estimated to have caused disease in 10 million people in 2017 and is one of the top 10 causes of death worldwide (1).

Investigation of TB outbreaks in TB high burden countries is often limited to the investigation of household and close contacts, especially children under the age of 5 years (2). TB control in low TB incidence countries aims at stopping TB transmission and thus focusses on investigation of TB outbreaks next to early diagnosis and treatment of TB. With international travel infectious diseases cross borders and cause disease outbreaks affecting people living in different countries. To control international disease outbreaks identification of cases in two or more countries has resulted in international disease outbreak investigations aiming at identifying additional cases and preventing further spread (3, 4). Also for TB, international outbreak investigations have been conducted (5–8).

The International Health Regulations (IHR) oblige countries to notify the World Health Organization of all events which may constitute a public health emergency of international concern within 24 h of assessment (9). In the European Union a similar system was created in 1998, the Early Warning and Response System, which is a tool with restricted access for monitoring public health threats (10). International TB outbreak investigations have started with a notification in EWRS (7). Both the IHR notification system and the EWRS allow for early notification and bring into permanent communication competent public health authorities in countries and others responsible for determining the measures, which may be required to protect public health.

The World Health Organization defines a disease outbreak as the occurrence of disease cases in excess of normal expectancy[1] Before the availability of molecular typing, outbreaks were defined as two or more TB cases with known exposure to each other by sharing enclosed airspace in the same period. Currently, information from molecular typing is added to epidemiological information to confirm linkage between patients.

Molecular typing methods for TB include IS*6110* restriction fragment length polymorphism, spoligotyping and mycobacterial interspersed repetitive units–variable number tandem repeat (MIRU-VNTR). These methods have been applied to outbreak investigations and provided useful additional information for TB control (5, 11–13). Since the complete genome sequence of *M. tuberculosis* was first described in 1998 (14), whole-genome sequencing (WGS) has been added to the toolbox for outbreak investigation. Several studies showed that WGS has a higher discriminatory power and subdivides clusters defined by other genotyping methods (15–17).

Recently WGS has been applied in the investigation of national outbreaks. In several investigations WGS was used to estimate the timing and directionality of transmission within clusters defined by spoligotyping and/or 24-loci MIRU-VNTR (13, 18–21); however, not always successfully (19, 20). In another investigation of an outbreak of extensively drug-resistant TB (XDR-TB) in London, the use of WGS confirmed the link between cases and guided early patient treatment (22). WGS has also helped in excluding cases from an investigation and thus to focus resources on the investigation of cases that were more likely to have been part of the transmission network (23). With WGS becoming more widely applied experience with using WGS for national outbreak investigations will quickly grow (24). WGS has also been applied for the investigation of international cross-border TB outbreaks (6–8). However, the added value of WGS for outbreak investigations remains unclear.

Objectives

- To assess how WGS has been applied in international TB outbreak investigations; and
- To determine the advantages and challenges of the application of WGS in international TB outbreak investigations.

Research Question

Is WGS a useful tool for international TB outbreak investigations, and what are the advantages and challenges?

METHODS

Study Design

In this systematic review, we examined studies reporting on an international *M. tuberculosis* complex outbreak investigation in humans using WGS. We included all study types in all types of populations.

Systematic Review Protocol

The review protocol was registered in PROSPERO, registration number CRD42018107259.

Search Strategy

The search strategies combined the concepts of WGS with surveillance/outbreak and TB and was set up on 13 August 2018 (**Appendix 1**). Controlled vocabulary (i.e., MeSH and Emtree terms) and natural vocabulary (i.e., keywords) in multiple field search combinations were used to represent the concepts in the search strategies. Automatic email updates were set up in all the databases to continue receiving new results from the designed searches. These alerts were monitored until 5 February 2019. Additional supplementary searches have been performed by backward and forward citation chasing of the included references on 4 February 2019. No language or date restrictions were applied.

Data Sources

We searched PubMed, EMBASE, and Scopus.

Eligibility Criteria

Records were eligible for inclusion if they reported on a study in humans, covered *M. tuberculosis* complex, applied WGS, and the

[1]http://www.who.int/environmental_health_emergencies/disease_outbreaks/en/.

outbreak investigation was performed by two or more countries. We included all study types in all types of populations.

Study Selection

Studies were imported into an EndNote X7 database and duplicates were removed. MW and CK independently screened the titles and abstracts to identify potentially eligible studies. The full text of potentially eligible studies was reviewed in duplicate by MW and CK against the eligibility criteria. Discrepancies were resolved by discussion between the reviewers.

Data Extraction

MW extracted data from selected studies using a predefined data extraction form. CK checked the data extraction. Inconsistencies were resolved by discussion. For each study, we extracted the author name, year, and countries involved in the outbreak investigation. Thereafter, we extracted information on: Why was WGS applied?; How was WGS applied?; Organizational issues; WGS methodology; What was learned/what were the implications of the WGS investigation?; and challenges and lessons learned. No formal study quality assessment was performed, as any description of an international outbreak investigation was relevant for our review with the main limitation of studies being that not all areas of interest were described as is reported in the results.

Definitions

We defined an international outbreak investigation as activities undertaken to establish the existence of an outbreak, describe the outbreak, and to identify the source, transmission mechanism, and contributory factors, as a basis for outbreak response involving two or more countries [adjusted from (25)].

Data Analysis

We summarize the extracted information using the themes: reason for WGS; WGS application; WGS methodology; organizational issues; implications of the WGS investigation; and challenges and lessons learned.

Ethics Statement

This review used published data and ethical review was not required.

RESULTS

Study Selection

The search strategy identified 572 unique records (**Figure 1**). Of these four were selected based on title and abstract. Studies were excluded because they did not cover: humans (39 records); *M. tuberculosis* complex (55 studies); WGS (140 studies); outbreak investigation (262); or two or more countries (69 studies). Three records were errata. After the full text assessment, three records fulfilled the eligibility criteria. One record was excluded from further analysis because it was not an outbreak investigation (17).

Synthesized Findings

The included studies reported on outbreak investigations involving European countries and Israel and covered three (6), four (8), and 12 (7) countries. The outbreak investigations included patients diagnosed between 2010–2014, 2015–2016, and 2016–2017, respectively.

Reason for WGS

In all three outbreak reports WGS was applied to study TB transmission. Walker et al. (7) further specified that the aim of the outbreak investigation, including the application of WGS, was to elucidate the origin of the cluster, identify possible locations of transmission, and interrupt further transmission.

WGS Application

In none of the reported international outbreak investigations WGS was used as a routine investigation method for TB in all involved countries. Therefore, studies used specific criteria to select TB cases for whom WGS data needed to be collected. Criteria included a specific spoligotyping and/or 24-loci MIRU-VNTR patterns and/or drug resistance profile. Popovici et al. (8) also used place as a criterion (i.e., connected to a university in Romania). Cases identified in the contact investigation were later added to the WGS investigation. Retrospective WGS was applied in all three studies; two studies also performed prospective WGS of strains identified during the investigation (7, 8).

WGS data were produced centrally in one laboratory and analysis was done centrally in the study by Fiebig et al. (6). In the other two studies WGS data production was done in a decentralized manner and analysis was centralized in one laboratory. In addition to WGS data, all studies also collected epidemiological information including travel information.

Organizational Issues

In the countries included in the outbreak investigations, three groups of professionals were involved: national and local public health authorities (6–8); experts from (national reference) laboratories (6–8); and clinicians (7).

None of the studies reported on issues related to shipping of strains, e.g., method, costs, and duration. The two studies that used prospective WGS sequencing in addition to retrospective sequencing information did not report on the time till WGS results were available (7, 8). Popovici et al. (8) reported, that it takes several months before the results are available if WGS is not routinely available in the country. None of the studies in which WGS information was exchanged between countries reported on the methods or tools used for WGS data exchange.

WGS Methodology

In the outbreak investigation reported by Fiebig et al. (6) WGS was performed in one laboratory, whereas in the other two outbreak investigations WGS was performed in several laboratories. All but the laboratory of the Public Health Agency of Sweden used an Illumina sequencing platform (6–8), **Table 1**. In two outbreak investigations (6, 7) reads were mapped to the *M. tuberculosis* H37Rv reference genome, whereas in Popovici et al. (8) *de novo* assembly was used in addition to mapping against a reference genome (unspecified). Fiebig et al. (6) specified the percentage of the reference genome covered and the mean genomic coverage, i.e., at least 45 times. The percentage of the reference genome covered was not reported by the other two

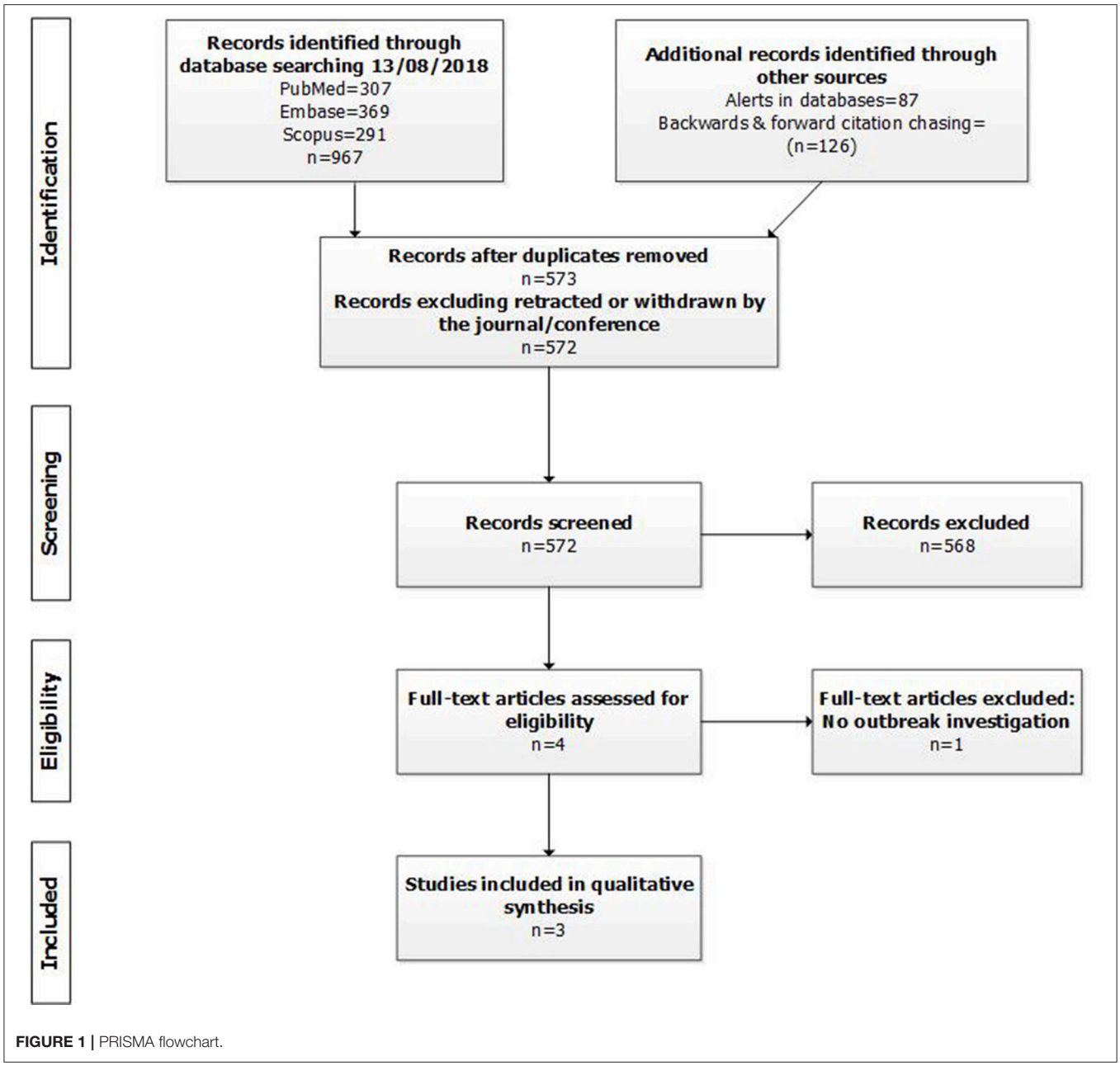


FIGURE 1 | PRISMA flowchart.

studies where WGS was performed in different laboratories. Walker et al. (7) aimed for a mean coverage of 20–50 times. Programs used for alignment and analysis were similar in the studies of Fiebig et al. (6) and Walker et al. (7). In all three studies, in-house scripts were used for variance calling and the analysis used a single nucleotide polymorphism (SNP) approach. The maximum difference in SNPs to define a cluster was not reported in Popovici et al. (8), and was five SNPs in Walker et al. (7) and 12 SNPs in Fiebig et al. (6).

Implications of WGS

According to all three studies, WGS provided useful information for the outbreak investigation (**Table 2**). The outbreak investigation studies reported that WGS refuted suspected transmission events based on epidemiological or MIRU-VNTR information and thus focussed the investigation. WGS also provided supporting evidence for epidemiological data. Walker et al. (7) reported that WGS helped in identifying the direction of transmission and in identifying additional links/missing cases. Furthermore, it provided information about the origin of the strain and where transmission is likely ongoing.

None of the studies provided evidence that WGS was essential for successful control of the outbreak. Also, no changes in TB prevention and control practices or in TB laboratory and surveillance were reported.

TABLE 1 | Whole-genome sequencing methodology applied to international tuberculosis outbreak investigations.

Whole-genome sequencing methodology	Fiebig (6)	Walker (7)	Popovici (8)
Sequencing platform	Illumina	Illumina and Ion Torrent (Sweden)	Illumina and Ion Torrent (Sweden)
Reference genome	Mapping against *M. tuberculosis* reference strain H37Rv	Mapping against *M. tuberculosis* reference strain H37Rv	Mapping against unspecified reference genome and *de novo* genome assembly
% of the reference genome covered	>99% of reference genome	Not reported	Not reported
Coverage depth	At least 45 times	Mean 20–50 times	Not reported
Programs used for alignment and analysis	SARUMAN exact alignment tool In-house Perl scrips for variance calling Bionumerics software (Applied Maths NV, Belgium)	Burrows-Wheeler Aligner version 0.7.12-r1039; Genome Analysis Toolkit; SAMtools Custom Perl scrips for variance calling PhylML 3.1 and Bionumerics 6.7	CLC Assembly Cell v 4.4.2 In-house script
Analysis approach	SNP mapping	SNP mapping	SNP mapping
Maximum SNP or allelic difference thresholds to define cluster	12 SNPs	5 SNPs	Not reported

SNP, single nucleotide polymorphism.

TABLE 2 | Implications of whole-genome sequencing in international tuberculosis outbreak investigations.

Implications of WGS investigation	Fiebig (6)	Walker (7)	Popovici (8)
Guiding contact investigation	No*	No	No
Identification of possible direction of transmission	No	Yes	Not reported
Identification of additional links or missing cases	No	Yes	Not reported
Identification of places of transmission	No	Yes	Not reported
Refuting suspected transmission based on epidemiological or MIRU-VNTR information	Yes	Yes	Yes
Supporting evidence for information from epidemiological data	Yes	Yes	Yes
Successful control of the outbreak	Not reported	Not reported	Not reported
Changes in TB prevention and control practices or TB laboratory and surveillance systems	Not reported	Not reported	Not reported

MIRU-VNTR, Mycobacterial interspersed repetitive units-variable number tandem repeat; TB, Tuberculosis; WGS, Whole-genome sequencing.
**Contact investigation was completed before initiation of the international outbreak investigation.*

Challenges and Lessons Learned

Several challenges were reported in the three international TB outbreak investigations of which most were not related to WGS. First, the collected information on epidemiological links was difficult to interpret and it was often not known whether absence of a link meant that there was indeed no link, or that it was unknown or not reported (6). Also, a challenge in transferring patient reports was noted (6). Collection of travel information is often not a routine component of an outbreak investigation, and it was reported as challenging (7). The only challenge specifically related to WGS was the timeframe of getting WGS data, if WGS is not routinely performed for tuberculosis strains in the country. Experience from the outbreak investigation reported by Popovici et al. (8) showed that it took several months to get the WGS results and to have a link confirmed.

The main lesson learned in all three outbreak investigations was the importance of establishing collaboration and coordination between institutions in different countries involved in the investigation. This also needs a secure system for the exchange of patient data among the involved countries.

DISCUSSION

Summary of Main Findings

Three studies reporting on international TB outbreak investigation using WGS were identified. The WGS methodology used for the outbreak investigation, i.e., sequencing platform, quality indicators such as genomic coverage, and scripts for variance calling, differed to some extent. In addition, the maximum difference in SNPs to define a cluster was different in the two studies that reported SNPs thresholds. WGS was a useful tool for international TB outbreak investigations according to the three studies. However, none of the studies provided evidence that WGS was essential for successful control of the outbreak or provided evidence on the cost-effectiveness of WGS for international outbreak investigations.

Reason for WGS

By applying WGS in international outbreak investigations researchers and experts hoped to obtain additional information on transmission that would help in controlling the outbreak. WGS can provide more information than any of the other typing

methods used for studying TB transmission since it has a higher resolution. It has been shown that WGS can divide clusters identified by other methods into sub-clusters (16, 17), and can identify transmission missed by conventional epidemiological investigations (26). Furthermore, WGS can provide supporting evidence, complementary to temporal- and contact tracing data, to identify the most likely direction of transmission (27, 28).

WGS Application

WGS was not a standard typing method in all countries involved in the international outbreak investigations (6–8). Thus, WGS information was not readily available for all cases and WGS had to be done specifically for strains suspected to be part of the outbreak. This required a decision on the type of cases for which WGS information was to be collected. Restricting the collection of WGS information to specified cases introduces a risk of missing transmission events. This risk might be relatively low if the cases are selected based on a specific MIRU-VNTR pattern. A population based study from the Netherlands reported 86% concordance between MIRU-VNTR and WGS, although the percentages of cases clustered by MIRU-VNTR was almost twice as high (25% by MIRU-VNTR vs. 14% by WGS). In addition, clustering was only shown by WGS and not by MIRU-VNTR for 8 of 76 isolates included in the WGS cluster (29). These potential transmission events would thus have been missed, if the cluster definition was based on MIRU-VNTR pattern only.

Recently, the number of laboratories able to perform WGS has increased rapidly in the European Union (30), providing more countries access to WGS and facilitating WGS of all identified TB cases. Therefore, application of criteria for selecting cases for WGS and thus potentially excluding cases may not be needed anymore in the near future.

Organizational Issues

International TB outbreak investigations require the involvement of different types of stakeholders. All three outbreak investigation studies described the involvement of public health authorities and laboratory experts. Depending on how the (public) health system in countries is organized an international outbreak investigation will need the involvement of both national and local level public health authorities.

All three studies concluded that collaboration and coordination between all institutions involved in the investigation is essential. Our organization, the European Centre for Disease Prevention and Control (ECDC) was involved in the coordination of two of the international outbreak investigations (7, 8). Given the mandate of ECDC, i.e., supporting the response to public health threats in the European Union (31), this supra national organization can play a role in the coordination of outbreak investigations next to other organizations such as the World Health Organization. To ensure effective and efficient international collaboration, mechanisms for collaboration and communication should be further developed.

To be able to rapidly and efficiently investigate potential international TB outbreaks mechanisms for exchange of samples and/or data (including patient data) should be in place. None of the included studies reported on mechanisms for sample or data exchange. Given that patient information would need to be exchanged these exchanges need to be done in a secure way ensuring that data protection and privacy regulations such as the European Union regulation 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (32) are adhered to. Currently, data can be exchanged among European Union countries in a secured way through the Early Warning and Response System (33, 34). Communication between countries can also be done in the framework of the International Health Regulation (9).

WGS Methodology

In two studies (7, 8) WGS data production was done in several laboratories. Since WGS is not standardized for TB (35) this entails a risk that WGS data produced by different laboratories are not 100% comparable. All laboratories will start with genomic DNA from *M. tuberculosis* but may use different protocols for library construction prior to sequencing and library preparation methodology has been shown to play an important role in WGS data quality and may thus influence the results (36, 37). Also, data analysis and interpretation has not been standardized.

Some WGS quality control indicators have been proposed and used (38). These include assessment of the quality of genomic DNA, average depth of genome coverage, and percent of reference genome covered. Fiebig et al. (6) reported on quality targets for percent of the reference genome covered and coverage depth, whereas the two studies that had WGS performed in different laboratories did not report on specific quality targets (7, 8). To ensure comparability of data generated by different laboratories to enable the investigation of outbreaks that go beyond the coverage area of one laboratory there is a need for minimal set of quality standards. The EU wide project EUSeqMyTB (35) will develop the minimal set of standards for WGS methodology to be used in routine European Union level TB molecular surveillance activities.

To ensure that results from laboratory tests are of high quality, reliable, and comparable, external quality assessment is used. Within the European Union, the European Reference Laboratory Network for TB organizes external quality assessment for TB diagnosis and resistance testing and for MIRU-VNTR (39, 40). Recently, the Network established an external quality assessment scheme for WGS. A first pilot was performed in 2015 using five samples with known mutations in genes associated with drug resistance. Participating laboratories were asked to report all the mutations detected in these genes and the results were compared to the results of the reference laboratory. In this first pilot study, most laboratories missed a number of mutations that had been identified by the reference laboratory and found a variety of additional mutations not found by the reference laboratory. In the second WGS external quality assessment round in 2016 participating laboratories were asked to report the WGS data they felt important. The results showed that reporting of mutations at specified loci identified as significantly associated with drug resistance was highly diverse by the participating laboratories. In 2017, participants were asked to identify any mutations strongly associated with drug resistance and to report their position in the respective gene. In addition, laboratories were asked to identify DNA specimens they considered either identical or genetically

closely related. This round was also the first where the WGS external quality assessment results were scored and certificates issued. The external quality assessment scheme for TB WGS developed by the European Reference Laboratory Network for TB seems to be one of the first attempts for assessing the quality of WGS for a specific pathogen although the need for external quality assessment or proficiency testing for WGS of pathogens has been identified earlier (41, 42). In general, only few experiences with external quality assessment schemes for WGS have been published (43, 44).

In the framework of the EUSeqMyTB pilot project (35) a comparison of different WGS analysis pipelines was undertaken using fastq files from a well-defined set of isolates. This analysis showed that some pipelines identify more SNPs then others. The main question that needs to be answered is whether different analysis pipelines result in different conclusions about transmission and relatedness.

In the identified international outbreak investigations, analysis of WGS data was performed centrally in one of the participating laboratories using a SNP-based in-house analysis pipeline. The use of in-house analysis pipelines prohibits easy comparison of results between studies. An alternative approach would be the use of a common nomenclature based on a standardized allele numbering system, which would facilitate exchange of information. For TB a core genome multilocus sequence typing (cgMLST) has been proposed and a web-based nomenclature server is available (45).

Implications of WGS

All three studies reported on advantages of using WGS in international outbreak investigations. Applying WGS in national outbreak investigations has also shown benefits. In the UK, it was shown that using WGS in a multidrug-resistant TB outbreak investigation allowed to exclude one-third of cases from the investigation (23). Resources could thus be focussed. Use of WGS has allowed verification of clusters, i.e., it confirmed that cases were part of a single transmission chain (19) and it identified missed transmission events (26). Reports on national outbreak investigations have also shown that WGS can indicate the direction of transmission (21, 28, 46). However, this does not seem to be the case in all settings (20).

Challenges and Lessons Learned

The main challenges reported in the three international outbreak investigations (6–8) were related to the collection and exchange of epidemiological and travel information. Above we discuss solutions for these challenges. If WGS was not routinely performed, it did take considerable time to get the WGS results since re-culturing of samples was required (8). The increase in the number of countries that have access to WGS may result in routine performance of WGS on all TB samples and thus timely availability of the information (30).

LIMITATIONS

We aimed to collect and abstract information from studies reporting on international outbreak investigations using WGS on: Why was WGS applied?; How was WGS applied?; Organizational issues; WGS methodology; What was learned/what were the implications of the WGS investigation?; and challenges and lessons learned. We searched PubMed, EMBASE, and Scopus but did not search for studies reporting on international outbreak investigations using WGS in the gray literature. We therefore might have missed studies reporting on the application of WGS in international TB outbreak investigations and thus not have identified all information on Why was WGS applied?; How was WGS applied?; Organizational issues; WGS methodology; What was learned/what were the implications of the WGS investigation?; and challenges and lessons learned.

We used a detailed data collection tool. Not all information was reported by the included studies (**Tables 1**, **2**). This may bias our analysis. Furthermore, only three studies were identified that reported on the use of WGS for international outbreak investigations. Since outbreak investigations are a routine activity for public health experts, more international outbreak investigations may have used WGS without being published in (scientific) reports. Thus, some important experiences may have been missed in this analysis.

CONCLUSIONS

WGS seems to be a promising tool for international outbreak investigations. WGS methodology needs to be standardized further, especially quality control, analysis, and interpretation, to better support cross-border collaboration in outbreak investigations. It also allows for the prediction of drug resistance and therefore testing practices have already changed in some countries (24), making WGS information also available for outbreak investigations. However, the advantages for public health still need to be determined. More specifically, does the use of WGS result in earlier detection of cases, which belong to the same transmission chain, and thus limit transmission and result in smaller clusters?

AUTHOR CONTRIBUTIONS

MW designed the systematic review, performed the systematic review, analyzed the results, and wrote the manuscript. CK performed the systematic review and critically reviewed the manuscript.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support received from the ECDC Library in defining the search strategy and performing the search. We also acknowledge the critical review by Marc Struelens.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2019.00087/full#supplementary-material

REFERENCES

1. World Health Organization. *Global Tuberculosis Report 2018*. Geneva: World Health Organization (2018). Available online at: http://apps.who.int/iris/bitstream/10665/137094/1/9789241564809_eng.pdf?ua=1
2. World Health Organization. *Recommendations for Investigating Contacts of Persons With Infectious Tuberculosis in Low- and Middle-Income Countries*. Geneva: World Health Organization (2012).
3. Nic Lochlainn L, Mandal S, de Sousa R, Paranthaman K, van Binnendijk R, Ramsay M, et al. A unique measles B3 cluster in the United Kingdom and the Netherlands linked to air travel and transit at a large international airport, February to April 2014. *Euro Surveill.* (2016) 21:30177–85. doi: 10.2807/1560-7917.ES.2016.21.13.30177
4. Abbara A, Brooks T, Taylor GP, Nolan M, Donaldson H, Manikon M, et al. Lessons for control of heroin-associated anthrax in Europe from 2009-2010 outbreak case studies, London, UK. *Emerg Infect Dis.* (2014) 20:1115–22. doi: 10.3201/eid2007.131764
5. Kiers A, Drost AP, van Soolingen D, Veen J. Use of DNA fingerprinting in international source case finding during a large outbreak of tuberculosis in The Netherlands. *Int J Tuberc Lung Dis.* (1997) 1:239–45.
6. Fiebig L, Kohl TA, Popovici O, Muhlenfeld M, Indra A, Homorodean D, et al. A joint cross-border investigation of a cluster of multidrug-resistant tuberculosis in Austria, Romania and Germany in 2014 using classic, genotyping and whole genome sequencing methods: lessons learnt. *Euro Surveill.* (2017) 22:30439. doi: 10.2807/1560-7917.ES.2017.22.2.30439
7. Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, et al. A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *Lancet Infect Dis.* (2018) 18:431–40. doi: 10.1016/S1473-3099(18)30004-5
8. Popovici O, Monk P, Chemtob D, Chiotan D, Freidlin PJ, Groenheit R, et al. Cross-border outbreak of extensively drug-resistant tuberculosis linked to a university in Romania. *Epidemiol Infect.* (2018) 146:824–31. doi: 10.1017/S095026881800047X
9. World Health Organization. *International Health Regulations* 2005, 3rd ed. Geneva: World Health Organization (2016).
10. Decision no 2119/98/EC of the European Parliament and of the Council of 24 September 1998 setting up a network for the epidemiological surveillance and control of communicable diseases in the Community. Official Journal of the European Communities;3 October 1998.
11. McElroy PD, Southwick KL, Fortenberry ER, Levine EC, Diem LA, Woodley CL, et al. Outbreak of tuberculosis among homeless persons coinfected with human immunodeficiency virus. *Clin Infect Dis.* (2003) 36:1305–12. doi: 10.1086/374836
12. Ma MJ, Yang Y, Wang HB, Zhu YF, Fang LQ, An XP, et al. Transmissibility of tuberculosis among school contacts: an outbreak investigation in a boarding middle school, China. *Infect Genet Evol.* (2015) 32:148–55. doi: 10.1016/j.meegid.2015.03.001
13. Black AT, Hamblion EL, Buttivant H, Anderson SR, Stone M, Casali N, et al. Tracking and responding to an outbreak of tuberculosis using MIRU-VNTR genotyping and whole genome sequencing as epidemiological tools. *J Public Health.* (2018) 40:e66–73. doi: 10.1093/pubmed/fdx075
14. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature.* (1998) 393:537–44. doi: 10.1038/31159
15. Nikolayevskyy V, Kranzer K, Niemann S, Drobniewski F. Whole genome sequencing of Mycobacterium tuberculosis for detection of recent transmission and tracing outbreaks: a systematic review. *Tuberculosis.* (2016) 98:77–85. doi: 10.1016/j.tube.2016.02.009
16. Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, et al. Standard genotyping overestimates transmission of *Mycobacterium tuberculosis* among immigrants in a low-incidence country. *J Clin Microbiol.* (2016) 54:1862–70. doi: 10.1128/JCM.00126-16
17. Jajou R, de Neeling A, Rasmussen EM, Norman A, Mulder A, van Hunen R, et al. A predominant variable-number tandem-repeat cluster of *Mycobacterium tuberculosis* isolates among asylum seekers in the Netherlands and Denmark, deciphered by whole-genome sequencing. *J Clin Microbiol.* (2018) 56:e01100-17. doi: 10.1128/JCM.01100-17
18. Seraphin MN, Didelot X, Nolan DJ, May JR, Khan MSR, Murray ER, et al. Genomic investigation of a *Mycobacterium tuberculosis* outbreak involving prison and community cases in Florida, United States. *Am J Trop Med Hyg.* (2018) 99:867–74. doi: 10.4269/ajtmh.17-0700
19. Norheim G, Seterelv S, Arnesen TM, Mengshoel AT, Tonjum T, Ronning JO, et al. Tuberculosis outbreak in an educational institution in Norway. *J Clin Microbiol.* (2017) 55:1327–33. doi: 10.1128/JCM.01152-16
20. Casali N, Broda A, Harris SR, Parkhill J, Brown T, Drobniewski F. Whole genome sequence analysis of a large isoniazid-resistant tuberculosis outbreak in London: a retrospective observational study. *PLoS Med.* (2016) 13:e1002137. doi: 10.1371/journal.pmed.1002137
21. Outhred AC, Holmes N, Sadsad R, Martinez E, Jelfs P, Hill-Cawthorne GA, et al. Identifying likely transmission pathways within a 10-year community outbreak of tuberculosis by high-depth whole genome sequencing. *PLoS ONE.* (2016) 11:e0150550. doi: 10.1371/journal.pone.0150550
22. Arnold A, Witney AA, Vergnano S, Roche A, Cosgrove CA, Houston A, et al. XDR-TB transmission in London: case management and contact tracing investigation assisted by early whole genome sequencing. *J Infect.* (2016) 73:210–8. doi: 10.1016/j.jinf.2016.04.037
23. Lalor MK, Casali N, Walker TM, Anderson LF, Davidson JA, Ratna N, et al. The use of whole-genome sequencing in cluster investigation of a multidrug-resistant tuberculosis outbreak. *Eur Respir J.* (2018) 51:1702313. doi: 10.1183/13993003.02313-2017
24. Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, Crook DW. Tuberculosis is changing. *Lancet Infect Dis.* (2017) 17:359–61. doi: 10.1016/S1473-3099(17)30123-8
25. *Guidelines for the Investigation and Control of Disease Outbreaks. Porirua, New Zealand: Institute of Environmental Science & Research Limited*. (2002) updated 2012.
26. Torok ME, Reuter S, Bryant J, Koser CU, Stinchcombe SV, Nazareth B, et al. Rapid whole-genome sequencing for investigation of a suspected tuberculosis outbreak. *J Clin Microbiol.* (2013) 51:611–4. doi: 10.1128/JCM.02279-12
27. Schurch AC, Kremer K, Daviena O, Kiers A, Boeree MJ, Siezen RJ, et al. High-resolution typing by integration of genome sequencing data in a large tuberculosis cluster. *J Clin Microbiol.* (2010) 48:3403–6. doi: 10.1128/JCM.00370-10
28. Hatherell HA, Colijn C, Stagg HR, Jackson C, Winter JR, Abubakar I. Interpreting whole genome sequencing for investigating tuberculosis transmission: a systematic review. *BMC Med.* (2016) 14:21. doi: 10.1186/s12916-016-0566-x
29. Jajou R, de Neeling A, van Hunen R, de Vries G, Schimmel H, Mulder A, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: a population-based study. *PLoS ONE.* (2018) 13:e0195413. doi: 10.1371/journal.pone.0195413
30. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, ECDC National Microbiology Focal Points and Experts Group. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European National Capacities, 2015-2016. *Front Public Health.* (2017) 5:347. doi: 10.3389/fpubh.2017.00347
31. *Regulation (EC) No 851/2004 of the European Parliament and Council of 21 April 2004 Establishing a European Centre for Disease Prevention and Control*. Official Journal of the European Union (2004) L142:1–11.
32. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation)*. Official Journal of the European Union (2016) 59:1–88.
33. Cox A, Guglielmetti P, Coulombier D. Assessing the impact of the 2009 H1N1 influenza pandemic on reporting of other threats through the Early Warning and Response System. *Euro Surveill.* (2009) 14:19397. doi: 10.2807/ese.14.45.19397-en
34. Guglielmetti P, Coulombier D, Thinus G, Van Loock F, Schreck S. The early warning and response system for communicable diseases in the EU: an overview from 1999 to 2005. *Euro Surveill.* (2006) 11:215–20. doi: 10.2807/esm.11.12.00666-en
35. Tagliani E, Cirillo DM, Kodmon C, van der Werf MJ, Consortium EU. EUSeqMyTB to set standards and build capacity for whole genome

sequencing for tuberculosis in the EU. *Lancet Infect Dis.* (2018) 18:377. doi: 10.1016/S1473-3099(18)30132-4
36. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, et al. Comparison of sample preparation methods used for the next-generation sequencing of *Mycobacterium tuberculosis*. *PLoS ONE*. (2016) 11:e0148676. doi: 10.1371/journal.pone.0148676
37. Jones MB, Highlander SK, Anderson EL, Li W, Dayrit M, Klitgord N, et al. Library preparation methodology can influence genomic and functional predictions in human microbiome research. *Proc Natl Acad Sci USA*. (2015) 112:14024–9. doi: 10.1073/pnas.1519288112
38. Ezewudo M, Borens A, Chiner-Oms A, Miotto P, Chindelevitch L, Starks AM, et al. Integrating standardized whole genome sequence analysis with a global Mycobacterium tuberculosis antibiotic resistance knowledgebase. *Sci Rep.* (2018) 8:15382. doi: 10.1038/s41598-018-33731-1
39. Nikolayevskyy V, Hillemann D, Richter E, Ahmed N, van der Werf MJ, Kodmon C, et al. External quality assessment for tuberculosis diagnosis and drug resistance in the European Union: a five year multicentre implementation study. *PLoS ONE.* (2016) 11:e0152926. doi: 10.1371/journal.pone.0152926
40. de Beer JL, Kodmon C, van Ingen J, Supply P, van Soolingen D, Global Network for Molecular Surveillance of T. Second worldwide proficiency study on variable number of tandem repeats typing of Mycobacterium tuberculosis complex. *Int J Tuberc Lung Dis.* (2014) 18:594–600. doi: 10.5588/ijtld.13.0531
41. Moran-Gilad J, Sintchenko V, Pedersen SK, Wolfgang WJ, Pettengill J, Strain E, et al. Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect Dis.* (2015) 15:174. doi: 10.1186/s12879-015-0902-3
42. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J Clin Microbiol.* (2016) 54:2857–65. doi: 10.1128/JCM.00949-16
43. Zhang R, Ding J, Han Y, Yi L, Xie J, Yang X, et al. The reliable assurance of detecting somatic mutations in cancer-related genes by next-generation sequencing: the results of external quality assessment in China. *Oncotarget.* (2016) 7:58500–15. doi: 10.18632/oncotarget.11306
44. Dubbink HJ, Deans ZC, Tops BB, van Kemenade FJ, Koljenovic S, van Krieken HJ, et al. Next generation diagnostic molecular pathology: critical appraisal of quality assurance in Europe. *Mol Oncol.* (2014) 8:830–9. doi: 10.1016/j.molonc.2014.03.004
45. Kohl TA, Harmsen D, Rothganger J, Walker T, Diel R, Niemann S. Harmonized genome wide typing of tubercle bacilli using a web-based gene-by-gene nomenclature system. *EBioMedicine.* (2018) 34:131–8. doi: 10.1016/j.ebiom.2018.07.030
46. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. *Lancet Infect Dis.* (2013) 13:137–46. doi: 10.1016/S1473-3099(12)70277-3

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

REVIEW
published: 04 June 2019
doi: 10.3389/fpubh.2019.00139

Whole Genome Sequencing Based Surveillance of *L. monocytogenes* for Early Detection and Investigations of Listeriosis Outbreaks

Ariane Pietzka[1], Franz Allerberger[2], Andrea Murer[1], Anna Lennkh[1], Anna Stöger[2], Adriana Cabal Rosel[2,3], Steliana Huhulescu[2], Sabine Maritschnik[2], Burkhard Springer[1], Sarah Lepuschitz[2], Werner Ruppitsch[2] and Daniela Schmid[2]*

[1] AGES - Austrian Agency for Health and Food Safety, Graz, Austria, [2] AGES - Austrian Agency for Health and Food Safety, Vienna, Austria, [3] European Public Health Microbiology training programme (EUPHEM), European Centre for Disease Prevention and Control (ECDC), Stockholm, Sweden

OPEN ACCESS

Edited by:
Vitali Sintchenko, University of Sydney, Australia

Reviewed by:
Alexandre Leclercq, Institut Pasteur, France
Qinning Wang, New South Wales Health Pathology, Australia

***Correspondence:**
Ariane Pietzka
ariane.pietzka@ages.at

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *01 February 2019*
Accepted: *16 May 2019*
Published: *04 June 2019*

Citation:
Pietzka A, Allerberger F, Murer A, Lennkh A, Stöger A, Cabal Rosel A, Huhulescu S, Maritschnik S, Springer B, Lepuschitz S, Ruppitsch W and Schmid D (2019) Whole Genome Sequencing Based Surveillance of L. monocytogenes for Early Detection and Investigations of Listeriosis Outbreaks. Front. Public Health 7:139. doi: 10.3389/fpubh.2019.00139

In Austria, all laboratories are legally obligated to forward human and food/environmental *L. monocytogenes* isolates to the National Reference Laboratory/Center (NRL) for *Listeria*. Two invasive human isolates of *L. monocytogenes* serotype 1/2a of the same pulsed-field gel electrophoresis (PFGE) pattern, previously unknown in Austria, were cultured for the first time in January 2016. Five further human isolates, obtained from patients with invasive listeriosis between April 2016 and September 2017, showed this PFGE pattern. In Austria the NRL started to use whole-genome sequencing (WGS) based typing in 2016, using a core genome MLST (cgMLST) scheme developed by Ruppitsch et al. 2015, which contains 1701 target genes. Sequence data are submitted to a publicly available nomenclature server (Ridom GmbH, Münster, Germany) for allocation of the core genome complex type (CT). The seven invasive human isolates differed from each other with zero to two alleles and were allocated to CT1234 (declared as outbreak strain). Among the Austrian strain collection of about 6,000 cgMLST-characterized non-human isolates (i.e., food/environmental isolates) 90 isolates shared CT1234. Out of these, 83 isolates were traced back to one meat processing-company. They differed from the outbreak strain by up to seven alleles; one isolate originated from the company's industrial slicer. The remaining seven CT1234-isolates were obtained from food products of four other companies (five fish-products, one ready-to-eat dumpling and one deer-meat) and differed from the outbreak strain by six to eleven alleles. The outbreak described shows the considerable potential of WGS to identify the source of a listeriosis outbreak. Compared to PFGE analysis, WGS-based typing has higher discriminatory power, yields better data accuracy, and allows higher laboratory through-put at lower cost. Utilization of WGS-based typing results of human and food/ environmental *L. monocytogenes* isolates by appropriate public health analysts and epidemiologists is indispensable to support a successful outbreak investigation.

Keywords: whole-genome sequencing, pulsed-field gel electrophoresis, outbreak investigation, public health laboratory capacity, public health surveillance

INTRODUCTION

Listeriosis is a relatively uncommon disease, which typically causes a severe disease in a high portion of cases and deaths in susceptible population subgroups (1, 2). Listeriosis is a foodborne illness of major public health concern because of the severity of its complications (infections of the central nervous system, septicemia, gastroenteritis and abortion), a hospitalization rate of 98.6% and a case-fatality ratio of 13,8%, as reported by the EU summary report on zoonoses, zoonotic agents and food-borne outbreaks from 2017 (3). The surveillance of listeriosis in the European Union/European Economic Area (EU/EEA) focuses on the severe invasive forms of the disease for which the risk groups are mainly elderly and immunocompromised persons, pregnant women and infants. In 2017, 2,480 confirmed cases of invasive listeriosis were reported by 28 EU/EEA countries, resulting in an overall notification rate of 0.48 per 100,000 population (3). The increasing trend in the number of listeriosis cases in the EU/EEA, probably also due to the increased population size of the elderlies (4, 5), is worrying and calls for utmost attention to be placed on the prevention and control of the disease and outbreaks. The European Center for Disease Prevention and Control (ECDC), the European Food Safety Authority (EFSA) and the European Union Reference Laboratory (EU-RL) for *L. monocytogenes* have set up a joint database collecting, on a voluntary basis, combined AscI/ApaI PFGE profiles for PFGE typing data for human, food, animal and environmental isolates from public heath institutes and food safety and veterinary authorities to enable detection of listeriosis outbreaks affecting several countries (6). However, technical development is evolving fast and whole genome sequencing (WGS)-based typing methods replaced pulsed-field gel electrophoresis (PFGE) as the gold-standard showing higher accuracy and a superior discriminatory power (7). The outbreak described here, illustrates impressively the considerable potential of WGS based typing to elucidate the source of a listeriosis outbreak.

BACKGROUND TO OUTBREAK INVESTIGATION

In Austria, laboratories have a legal obligation to forward human and food/environmental *L. monocytogenes* isolates derived from official controls as well as from ownchecks to the NRL. In January 2016, two human isolates of *L. monocytogenes* serotype 1/2a of the identical pulsed-field gel electrophoresis (PFGE) pattern, previously unknown in Austria, were cultured for the first time. Two environmental isolates of unknown origin, were obtained in January and February 2016, and another 24 food/environmental isolates were obtained between September 2016 and December 2017. In addition five further human isolates from patients with invasive listeriosis, isolated between April 2016 and September 2017 were obtained. All isolates showed this new PFGE pattern. The food and environmental isolates originated from six different laboratories. From January to August 2018, further 69 food/environmental isolates possibly related to the outbreak were sent to the National Reference Laboratory (NRL) for *Listeria*. In summary, a total of 95 non-human isolates together and seven human outbreak isolates were typed by using WGS cgMLST analysis. No reliable information was available on patients' relevant food consumption.

On 25 January 2018, Austria launched an Urgent Inquiry (UI-460) in The Epidemic Intelligence Information System (EPIS), a web-based communication platform that allows nominated public health experts to exchange technical information to assess whether current and emerging public health threats have a potential impact in the European Union. Aim of the outbreak investigation was to identify the source(s) and to recommend the appropriate public health measures for preventing further cases. Thirteen countries (Denmark, Finland, France, Germany, Ireland, Italy, Luxembourg, The Netherlands, Norway, Slovenia, Spain, Sweden and the United Kingdom) answered via the platform and eight countries reported cases with at least six allelic differences to the Austrian outbreak cluster. Raw data of the sequences were provided from the countries, which allowed a direct comparison with the Austrian database. Eight non-human strains isolated in France and the Netherlands were reported in the European Union Reference Laboratory for *Listeria monocytogenes* technical report [EURL Lm 2018 (8)] to form a cgMLST cluster with five to seven pairwise allele differences against the outbreak strain.

MATERIALS AND METHODS

Origin of Isolates, Cultivation, and Genomic DNA Isolation

In Austria, *Listeria* isolates obtained from food and environmental samples, as well as human isolates, must be sent to the NRL for Listeria by legislation. The non-human isolates are anonymized, provided with unique identifier and information on the type of food matrix only (e.g., meat-product, diary-product, vegetable-product, food-environment) by the sending primary food laboratories. Isolates are cultivated on RAPID'L.Mono™ agar plates (Biorad, Munich, Germany) for species confirmation and subsequently subcultured overnight on Columbia Broth (BD Difco™, Heidelberg, Germany) for extraction of high molecular weight genomic DNA using the HMW MagAttract kit (Qiagen, Hilden, Germany) according to the instructions of the manufacturer for Gram positive bacteria.

Whole Genome Sequencing and Data Analysis

Whole genome sequencing was performed as described previously (9). Briefly, for sequencing, an Illumina MiSeq platform (Illumina Inc., San Diego, CA, USA) was used. Library preparation was carried out using Nextera XT according to the instructions of the manufacturer (Illumina Inc., San Diego, CA, USA). For assembly into draft genomes, raw reads were *de novo* assembled using SPAdes version 3.11.1 (10). Contigs were filtered for a minimum coverage of 5-fold and minimum length of 200 bp, which resulted in 26–187 contigs at a coverage of 46–148-fold. Classical multilocus sequence typing (MLST) data

according to Ragon et al. (11) and genoserotyping data according to Hyden et al. (12) were *de novo* extracted from WGS sequence data. Assessment of the core genome multilocus sequence typing (cgMLST) results was done using Ridom SeqSphere+ software version 5.1.0 as described by Ruppitsch et al. (13). All isolates had 98.1–99.8% good targets and a minimum spanning tree (MST) was generated in Ridom SeqSphere+ version 5.1.0 for visualization of strain relatedness. For comparison and data harmonization SeqSphere+ results were compared to the Pasteur cgMLST scheme (7) and GenomeGraphR (14). The sequences have been deposited in DDBJ/EMBL/GenBank under the project number PRJNA434392. Raw sequence data for each strain were deposited under SRA accession numbers (**Table 1**).

SNP analysis was done with GenomeGraphR Beta 2.7 [Sanaa et al. (14)] using the default settings and a cluster threshold definition of 12 SNPs. All strains were compared with the isolates present in the database.

RESULTS

Austria reported a suspected outbreak due *to L. monocytogenes* serotype 1/2a of the same PFGE pattern, including seven patients of invasive listeriosis, having occurred in eastern Austria between 2015 and 2017. The cgMLST typing of the seven human invasive isolates revealed a genetically tight cluster, complex type 1234 (CT1234), which corresponds to CT1170 of Institut Pasteur cgMLST scheme [Moura et al. (7)], with zero to two allelic differences from each other. SNP analysis revealed that our clinical isolates differed from each other by 1–4 SNPs. In addition, the closest clinical strain clustering with our isolates differed by 11–12 SNPs and therefore confirmed the CT1234.

On 26 January 2018, the Austrian Ministry of Health mandated the Austrian Agency for Health and Food Safety (AGES) to investigate this suspected outbreak. A confirmed outbreak case was defined as a patient with invasive listeriosis, positive for *L. monocytogenes* cgMLST CT1234 isolate, which differed by ≤2 alleles from a representative outbreak isolate by using cgMLST, and with a disease onset on or after 1 January 2015.

The aforementioned seven patients fulfilled the definition of a confirmed outbreak case. Patients were 29–97 years old (mean: 68; median: 73), five females and two males, with disease onset between November 2015 and September 2017 and residence in three of the nine Austrian provinces. **Figure 1** depicts the outbreak cases by month of diagnosis and province of residence.

Among the Austrian genome database of about 6,000 non-human isolates (i.e., food/environmental isolates, collected between 2015 and 2018), 90 isolates shared genoserotype IIa, MLST CC155, and cgMLST CT1234. Out of these, 83 isolates were traced back to a meat-processing company (companyA; CoA) in eastern Austria. These food/environmental isolates differed from the outbreak strain by zero to seven alleles and one isolate, originated from the company's industrial slicer. The remaining seven CT1234 isolates were obtained from food products of four other companies (five fish-products, one

TABLE 1 | Accession numbers of sequences available at NCBI Sequence Read Archive (SRA).

ID	Accession no.	cgMLST	MLST CC	Genoserotype
5F_CoA	SRR6740436	1234	155	IIa
6F_CoA	SRR6740437	1234	155	IIa
3E_CoA	SRR6740438	1234	155	IIa
4E_CoA	SRR6740439	1234	155	IIa
1E_uk	SRR6740440	1234	155	IIa
2E_uk	SRR6740441	1234	155	IIa
7H	SRR6740442	1234	155	IIa
4F_CoA	SRR6740443	1234	155	IIa
1F_CoA	SRR6740444	1234	155	IIa
2F_CoA	SRR6740445	1234	155	IIa
15F_CoA	SRR6740446	1234	155	IIa
16F_CoA	SRR6740447	1234	155	IIa
9F_CoA	SRR6740448	1234	155	IIa
10F_CoA	SRR6740449	1234	155	IIa
7F_CoA	SRR6740450	1234	155	IIa
8F_CoA	SRR6740451	1234	155	IIa
13F_CoA	SRR6740452	1234	155	IIa
14F_CoA	SRR6740453	1234	155	IIa
11F_CoA	SRR6740454	1234	155	IIa
12F_CoA	SRR6740455	1234	155	IIa
3F_CoA	SRR6740456	1234	155	IIa
28F_CoA	SRR6740457	1234	155	IIa
19F_CoA	SRR6740458	1234	155	IIa
18F_CoA	SRR6740459	1234	155	IIa
17F_CoA	SRR6740460	1234	155	IIa
4H	SRR6740461	1234	155	IIa
3H	SRR6740462	1234	155	IIa
2H	SRR6740463	1234	155	IIa
1H	SRR6740464	1234	155	IIa
6H	SRR6740465	1234	155	IIa
5H	SRR6740466	1234	155	IIa
90F_CoA	SRR8184623	6743	37	IIa
56F_CoA	SRR8185109	1234	155	IIa
55F_CoA	SRR8185110	1234	155	IIa
58F_CoA	SRR8185111	1234	155	IIa
57F_CoA	SRR8185112	1234	155	IIa
52F_CoA	SRR8185113	1234	155	IIa
51F_CoA	SRR8185114	1234	155	IIa
40F_CoA	SRR8185115	1234	155	IIa
39F_CoA	SRR8185116	1234	155	IIa
38F_CoA	SRR8185117	1234	155	IIa
37F_CoA	SRR8185118	1234	155	IIa
36F_CoA	SRR8185119	1234	155	IIa
35F_CoA	SRR8185120	1234	155	IIa
34F_CoA	SRR8185121	1234	155	IIa
33F_CoA	SRR8185122	1234	155	IIa
32F_CoA	SRR8185123	1234	155	IIa
31F_CoA	SRR8185124	1234	155	IIa
72F_CoA	SRR8185125	1234	155	IIa

(Continued)

TABLE 1 | Continued

ID	Accession no.	cgMLST	MLST CC	Genoserotype
71F_CoA	SRR8185126	1234	155	IIa
74F_CoA	SRR8185127	1234	155	IIa
78F_CoA	SRR8185128	1234	155	IIa
76F_CoA	SRR8185129	1234	155	IIa
27F_nonCoA	SRR8185130	1234	155	IIa
77F_CoA	SRR8185131	1234	155	IIa
73F_CoA	SRR8185132	1234	155	IIa
54F_CoA	SRR8185133	1234	155	IIa
53F_CoA	SRR8185134	1234	155	IIa
89F_CoA	SRR8185135	1234	155	IIa
69F_CoA	SRR8185136	1234	155	IIa
70F_CoA	SRR8185137	1234	155	IIa
65F_CoA	SRR8185138	1234	155	IIa
66F_CoA	SRR8185139	1234	155	IIa
67F_CoA	SRR8185140	1234	155	IIa
68F_CoA	SRR8185141	1234	155	IIa
61F_CoA	SRR8185142	1234	155	IIa
62F_CoA	SRR8185143	1234	155	IIa
63F_CoA	SRR8185144	1234	155	IIa
64F_CoA	SRR8185145	1234	155	IIa
47F_CoA	SRR8185146	1234	155	IIa
48F_CoA	SRR8185147	1234	155	IIa
45F_CoA	SRR8185148	1234	155	IIa
46F_CoA	SRR8185149	1234	155	IIa
43F_CoA	SRR8185150	1234	155	IIa
44F_CoA	SRR8185151	1234	155	IIa
41F_CoA	SRR8185152	1234	155	IIa
42F_CoA	SRR8185153	1234	155	IIa
60F_CoA	SRR8185154	1234	155	IIa
49F_CoA	SRR8185155	1234	155	IIa
50F_CoA	SRR8185156	1234	155	IIa
59F_CoA	SRR8185157	1234	155	IIa
83F_CoA	SRR8185158	5753	517	IIb
29F_CoA	SRR8185159	6252	155	IIa
25F_nonCoA	SRR8185160	1234	155	IIa
24F_nonCoA	SRR8185161	1234	155	IIa
75F_CoA	SRR8185162	1234	155	IIa
26F_nonCoA	SRR8185163	1234	155	IIa
21F_nonCoA	SRR8185164	1234	155	IIa
5E_CoA	SRR8185165	1234	155	IIa
23F_nonCoA	SRR8185166	1234	155	IIa
22F_nonCoA	SRR8185167	1234	155	IIa
81F_CoA	SRR8185168	6399	451	IIa
82F_CoA	SRR8185169	6424	1	IVb
87F_CoA	SRR8185170	1234	155	IIa
84F_CoA	SRR8185171	1234	155	IIa
79F_CoA	SRR8185172	1234	155	IIa
88F_CoA	SRR8185173	1234	155	IIa
80F_CoA	SRR8185174	1234	155	IIa
85F_CoA	SRR8185175	1234	155	IIa
30F_CoA	SRR8185176	1234	155	IIa
86F_CoA	SRR8185177	1234	155	IIa
20F-CoA	SRR8186973	1234	155	IIa

(Continued)

TABLE 1 | Continued

ID	Accession No.	cgMLST	MLST CC	Genoserotype

BioProject ID: PRJNA434392; Title: LIST_2018_CT1234. E_uk, environmental isolate of unknown origin (at the time of submission); E_CoA, environmental isolate, company A-associated; F_CoA, food-isolate, company A-associated; H, human isolate.

ready-to-eat dumpling and deer-meat product each) and differed from the main outbreak strain by six to eleven alleles.

Figure 2 illustrates the non-human isolates of the AGES *Listeria* strain collection by CT1234 allocation and linkage to the meat processing company A.

Figure 3 illustrates the MST of the seven human outbreak isolates, of food and environmental isolates of *L. monocytogenes*, comprising company A associated and non-associated CT1234-isolates (CoA, nonCoA), and company A associated, nonCT1234-isolates.

In November 2017, a 47-year old male developed signs and symptoms compatible with a non-invasive listeriosis (i.e., febrile gastroenteritis) 7 h after consumption of a pizza with sliced ham topping in a restaurant in the Austrian province Tyrol. No patient isolate of *L.monocytogenes* was available. The official sample taken from the sliced pizza ham at the restaurant tested positive for *L. monocytogenes* CT1234 with one allelic difference from the outbreak strain (**Figure 3**: 28F_CoA). Trace-back analyses identified the origin of the ham pizza topping from meat-processing company A.

Public Health Measures

Company A implemented control measures including intensified environmental disinfection, installation of a new slicer and continuous investigation of environmental swabs and newly processed food products for *Listeria*. All food batches had to be negative for *L. monocytogenes* before being released to the market. During these activities, further four strains of *L. monocytogenes* were found in the tested food products of the company. There were of complex types CT6252 (genoserotype IIa, MLST CC155), CT6399 (genoserotype IIa, MLST CC451), CT6424 (genoserotype IVb, MLST CC1), CT6743 (genoserotype IIa, MLST CC37), different from the outbreak CT, CT1234 (**Figure 3**: food isolates 81F, 82F, 83F, and 90F). After August 2018, the public health authorities found no further *L. monocytogenes* positive food products.

DISCUSSION

Investigation of listeriosis outbreaks is difficult due to the multitude of possible food vehicles including a broad range of ready-to-eat foods. Pulsed-field gel electrophoresis (PFGE) was the gold standard for strain typing (6) but has become

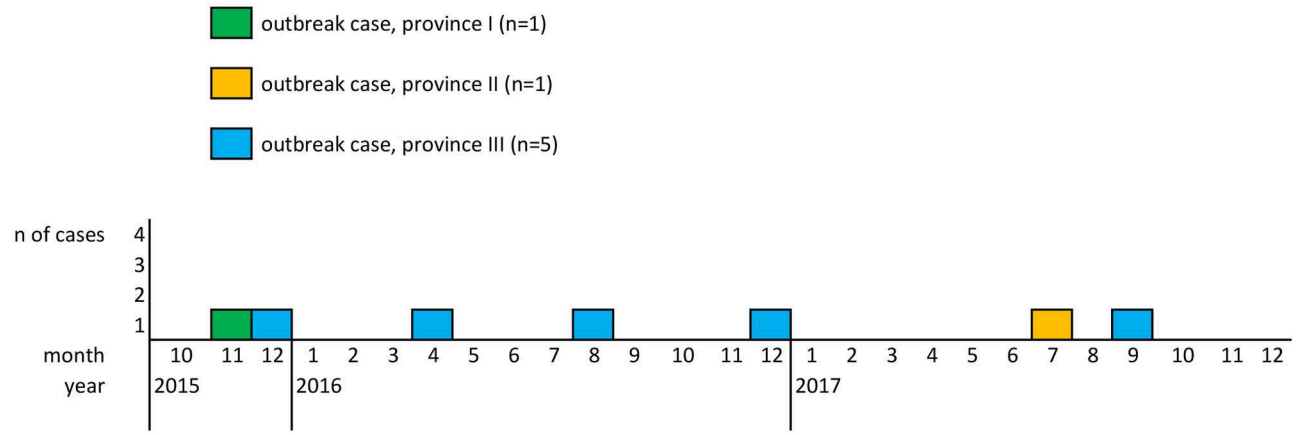


FIGURE 1 | Outbreak cases of *L. monocytogenes* CT1234 by month of disease onset and province of residence, 2015–2017, Austria.

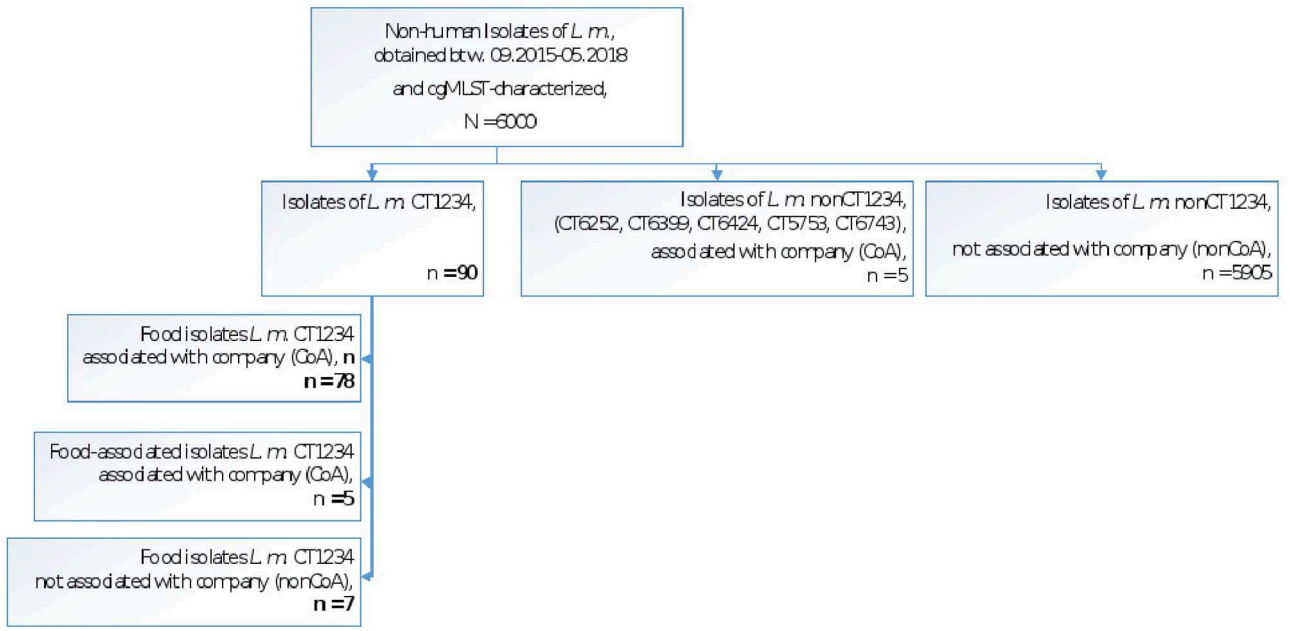


FIGURE 2 | Non-human *L. monocytogenes* isolates of the AGES strain collection (*L.m.* CT1234, *L.m.* nonCT1234) with and without company A linkage (CoA, nonCoA).

obsolete with the advent of WGS. WGS is highly discriminatory and superior for allocating listeriosis cases to an outbreak (7, 13). Due to this superiority of WGS it is time to stop PFGE (Pulsenet Network resolution, ECDC). A dictionary between PFGE and MLST will allow to screen the PFGE database for previous strains using WGS specific ST or CC (15). However, with the limitations of current WGS technology we cannot create PFGE patterns from WGS data and therefore cannot create a PFGE-WGS dictionary. Despite the availability of technical literature on methods for outbreak investigations, there are no pre-specified formulae to dictate the path that an outbreak investigation is supposed to take (16). Investigations of listeriosis outbreaks provide a unique opportunity to gain new scientific knowledge on the occurrence of *L. monocytogenes* in the food-processing setting.

In contrast to Europe, the United States have a zero tolerance policy for *L. monocytogenes* in ready-to-eat foods (17). Commission Regulation (EC) No 2073/2005 of 15 November 2005 on microbiological criteria for foodstuffs requires a limitation of $<$ 100 CFU/g in ready-to-eat food products able to support the growth of *L. monocytogenes* (other than those intended for infants and for special medical purposes) during the shelf life, when products are placed on the market; these particular food products must be tested for the absence of *L.*

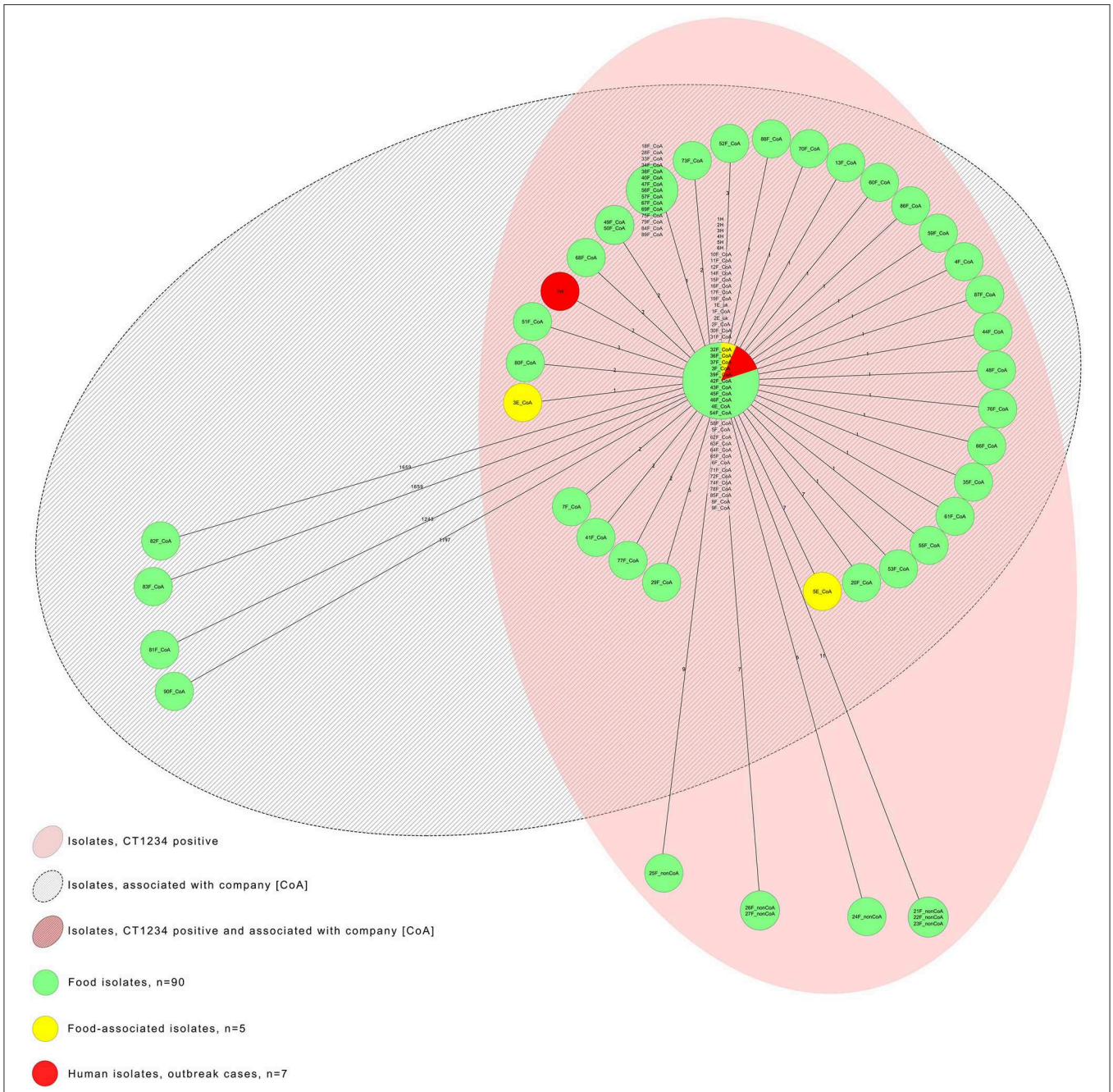


FIGURE 3 | Adapted minimum spanning tree of the seven human outbreak isolates, the food and environmental isolates of *L. monocytogenes*, including the company A associated and non-associated CT1234-isolates (CoA CT1234-isolates and nonCoA CT1234-isolates; $n = 90$), and the five company A associated nonCT1234-isolates; $N = 102$.

monocytogenes in 25 g before leaving the immediate control of the producing food business operator. Otherwise a challenge test which ensures that the limit of 100 cfu/g at the end of shelf-life will not be exceeded, has to be shown (18). Although the infective dose of *L. monocytogenes* is unknown and population subgroups differ in vulnerability to *Listeria monocytogenes* in food, the present European legislation should be sufficient in controlling foodborne listeriosis.

As a consequence of an earlier outbreak of listeriosis in Austria (19, 20), the Federal Ministry of Health had classified 600 food-producing facilities as being at high risk of *Listeria* contamination and ordered the provinces to conduct inspections on various control measures in a Key Activity Action Campaign entitled "Schwerpunktaktion SPA-A-600" in 2014. The province to which company A was assigned had neglected to complete this requirement. From 2015 to 2017, only one official sample of

sliced bacon was obtained on 22 November 2017 by the local food authority at the meat processing company A. The final report, outlining the presence of *L. monocytogenes* in numbers below 100 CFU/g, was not issued until 1 February 2018. Surprisingly, no challenge test was performed for this food-product, especially considering that a similar outbreak caused by the consumption of bacon that was contaminated with *L. monocytogenes* caused four fatalities in Bavaria. At the time in 2016, this outbreak led to a public recall and public warning in Austria (21).

For two decades, PFGE was the reference method for *L. monocytogenes* surveillance and outbreak investigation (2, 22). It is still used for screening but is increasingly replaced by WGS based typing methods (7, 14, 21, 23–26). WGS based typing outperforms PFGE with respect to the discriminatory power, information content, throughput, reproducibility, costs and inter-laboratory data exchange. However, it is important to keep in mind that the differences between the cgMLST schemes of Moura et al. and Ruppitsch et al. can have an impact on the cluster detection (24). For communication on detected clusters it is important to know which core genome scheme, assembler, and assembler version and sequencing technology was used and which average sequencing coverage was achieved.

Based on the current cgMLST analysis of the human outbreak isolates, a difference from each other by only zero to two alleles, and of the majority of the outbreak associated food-isolates, a difference by zero to four alleles should be considered to increase the specificity of linking isolates to *L. monocytogenes* outbreaks. WGS data not only allow to infer phylogenetic relationships but also to filter for additional information like serotypes (12), virulence- and resistance-genes (7).

Although it is known that SNP analysis provides maximal discriminatory power, results are difficult to standardize and interpret (27). Moreover, the analysis based on single nucleotide polymorphisms (SNPs) showed here identical results to the ones obtained by cgMLST. Expansion of the classical MLST principle to a genome wide gene-by-gene comparison allowed the establishment of databases based on well-defined core genome or whole genome MLST schemes (7, 13, 24, 28). The setup of open accessible databases (*Listeria monocytogenes* cgMLST at https://www.cgmlst.org/ncs/schema/690488/, BIGSdb-*Lm* at http://bigsdb.pasteur.fr/listeria) allows the comparison and sharing of data between public health laboratories worldwide and facilitates international source tracking and multinational outbreak investigation (29, 30). These new WGS databases, although only 3 years old, already harbor nearly twice the number of strain complex types (CT) than the >20 year old PulseNet PFGE database demonstrating again the higher discriminative power of WGS based typing. Compared to PFGE, these major improvements in *L. monocytogenes* typing allow a faster and more discriminative detection of clusters and reduce unnecessary epidemiological investigations. *L. monocytogenes* is one of the pathogens for which a rapid transition from traditional typing methods to WGS-based typing methods is presently occurring in the public health laboratories of the EU/EEA as well as the PulseNet International network, and it is the first food- and waterborne pathogen for which a comprehensive WGS-assisted real-time surveillance is planned to be established at the EU/EEA level (31). Due to the superiority of WGS for real-time surveillance in a One Health approach, the PulseNet International network and EU/EEA health and food safety authorities move to cgMLST and wgMLST analysis (24, 30).

CONCLUSIONS

Compared to PFGE analysis, WGS based typing has a higher discriminatory power, yields better data accuracy, and allows higher laboratory through-put at lower cost, as proven in the current outbreak investigation (26). The meaningful use of WGS based typing data for a successful investigation of a listeriosis outbreak and the appropriate public health measures, requires intense collaboration between the public health and food safety authorities, food microbiologists, typing experts and epidemiologists.

AUTHOR CONTRIBUTIONS

AP, DS, FA, and WR contributed conception and design of the study, data analysis and interpretation and writing of the manuscript. AM, AL, AS, AC, SM, and SL performed data analysis. SH and BS contributed to the interpretation of data of the work and critically revised the content of the study. All authors contributed to manuscript revision, read and approved the submitted version.

FUNDING

Part of the sequencing-work was funded by a grant awarded under the One Health European Joint Programme (OHEJP) JRP7 LISTADAPT.

REFERENCES

1. Allerberger F, and Wagner, M. Listeriosis: a resurgent foodborne infection. *Clin Microbiol Infect.* (2010) 16:16–23. doi: 10.1111/j.1469-0691.2009.03109.x
2. Allerberger F, Bagó Z, Huhulescu S, and Pietzka A. Listeriosis: the dark side of refrigeration and ensiling. In: Sing A, editor. *Zoonoses - Infections Affecting Humans and Animals. Focus on Public Health Aspects.* Heidelberg, DE: Springer Verlag (2015), p. 249–86.
3. EFSA and ECDC (European Food Safety Authority and European Centre for Disease Prevention and Control). The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017. *EFSA J.* (2018) 16:5500. doi: 10.2903/j.efsa.2018.5500
4. European Centre for Disease Prevention and Control. Listeria. In: *ECDC Annual epidemiological report for 2016.* Stockholm: European Centre for Disease Prevention and Control (2016).
5. EFSA Panel on Biological Hazards, Ricci A, Allende A, Bolton D, Chemaly M, Davies R. et al. *Listeria monocytogenes* contamination of ready-to-eat foods and the risk for human health in the EU. *EFSA J.* (2018) 16:5134. doi: 10.2903/j.efsa.2018.5134

6. Félix B, Danan C, Van Walle I, Lailler R, Texier T, Lombard B, et al. Building a molecular *Listeria monocytogenes* database to centralize and share PFGE typing data from food, environmental and animal strains throughout Europe. *J Microbiol Methods.* (2014) 104:1–8. doi: 10.1016/j.mimet.2014.06.001
7. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol.* (2016) 2:16185. doi: 10.1038/nmicrobiol.2016.185
8. European Union Reference Laboratory for Listeria monocytogenes technical reports to EFSA and DG SANTE related to UI-460 ST155. Technical report ST155 UI-460 multicountry outbreak of Listeria monocytogenes Version 1, Foodborne pathogens department, Laboratory for Food Safety, Anses, Salmonella Et Listeria. (2018).
9. Lepuschitz S, Mach R, Springer B, Allerberger F, Ruppitsch W. Draft genome sequence of a community-acquired methicillin-resistant *Staphylococcus aureus* USA300 Isolate from a River Sample. *Genome Announce.* (2017) 5:e01166–17. doi: 10.1128/genomeA.01166-17
10. Nurk S, Bankevich A, Antipov D, Gurevich AA, Korobeynikov A, Lapidus A, et al. Assembling genomes and mini-metagenomes from chimeric MDA products. *J Comput Biol.* (2013) 20:714–37. doi: 10.1089/cmb.2013.0084
11. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le Monnier A, & Brisse S. A new perspective on Listeria monocytogenes evolution. *PLoS Pathog.* 4:e1000146. doi: 10.1371/journal.ppat.1000146
12. Hyden P, Pietzka A, Lennkh A, Murer A, Springer B, Blaschitz M, et al. Whole genome sequence-based serogrouping of *Listeria monocytogenes* isolates. *J Biotechnol.* (2016) 235:181–6. doi: 10.1016/j.jbiotec.2016.06.005
13. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, et al. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Listeria monocytogenes*. *J Clin Microbiol.* (2015) 53:2869–76. doi: 10.1128/JCM.01193-15
14. Sanaa M, Pouillot R, Vega FG, Strain E, Van Doren JM. GenomeGraphR: a user-friendly open-source web application for foodborne pathogen whole genome sequencing data integration, analysis, and visualization. *PLoS ONE.* (2019) 14:e0213039. doi: 10.1371/journal.pone.0213039
15. Maury MM, Tsai YH, Charlier C, Touchon M, Chenal-Francisque V, Leclercq, A. et al. (2016). Uncovering Listeria monocytogenes hypervirulence by harnessing its biodiversity. *Nat Genet.* 48:308–313. doi: 10.1038/ng.3501
16. Jamasji-Pavri A. Transaction cost economics and principal-agent theory: insights into investigations of outbreaks of infectious diseases. In: Roberts JA, editor. *The Economics of Infectious Disease.* Oxford: Oxford University Press, (2006) p. 261–79.
17. Montville TJ, Matthews KR. *Food Microbiology: An Introduction. 2nd ed.* Washington DC: ASM Press, (2008).
18. European Commission. Commission Regulation (EC) No 2073/2005 of 15 November 2005 on microbiological criteria for foodstuffs. *Official Journal.* (2005) 338:1–26.
19. Fretz R, Pichler J, Sagel U, Much P, Ruppitsch W, Pietzka AT, et al. Update: multinational listeriosis outbreak due to 'Quargel', a sour milk curd cheese, caused by two different *L. monocytogenes* serotype 1/2a strains, 2009-2010. *Euro Surveill.* (2010) 15:19543. doi: 10.2807/ese.15.16.19543-en
20. Pichler J, Appl G, Pietzka A, Allerberger F. Lessons to be Learned from an Outbreak of Foodborne Listeriosis, Austria 2009–2010. *Food Prot Trends.* (2011) 31:268–73.
21. Ruppitsch W, Prager R, Halbedel S, Hyden P, Pietzka A, Huhulescu S. Ongoing outbreak of invasive listeriosis, Germany, 2012 to 2015. *Euro Surveill.* (2015) 20:30094. doi: 10.2807/1560-7917.ES.2015.20.50.30094
22. Swaminathan B, Barrett TJ, Hunter SB, Tauxe RV, & CDC PulseNet Task Force PulseNet: the molecular subtyping network for foodborne bacterial disease surveillance, United States. *Emerg Infect Dis.* 7:382–9. doi: 10.3201/eid0703.010303
23. Chen Y, Gonzalez-Escalona N, Hammack TS, Allard MW, Strain EA, Brown EW. Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. *Appl Environ Microb.* (2016) 82:6258–72. doi: 10.1128/AEM.01532-16
24. Van Walle I, Björkman JT, Cormican M, Dallman T, Mossong J, Moura A, et al. Retrospective validation of whole genome sequencing-enhanced surveillance of listeriosis in Europe, 2010 to 2015. *Euro Surveill.* (2018) 23:1700798. doi: 10.2807/1560-7917.ES.2018.23.33.17 00798
25. Schmid D, Allerberger F, Huhulescu S, Pietzka A, Amar C, Kleta, S. et al. Whole genome sequencing as a tool to investigate a cluster of seven cases of listeriosis in Austria and Germany, 2011-2013. *Clin Microbiol Infect.* (2014) 20:431–6. doi: 10.1111/1469-0691.12638
26. Reimer A, Weedmark K, Petkau A, Peterson C, Walker M, Knox, N. et al. Shared genome analyses of notable listeriosis outbreaks, highlighting the critical importance of epidemiological evidence, input datasets and interpretation criteria. *Microb Genom.* (2019) 5: 237. doi: 10.1099/mgen.0.000237
27. Pightling AW, Pettengill JB, Luo Y, Baugher JD, Rand H, and Strain E. Interpreting whole-genome sequence analyses of foodborne bacteria for regulatory applications and outbreak investigations. *Front Microbiol.* (2018) 10:1482. doi: 10.3389/fmicb.2018.01482
28. Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen, N. et al. Real-Time Whole-Genome Sequencing for Surveillance of Listeria monocytogenes, France. *Emerg Infect Dis.* 23:1462–70. doi: 10.3201/eid2309.170336
29. Schjørring S, Gillesberg Lassen S, Jensen T, Moura A, Kjeldgaard JS, Müller L, et al. Cross-border outbreak of listeriosis caused by cold-smoked salmon, revealed by integrated surveillance and whole genome sequencing (WGS), Denmark and France, 2015 to 2017. *Euro Surveill.* (2017) 22:17-00762. doi: 10.2807/1560-7917.ES.2017.22.50.17-00762
30. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill.* (2017) 22:30544. doi: 10.2807/156 0-7917
31. Revez J, Espinosa L, Albiger B, Leitmeyer KC, and Struelens MJ. ECDC microbiology focal points and experts group. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of European national capacities, 2015-2016. *Front Public Health.* (2017) 5:347. doi: 10.3389/fpubh.2017.00347

Conflict of Interest Statement: We declare that none of the authors have any commercial and financial relationship to the company Ridom GmbH (Münster, Germany), developer of the Ridom SeqSphere+ software, mentioned in the manuscript.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

REVIEW
published: 20 June 2019
doi: 10.3389/fpubh.2019.00164

Genomic Delineation of Zoonotic Origins of *Clostridium difficile*

Daniel R. Knight[1] *and Thomas V. Riley*[1,2,3,4]*

[1] *Medical, Molecular, and Forensic Sciences, Murdoch University, Perth, WA, Australia,* [2] *School of Medical and Health Sciences, Edith Cowan University, Joondalup, WA, Australia,* [3] *School of Biomedical Sciences, The University of Western Australia, Nedlands, WA, Australia,* [4] *PathWest Laboratory Medicine, Department of Microbiology, Nedlands, WA, Australia*

Clostridium difficile is toxin-producing antimicrobial resistant (AMR) enteropathogen historically associated with diarrhea and pseudomembranous colitis in hospitalized patients. In recent years, there have been dramatic increases in the incidence and severity of *C. difficile* infection (CDI), and associated morbidity and mortality, in both healthcare and community settings. *C. difficile* is an ancient and diverse species that displays a sympatric lifestyle, establishing itself in a range of ecological niches external to the healthcare system. These sources/reservoirs include food, water, soil, and over a dozen animal species, in particular, livestock such as pigs and cattle. In a manner analogous to human infection, excessive antimicrobial exposure, particularly to cephalosporins, is driving the expansion of *C. difficile* in livestock populations worldwide. Subsequent spore contamination of meat, vegetables grown in soil containing animal feces, agricultural by-products such as compost and manure, and the environment in general (households, lawns, and public spaces) is contributing to a persistent community source/reservoir of *C. difficile* and the insidious rise of CDI in the community. The whole-genome sequencing era continues to redefine our view of this complex pathogen. The application of high-resolution microbial genomics in a One Health framework (encompassing clinical, veterinary, and environment derived datasets) is the optimal paradigm for advancing our understanding of CDI in humans and animals. This approach has begun to yield critical insights into the genetic diversity, evolution, AMR, and zoonotic potential of *C. difficile*. In Europe, North America, and Australia, microevolutionary analysis of the *C. difficile* core genome shows strains common to humans and animals (livestock or companion animals) do not form distinct populations but share a recent evolutionary history. Moreover, for *C. difficile* sequence type 11 and PCR ribotypes 078 and 014, major lineages of One Health importance, this approach has substantiated inter-species clonal transmission between animals and humans. These findings indicate either a zoonosis or anthroponosis. Moreover, they challenge the existing paradigm and the long-held misconception that CDI is primarily a healthcare-associated infection. In this article, evolutionary, and zoonotic aspects of CDI are discussed, including the anthropomorphic factors that contribute to the spread of *C. difficile* from the farm to the community.

Keywords: evolution, transmission, *Clostridium difficile*, one health, livestock, zoonosis

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control, Sweden

Reviewed by:
David Eyre, University of Oxford, United Kingdom
Sergio Alvarez-Perez, KU Leuven, Belgium

***Correspondence:**
Thomas V. Riley
T.Riley@murdoch.edu.au

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *11 April 2019*
Accepted: *03 June 2019*
Published: *20 June 2019*

Citation:
Knight DR and Riley TV (2019) Genomic Delineation of Zoonotic Origins of Clostridium difficile. Front. Public Health 7:164. doi: 10.3389/fpubh.2019.00164

INTRODUCTION

Last year was the 40th anniversary of the publication in 1978 of a series of papers from several research groups that provided proof that *Clostridium difficile* caused pseudomembranous colitis (1–4). While the spectrum of gastrointestinal disease caused by *C. difficile* has broadened significantly since then, for much of those 40 years *C. difficile* was thought of as causing disease almost exclusively within high-risk hospitalized patient populations (5). In evolutionary terms, 40 years is a negligible length of time. The *Clostridia* are an ancient prokaryotic lineage, estimated to have diverged from the bacterial domain 2.34 Ga (billion years) ago around the time when concentrations of molecular oxygen in the atmosphere began to increase (6). With the advances of next-generation sequencing, the taxonomy of the *Clostridia* is currently undergoing a major revision. Indeed, given the significant differences between *C. difficile* and some other pathogenic clostridia, it has been proposed that it be renamed *Clostridioides difficile* (7). While this has caused some angst in the *C. difficile* community, both names are currently viewed as being "validly published" and therefore acceptable (8).

In recent years, the vast majority of emerging or re-emerging infections have been vector-borne or zoonoses—animal diseases that are transmissible to humans (9). Most attention has focused on viral infections because of highly publicized outbreaks; SARS, avian influenza, and Ebola. However, disease associated with *C. difficile* infection (CDI) has killed more people worldwide in the last 15 years than all these viral infections combined, around 30,000 per year in the USA alone according to the CDC (10). CDI should always have been considered a zoonosis, either direct or indirect. In some definitions of zoonoses, non-human animal hosts play an essential role in maintaining the infection in nature and humans are only incidental hosts. In CDI, all animals (human and non-human) are likely hosts; the wide variety of animals from which *C. difficile* has already been isolated suggests this (11).

What then is the natural history of CDI following exposure to *C. difficile*? *C. difficile* is ubiquitous in the environment. *C. difficile* colonizes the gastrointestinal tracts of all animals during the neonatal period, multiplies, and is excreted, but cannot/does not compete well when other bacterial species start to colonize. The exact timing of this change is not clear, but it is probably linked to changes in diet in babies, i.e., weaning. Through a process known as colonization resistance, a well-developed microbiota provides protection against overgrowth of *C. difficile* by inhibiting germination, vegetative growth, and toxin production (12). In human and non-human animals, antimicrobial exposure creates an environment that could be thought of as mimicking the neonatal gut—characterized by an underdeveloped microbiota and consequently reduced or absent colonization resistance. In such a compromised host gut, *C. difficile* spores rapidly germinate and begin to produce potent cytotoxins (toxin A and toxin B) which cause extensive colonic inflammation and epithelial tissue damage, the net effect being a rapid fluid loss into the intestinal lumen which manifests as diarrhea (13). Some strains also produce a binary toxin, an ADP-ribosyltransferase that causes actin cytoskeletal disruption, and is associated with more severe CDI, a higher case-fatality rate and refractory disease (14).

When those antimicrobials were cephalosporins in the 1980s and 90s, antimicrobials to which *C. difficile* is intrinsically resistant, there was an expansion of CDI in hospitals that continues today. Since the 1990s in North America, cephalosporins have been licensed for use in food animals. There has been an amplification of *C. difficile* in food animals since then, with subsequent contamination of meat, and vegetables grown in soil containing animal feces. In some animals such as piglets, there is overt disease with significant impact on industry. "Animal" strains of *C. difficile* are now infecting humans. *C. difficile* ribotype (RT) 027 was found in animals in North America in the early 2000s (15) but probably moved from animals to humans a decade earlier around the time that RT027 developed resistance to fluoroquinolone antimicrobials (16). This strain was likely to have initially caused infections in the community at a time when community-acquired (CA) CDI [defined as cases with symptom onset in the community or $\leq$48 h after admission to a healthcare facility (17)] was thought infrequent, and diarrhea in the community was rarely investigated. The mutation to fluoroquinolone resistance and high use of fluoroquinolones drove RT027 spread, in North America and later Europe, once it entered the hospital system (16). A similar process now appears to be occurring with *C. difficile* RT078, another animal strain that has increased significantly as a cause of CA-CDI in Europe over the last 10 years (18, 19). *C. difficile* continues to expand in food animal populations, driven by cephalosporin use, and animal strains of *C. difficile* are driving the worldwide increase in CA-CDI.

The whole-genome sequencing era continues to redefine our view of this complex pathogen. The application of high-resolution microbial genomics in a One Health framework (encompassing clinical, veterinary, and environment derived datasets) is the optimal paradigm for advancing our understanding of CDI in humans and animals. This approach has begun to yield critical insights into the genetic diversity, evolution, AMR, and zoonotic potential of *C. difficile*. In this review, evolutionary and zoonotic aspects of CDI are discussed, including the anthropomorphic factors that contribute to the spread of *C. difficile* from the farm to the community.

Community-Acquired CDI

Surveillance data indicate that CA-CDI comprises a significant fraction of total CDI cases and that the incidence of CA-CDI has been increasing globally (20). In the United States, CA-CDI accounts for around a third of all CDI cases and increased 4-fold during the period 1991–2005 (18, 21–24). In another US study, comparable incidence rates for CA-CDI and hospital-associated CDI (HA-CDI) were reported (11.2 cases/100,000 person-years and 12.1 cases/100,000 person-years, respectively) (18). A recent European multi-center study (97 hospitals in 34 European countries) found 14% of 506 cases were classified CA-CDI (25). In Australia, data from 2011 to 2012 showed CA-CDI accounted for up to a quarter of all cases (26% of 5,109 CDI cases) and has been increasing in recent years (26–28). More recent studies from the USA report higher proportions

of CA-CDI around 40% (24). Many studies have noted that individuals with CA-CDI often do not have the "classical" risk factors for CDI acquisition and are generally younger, healthy, and female, without contact with hospitalized patients nor prior antimicrobial exposure (5, 20, 29). In up to 40% of CA-CDI cases, infection is more severe and there are adverse outcomes (hospitalization, treatment failure, complications, colectomy, and recurrence) (19, 30). Notably, *C. difficile* strains acquired in the community can differ in genotype from predominant hospital strains (31), however, *C. difficile* RT078 (see below) has emerged as a significant pathogen associated with both HA- and CA-CDI in the Northern Hemisphere (21, 24, 32–35).

Zoonotic and Environmental Sources of *C. difficile*

C. difficile shows remarkable adaption to life within a diverse array of natural and host environments, including its primary habitat the mammalian gastrointestinal tract (as a commensal and/or pathogen), and several secondary habitats such as water, soil, and compost. We have previously reviewed aspects of *C. difficile* prevalence, pathogenicity and antimicrobial resistance (AMR) in non-human reservoirs (36), as have others including excellent reviews by Rodriguez et al. (11) and Candel-Pérez et al. (37). Here we will briefly summarize the key prevalence and molecular data that suggest a zoonotic origin for CDI. **Figure 1** summarizes *C. difficile* prevalence data in farm animals, food and the environment taken from 86 studies in 23 countries worldwide (15, 38–122). In many of these studies, differences in *C. difficile* prevalence, strain lineage, toxigenic status, and AMR were identified. These were influenced by a variety of factors including the age of the animal, geographic region, methods used for isolation (e.g., sample type, spore selection, enrichment vs. no enrichment) and veterinary and agricultural practices [see recent reviews (11, 37)].

C. difficile is known to colonize numerous food-producing animals including pigs, cattle, sheep, lambs, and poultry. Neonatal animals are viewed as significant reservoirs for *C. difficile* (**Figure 1**). Prevalence in domestic pigs and piglets averages around ∼43%, ranging from 0% [Belgium and Switzerland (98, 103)] to ∼50% [USA and Slovenia (61, 70)] and 100% [Spain and The Netherlands (62, 68)]. In cattle and calves, *C. difficile* prevalence averages around 14%, ranging from 0.5% [Switzerland (98)] to ∼20% [Italy, Belgium and the USA (43, 46, 103)] to ∼50% [Australia and Canada (38, 40)]. On average, a lower prevalence has been reported in ovine hosts [sheep and lambs, ∼6% (77)] with prevalence in poultry [hens, broiler chickens] varying considerably [0.3% in the USA (82), to 29.0% in Zimbabwe (83) and 62% in Slovenia (80), mean ∼19%]. Due to an absence of colonization resistance afforded by a mature intestinal microflora, during the first weeks of life neonatal pigs and calves are susceptible to disease caused by *C. difficile*. Although data is limited for calves (46) the pathophysiology of CDI in piglets is well-described; diarrhea, dehydration, weight loss, enteritis histologically similar to human lesions, and high mortality (123–125).

Other non-human animal reservoirs of *C. difficile* include cats and dogs (prevalence 0–100%), horses and foals (3–33%) and numerous wild animal species including rabbits, zebra, kangaroos, birds, shrews, Kodiak bears, racoons, camels, donkeys, feral swine, elephants, ibex, molluscs, tamarin monkeys, chimpanzees and, most recently, polar bears (0–100%) (37, 126, 127). The most common *C. difficile* lineage identified in many of these animal studies is multilocus sequence type (MLST, ST) 11, predominated by RT078 and its close relatives RTs 033, 045, 066, 126, 127, and 288 (all binary toxin positive, toxinotype V and cause CDI in humans) (**Figure 1**). Surprisingly, in Australia, the predominant RT found in pig herds is RT014, one of the most common strains causing CDI in humans worldwide (128) (see below).

C. difficile has been recovered from meats and plant-based foods sourced from processing plants, shops, farms and markets throughout Europe, North America and the Middle East (**Figure 1**). These include retail meat (veal, beef, pork, lamb, chicken, goat, buffalo, and turkey), seafood (salmon, perch, clams, shrimp, and mussels), and salads and vegetables (lettuce, pea sprouts, ginger, carrots, potatoes, onions, and spinach). As is the case with farm animals, the prevalence of *C. difficile* in food varies widely with food type and geographic origin. A high prevalence of *C. difficile* in retail pork, beef, and chicken has been reported in the USA (42%) but studies elsewhere report a much lower prevalence (Taiwan, 23%; Cote d'Ivoire, 14%), especially in Europe (∼3.0%) (105, 129, 130). The prevalence of *C. difficile* in seafood varies considerably from ∼5.0% in Canada, USA and Wales (99, 108, 118) to ∼50.0% in Italy where its presence has been tentatively linked to sewage contamination in local rivers (95). Similarly, the prevalence of *C. difficile* on vegetables varies from 3 to 8% in North America and Europe [ready to eat salads (85, 101, 107, 109, 111, 118)] to 20–56% in Australia [organic beetroot and potatoes (84)] reflecting, possibly, different methods of processing. The molecular epidemiology of *C. difficile* recovered from food largely mirrors that of farm animals (ST11 RTs and common healthcare-associated lineages including 014 and 027, **Figure 1**).

Farm to Fork: Agricultural Practices Presenting a Risk for CA-CDI

In its spore form, *C. difficile* persists in various different natural ecosystems [soil, rivers, oceans, lakes, and sediments (114–116, 118, 119)], animals and food (11), and many abiotic environments for example toilets, floors, sinks, and soles of shoes (112, 113, 131). The high transmissibility of the spore (132) combined with its inherent resilience to desiccation, extremes of temperature, and disinfection (133) facilitates the transmission of *C. difficile* between these ecosystems. *C. difficile* spores could be transmitted from the farm environment to humans through a number of mechanisms including direct contact, airborne dispersal, avian, rodent or arthropod vectors (134–137), contamination of meat with feces during slaughter (53, 138) and via animal effluent or effluent by-products such as compost (139). However, CDI is a complex phenomenon encompassing pathogen, host, anthropomorphic and environmental factors,

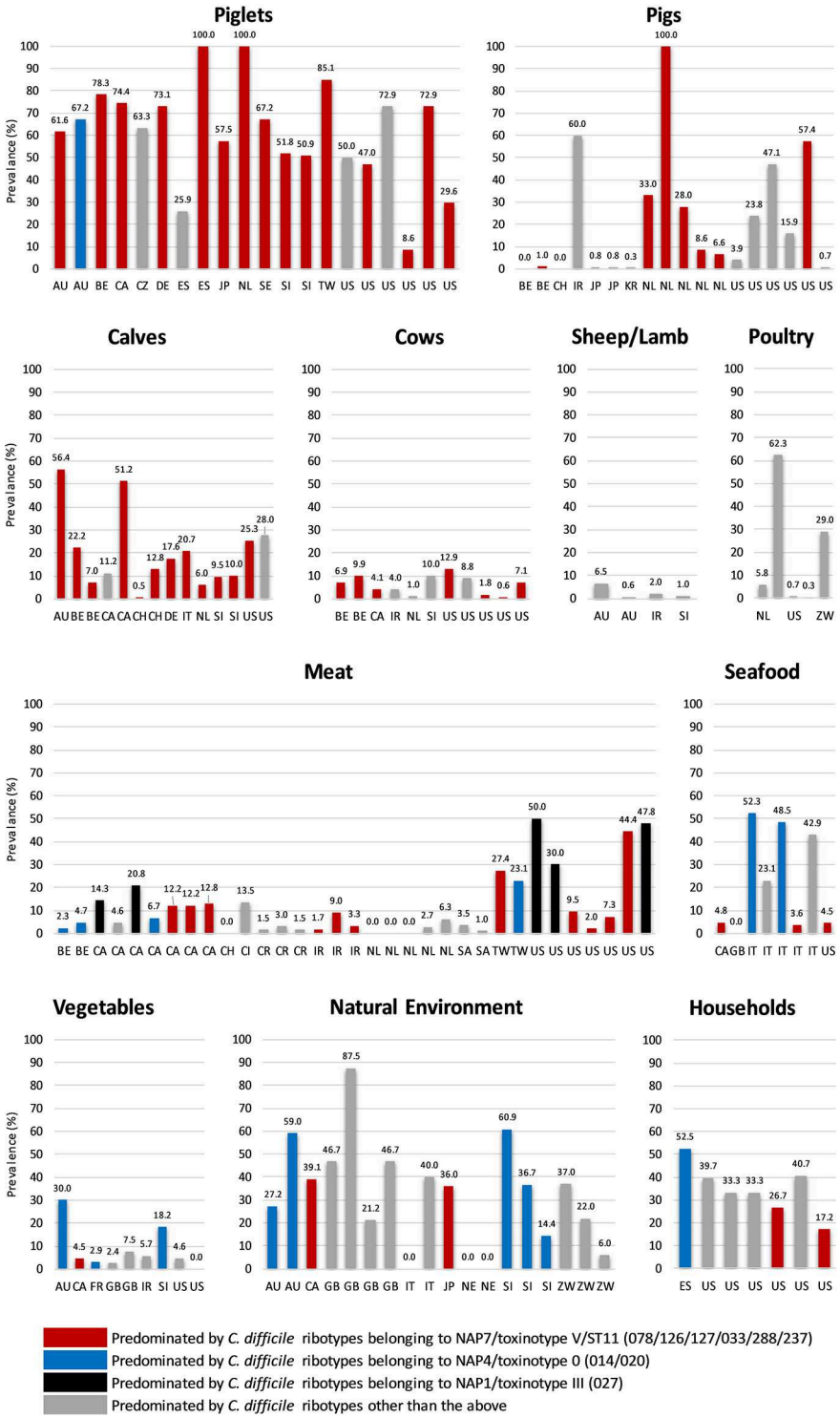


FIGURE 1 | Global prevalence of *C. difficile* in farm animals, food and the environment. Data were taken from 86 studies in 23 countries worldwide (15, 38–122). Categories: Poultry (hens and broilers, Seafood (salmon, perch, clams, shrimp, and mussels), Meat (veal, beef, pork, lamb, chicken, goat, buffalo, and turkey), Vegetables (salads, lettuce, pea sprouts, ginger, carrots, beetroot, potatoes, onions, and spinach), Household (sandbox, shoes, toilet, vacuum, sink, floor), and Natural Environment (compost, lawn, soil, sediment, lake, and river). Two-letter country codes (International Organization for Standardization, ISO): AU, Australia; BE, Belgium; CA, Canada; CH, Switzerland; CI, Ivory Coast; CR, Costa Rica; CZ, Czech Republic; DE, Germany; ES, Spain; FR, France; GB, Great Britain and Northern Ireland; IR, Iran; IT, Italy; JP, Japan; KR, Korea; NE, Nigeria; NL, The Netherlands; SA, Saudi Arabia; SE, Sweden; SI, Slovenia; TW, Taiwan; US, United States of America; ZW, Zimbabwe. NAP, North American Pulse Type. RT027 and all ST11 RTs listed are binary toxin-positive.

and our understanding of CDI transmission dynamics between production animals and humans is nowhere near perfect. Within Australia, two agricultural practices have been identified which present a credible risk for transmission of *C. difficile* causing CA-CDI: (i) slaughtering of neonatal animals destined for human consumption, and (ii) the recycling of effluent for agricultural purposes such as manufacturing compost which is then disseminated into the community setting (140, 141).

The prevalence of *C. difficile* in Australian veal calves is high although this decreases significantly with increasing age of the animal; 56% from <7-day-old calves, 3.8% in 2–6 month-old calves, and 1.8% in adult cattle (38). The *C. difficile* population within these cattle was dominated by ST11 RTs that all cause disease in humans. Moreover, at slaughter, the prevalence of *C. difficile* in calve feces was 60.0% and a significant proportion of calf carcasses (25.3%) was positive (with a spore concentration of 33 CFU/cm^2), as a result of spore contamination from gastrointestinal contents during the slaughter process (138). As before, clinically important ST11 RT lineages dominated (138). Australia is one of the very few countries that cull male neonatal dairy calves (veal calves), a practice that exists because they are born male and considered surplus to industry requirements. With *C. difficile* prevalence highest in this neonatal period (127), the unique slaughter age of these animals presents a significant and perhaps under-appreciated risk for contamination of carcasses during the slaughter process. Further, *C. difficile* spores contaminating carcases would likely survive chilling, freezing, and cooking processes (142–145) and may compromise the quality of veal for domestic and export markets. To date, *C. difficile* has not been recovered from retail meat in Australia although only limited surveys have been undertaken mainly on meat from adult animals. Consumer demand for newborn veal in Australia is low and thus there is likely to be limited exposure of consumers to contaminated meats. However, Australia is the third largest beef and veal producer in the world (146), exporting 1.9 million tons of beef and veal per annum to over 100 countries, particularly in Africa, Asia and the Middle East. It is possible that contaminated Australian veal may be contributing to CDI in these regions, however, with the exception of Taiwan where ST11 strains are commonly reported in humans with CDI and farm animals (64, 102, 147), CDI surveillance is lacking in many of these countries. Whatever the level of risk to the domestic and export consumer, it is possible that it can be significantly mitigated by increasing the age that the animal is slaughtered to >3 or more weeks (38).

In the case of Australian piglets and dissemination of the major healthcare-associated lineage RT014, a growing body of evidence points to zoonotic transmission extending from the farrowing shed to the community. First, Australian piglets are major amplification reservoirs for *C. difficile* (67% prevalence nationwide with RT014 comprising 23% of isolates) (52). Second, whilst suckling age piglets are not slaughtered for meat on a large scale, *C. difficile* spores are abundant in treated biosolids, effluent, and piggery wastewater (121, 148–150). These by-products of the pig industry are subsequently recycled to pasture and agriculture for composting and direct irrigation/fertilization of crops and lawn. Third, *C. difficile* has been recovered from 30% of "high-street" retail compost samples in Australia (122), 59% of new roll-on lawn samples in Australia (151) and 20% of various root vegetables from mainstream and organic markets (84). Both lawn and organic vegetables are invariably grown in compost/soil containing animal manure. In these studies, RT014 comprised 7, 39, and 10% of isolates, respectively. Finally, the use of potent, late generation cephalosporins in human and veterinary medicine is a major driver of (i) *C. difficile* colonization and onset of disease in pigs; (ii) amplification and persistence of *C. difficile* in piggeries; (iii) spill-over of spores into the environment; and (iv) onset of CDI in the community (135, 140, 141).

GENOMIC INSIGHTS INTO THE EVOLUTION AND TRANSMISSION OF *C. DIFFICILE* IN ANIMALS AND HUMANS

Microevolution in the *C. difficile* Core Genome

The next generation sequencing era has seen the development of exquisitely sensitive, cost-effective, and rapid, benchtop whole-genome sequencing (WGS) technologies. Combined with new WGS-based genotyping tools, these technologies are shaping the future of infectious diseases surveillance. Core genome single nucleotide variant (SNV) analysis is an ultra-fine scale discriminatory method that uses WGS to detect transmission and outbreaks of bacterial pathogens (152, 153). SNV analysis is restricted to the non-repetitive, non-recombinative core genome which contains essential genes common to all isolates under analysis that are often vertically inherited and most likely to have the strongest signal-to-noise ratio for inferring phylogeny (152, 153). For *C. difficile*, SNV analysis uses a fixed-rate molecular clock derived from serial isolation of strains from clinical cases, estimated to be in the region of 1.47×10^{-7} to 5.33×10^{-7} mutations per site per year, to identify signatures of plausible clonal transmission (154, 155). This equates to 1–2 SNVs per genome per year. For studies of *C. difficile* transmission, a clonal group is therefore defined as two or more strains differing by <2 SNVs in their core genome, with ≥10 SNVs used as a threshold for genetically distinct isolates (154–157). For longer-term ecological studies, these thresholds may not hold true as the genetically quiescent nature of *C. difficile* spores may result in underestimating the evolutionary distance between strains (19).

The ultra-fine scale resolution of this technique is superior to conventional *C. difficile* typing methods including PCR ribotyping, pulsed-field gel electrophoresis (PFGE), MLST, Rep-PCR, toxinotyping, and amplified fragment-length polymorphism (AFLP) fingerprinting (152). It also shows discriminatory power comparable, and in some cases superior, to multilocus variable-number tandem repeat analysis (MLVA) (152, 157) and the recently developed core genome MLST scheme (158). **Supplementary Table 1** provides a summary of bioinformatics tools and algorithms involved in a *C. difficile* SNV pipeline.

For *C. difficile*, SNV-based typing has been used to study the microevolution of CDI in the hospital setting (154)

and to investigate localized transmission and international dissemination of major clinically important lineages such as RTs 027 (16) and 017 (159). But as outlined below, this approach has also been used to delineate cryptic transmission pathways of *C. difficile* between animals, humans, and their shared environment. In doing so, these genomic studies have redefined our understanding of the ecology and evolution of this complex species.

C. difficile RT078

C. difficile RT078 belongs to evolutionary clade 5 and is the principal ST11 sublineage (**Figure 2**). Between 2005 and 2008, RT078 rose from 11th to become the 3rd most frequently encountered RT in European hospitals (25), an increase particularly evident in the Netherlands where, from 2005 to 2008, Dutch hospitals would see the total prevalence of RT078-associated cases increase from 3 to 13% (32). These RT078 cases of CDI were in younger patients and with community-onset (32, 33). Comparable rates have been found in North America (21, 24, 35) with one study reporting 46% of all RT078 isolates were community acquired (160). As with many toxigenic *C. difficile* RTs, RT078 can be carried asymptomatically (161, 162). *C. difficile* RT078 has established significant reservoirs in North American, European, and Asian pigs and cattle and is often reported as the dominant type irrespective of age, diarrheal status or other farm-specific factors (37, 127). In an important Dutch study of *C. difficile* spore acquisition, Hopman et al. (68) demonstrated that piglets delivered by cesarean-section were *C. difficile*-negative yet were rapidly colonized with *C. difficile* RT078 spores within 48 h.

The virulence potential of RT078 has been likened to that of epidemic RT027 with which it shares similar genetic features. These include the major virulence genes *tcdA*, *tcdB,* and *cdtA/B* involved in toxin production, and an aberrant toxin regulator gene *tcdC* (deletions, nonsense mutations, and premature stop codons) leading to a reduction in log phase repression of toxin expression. The role for the latter in the observed hyper-virulent disease phenotype seen also in RT078 infections i.e., more toxin, increased mortality and morbidity, remains speculative (32, 163–167). *C. difficile* RT078 strains are often multidrug-resistant (MDR) (161, 168) and, compared to other RTs, including RT027, show remarkable resilience to extremes of temperature (80 to 96°C) and water treatment processes (142, 143, 145). It has also been proposed that the emergence and global dissemination of RT078 in humans is linked to an enhanced ability to metabolize the food additive trehalose (169). These virulence and survival traits may explain the successful dissemination of this lineage in production animals and humans worldwide. Unsurprisingly, it has received major attention as a potentially zoonotic lineage.

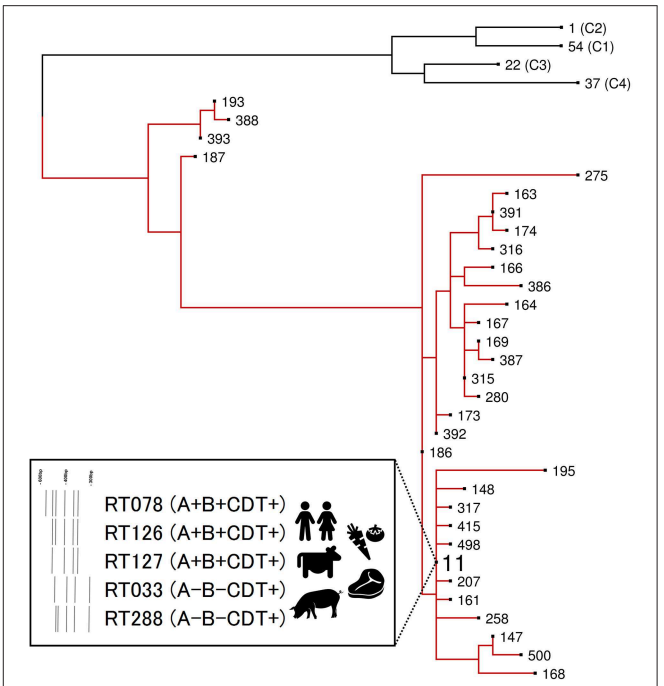


FIGURE 2 | Evolutionary clade 5. Maximum-likelihood phylogeny based on concatenated MLST allele sequences (seven loci, 3,501 bp) for 32 known clade 5 STs. For global phylogenetic context, well-characterized representatives of major MLST clades C1 (ST54, RT012), C2 (ST1, RT027), C3 (ST22, RT023), and C4 (ST37, RT017) are also shown. Branches for clade 5 are shown in red. The inset box highlights major ST11 RT sublineages of clinical and agricultural/livestock importance with associated toxin gene profiles and graphical representation of 16s−23s rRNA PCR ribotyping banding patterns.

Zoonotic Transmission of *C. difficile* RT078 Between Humans and Animals

Genetic studies using MLST, MLVA, Rep-PCR and AFLP fingerprinting have all provided significantly higher strain resolution of RT078 populations compared to conventional PCR ribotyping. In 2010, Bakker et al. (170) found 85% of RT078 isolates of human and porcine origin in the Netherlands were genetically related and, in many instances, indistinguishable by high-resolution MLVA. In 2012, Stabler et al. (171) used MLST to analyse 385 *C. difficile* isolates from different geographical locations (Europe, North America, and Australia) and sources (human, food, and animal). Strains of RT078 from humans, food and animals, some from different countries and continents, were indistinguishable (all sharing seven identical housekeeping genes, ST11) (171). More recent work from Taiwan showed RT078 isolated from pig farms shared identical Rep-PCR fingerprints as RT078 strains derived from humans with CDI in hospitals in the same region (64). Similarly, in Spain, RT078 of human and animal origin were clustered together by AFLP (172). Evidence from Japan suggests RT078 has been introduced from Europe. Usai et al. found Japanese pig RT078 strains clustered (by MLVA) with European human and pig RT078 strains (86), and Niwa et al. found a single MLVA cluster of RT078 responsible for five cases of colitis in Japanese racehorses (173). Both pigs and racehorses are internationally traded in Japan; thus, RT078 may have been imported into Japan from Europe via live animals.

Natural and diverse reservoirs of RT078 support the hypothesis that CDI may have a zoonotic origin. To date, a few key WGS-based studies have led to significant advancement in understanding the true zoonotic potential and evolution of the RT078 and its close relatives. In 2013, Knetsch et al. (161)

used core genome SNV typing to compare 65 *C. difficile* RT078 isolates of human and porcine origin sourced over a 10-year period in The Netherlands. Using Bayesian techniques, an RT078 population-specific mutation rate was estimated to be 2.72×10^{-7} substitutions per site per year, equating to around 1 SNV per genome per year—a figure comparable with earlier estimates (154, 155). A core genome phylogeny showed isolates of human and porcine origin clustering together. Notably, the analysis showed a pair of human and pig isolates from the same pig farm in The Netherlands to be indistinguishable (zero SNVs difference in their core genome). Working in pig husbandry or living in (or visiting) areas with a high density of pigs increased the risk of acquiring *C. difficile* due to exposure to pig feces (161). Whilst the transmission of RT078 between a pig and pig farmer within the confines of a pig-rearing facility might not be that surprising, it was nonetheless the first ever confirmation that interspecies transmission of *C. difficile* had occurred (161). The exact mode of transmission between these species remains unclear. Whilst these data appear to support the theory that CDI is a zoonosis, a common environmental source, asymptomatic carriage and/or zooanthroponotic (human to animal) transmission cannot be ruled out.

In 2017, the same authors (174) extended these findings. They investigated microevolution in the core genome of 248 *C. difficile* RT078 strains sourced from humans and animals in 22 countries. This study provided the first estimate of the global RT078 population structure and yielded new insights into the potential and extent for zoonotic spread. Extensive clustering of *C. difficile* RT078 from human clinical cases and food animals was observed, with clear instances of interspecies clonal transmission, only this time, the significant clustering of clones supported evidence of bidirectional spread of *C. difficile* RT078 between production animals and humans. Moreover, there was only limited geographic clustering with clones of *C. difficile* RT078 spread multiple times across multiple towns, countries and continents, in particular between North America and Europe: one example was the transmission of an RT078 clone between an animal in Canada and humans in the United Kingdom. This indicated interspecies transmission of *C. difficile* RT078 was not restricted to a local population of humans and production animals, as previously shown in the 2014 Dutch study. Together, these data revealed a highly linked intercontinental transmission network of *C. difficile* RT078 between humans and animals and provided further evidence that CDI has a significant zoonotic component (174). Yet it also showed that, in contrast to another classic enteric pathogen *Salmonella enterica* which has distinct animal- and human-associated populations, *C. difficile* RT078 appeared to be a clonal population moving frequently (and likely over long time periods) between production animal and human hosts, with no geographical constraints.

ST11 Is a Heterogeneous Lineage of Major One Health Importance

ST11 is an ancient evolutionary lineage comprising at least a dozen CDT^+ ribotypes that cause CDI in humans with significant ecological niches in production animals worldwide (175) (**Figure 2**). As is apparent, and for good reasons, there has been a strong focus on the ST11 sub-lineage RT078, however, until recently, little was known about the evolutionary history and zoonotic potential of other ST11 RTs. Our recent study (175) addressed this knowledge gap, using WGS to investigate population structure and clonal transmission in over 200 strains of major ST11 RTs 078, 126, 127, 033, and 288 sourced from human and veterinary/environmental origin across Australia, Asia, Europe, and North America. A core genome phylogeny showed the global ST11 population structure largely mirrored RT sub-lineage, with discrete evolutionary clusters congruent with RTs 078/126, 127, 033/288. Core genome SNV analysis found multiple instances of inter- and intra-species clonal transmission in all RT sub-lineages. Interspecies clonal groups comprised *C. difficile* isolates derived from health care facilities and farm animals spread across different states, countries, and continents, often without any healthcare association. Our findings independently confirm and extend the work of Knetsch et al. (161, 174) revealing a globally-disseminated network of *C. difficile* ST11 clones with the capability and proclivity for reciprocal zoonotic and/or anthroponotic transmission. Moreover, this study showed for the first time that non-RT078 ST11 strains such as RTs 126, 127, 033, and 288 also display a high zoonotic potential and should also be considered lineages of emerging One Health importance.

Antimicrobial Resistance and ST11 Evolution

Antimicrobials are a crucial component in the pathogenesis of CDI; they play a central role in the establishment of infection and, paradoxically, remain the preferred option for treatment (176).

AMR is, therefore, a key factor driving epidemiological changes in CDI (1). As we have seen with virulent *C. difficile* RT027 epidemic lineage, outbreaks emerge when the inherent resistance of *C. difficile* to cephalosporins is combined with acquired resistance to high-risk antimicrobials known to incite CDI, such as fluoroquinolones (16). In all the above WGS-based studies of RT078 and ST11, substantial AMR repertoires were identified. In the Dutch study (161), interspersed throughout the RT078 phylogeny were clones common to humans and livestock harboring identical mobile genetic elements (MGEs) conferring resistance to streptomycin (Tn*6235*, *aphA1*$^+$) and tetracycline (Tn*6190*, *tetM*$^+$) (161). In the later study by Knetsch et al. (174), the global population of RT078 contained a broad array of AMR genes encoding resistance to aminoglycosides and streptothricin (*aph3′-III, ant6′-Ib, Sat4A*), erythromycin (*ermB*$^+$), and tetracycline (*tetM, tetO, tet32, tet40, tet44*). The gene *cdeA* encoding a multidrug efflux transporter was found in all isolates (174).

In our ST11 study (175), half of all strains showed phenotypic resistance to one or more of tetracycline, moxifloxacin, erythromycin, and clindamycin, of which a quarter, predominantly RTs 126/078, were resistant to $\geq$3 of these agents. Underscoring this resistance was an array of AMR genetic loci including chromosomal mutations in *gyrA/B* (fluoroquinolone resistance) and MGEs conferring resistance to

macrolides and lincosamides (Tn*6194*; *ermB*$^+$), and tetracycline (Tn*6190*; *tetM*$^+$ and Tn*6164*; *tet44*$^+$), the latter a 106 kb genetic island apparently specific to RT078 (177). This was the first such report of Tn*6194* from animals in the world. This element is the most common *ermB*-containing element found in human clinical isolates in Europe and is a defining genetic feature of epidemic RT027 (16, 178, 179). A phenotypically silent *vanB2* transposon (likely from *Enterococcus faecalis*) was also found in a *C. difficile* RT033 strain isolated from an Australian veal calf at slaughter (180). Another common ruminant species *Erysipelothrix rhusiopathiae* appeared to be the origin of the numerous aminoglycoside resistance gene clusters present in all ST11 sub-lineages.

In a compelling new study, Dingle et al. (181) present a strong case for antimicrobial selection influencing the recent evolutionary history of *C. difficile* RT078. A time-scaled phylogeny built from the core genome of over 400 international *C. difficile* RT078 strains revealed three major clonal expansions (a rapid, recent international spread of RT078 clones). Two-thirds of all RT078 were tetracycline resistant. Remarkably, a common ancestor of each clonal expansion had independently evolved tetracycline resistance via the acquisition of distinct *tetM* alleles carried on closely related Tn*916-like* elements, an analogous situation to the emergence of fluoroquinolone resistance in RT027 (16). The parallel *tetM* associated clonal expansions were estimated to have occurred sometime around the year 2000, at a time when the number of RT078-associated clinical cases (at least in Europe) started to increase. Moreover, the three *tetM* alleles show significant homology (97–100% sequence identity) with *tetM* genes belonging to established zoonotic species such as *E. faecalis*, *Escherichia coli*, and *Streptococcus suis*—further supporting an agricultural origin for RT078. The authors note that *S. suis* has striking parallels with *C. difficile* RT078—it is a globally disseminated human pathogen which has established substantial reservoirs in pigs and has displayed recent increases in tetracycline resistance (182, 183). In summary, these phylogenetic data are consistent with an evolutionary response to tetracycline selective pressure. The inappropriate and overuse of tetracycline in animal husbandry is well-recognized (184). This selective pressure, combined with the rapid, international spread of *C. difficile* RT078 via the food chain and other agricultural vectors provides a plausible explanation for the clinical prominence of this lineage in humans.

Interspecies Transfer of *C. difficile* RT014 Between Humans and Animals

C. difficile RT014 is a toxigenic ($A^+B^+CDT^-$) and highly successful lineage of *C. difficile* belonging to MLST clade 1. RT014 is consistently among the most common RTs causing CDI in European healthcare systems, and in Australia it has been the most prevalent RT causing human infection for many years, accounting for ∼25% of all CDI cases (10, 185–188). The zoonotic potential of this RT was initially thought to be quite low as its prevalence in production animals in Europe was low and it was absent from livestock in Asia. In Australia, there was a completely different and intriguing story. In 2013, a nationwide cross-sectional study of *C. difficile* in 21 pig farms in Australia found RT014 to be the most prevalent RT in neonatal pigs aged <14 days, accounting for 23% ($n = 26/154$) of isolates (52). With rates of CDI in Australia increasing markedly in recent years (24% in 2011–2012 alone) and a significant rise in CA-CDI (26), the establishment of significant RT014 reservoirs in porcine populations in Australia suggests zoonotic transmission as a plausible source of human infection.

To examine the true extent of genetic relatedness, a collection of 40 contemporaneous isolates of RT014 of human and porcine origin in Australia were subjected to WGS (128). A total of three distinct STs were identified in this RT014 collection (STs 2, 13, and 49), and in each, human and porcine populations were intermingled, signaling a very recent shared ancestry. A phylogeny based on evolution in 1,260 core orthologous genes (1,019,160 bp, ∼25% of bases in an average *C. difficile* genome) showed geographically and temporally unconstrained clustering of human and animal *C. difficile* RT014 strains in all three STs again supporting a close genetic relationship. Finally, a phylogeny-based on evolution in non-recombinant 1,287 core genome SNVs provided ultra-fine scale resolution of the RT014 population, identifying multiple instances of plausible interspecies clonal transmission. In total, 42% of *C. difficile* RT014 strains from humans with CDI showed a clonal relationship (differing by no more than two SNVs in their core genome) with one or more RT014 strains derived from pigs. Remarkably, many RT014 clones originated from pigs and humans in states separated by thousands of kilometers, collected many months apart, and half of the human isolates in these clonal groups originated from cases classified as CA-CDI, representing the acquisition of CDI outside of the hospital system (**Figure 3**). Long range transmission of *C. difficile* RT014 clones suggests direct contact between humans and colonized livestock is perhaps unlikely, and there was no evidence here. Given what we know of the *C. difficile* colonization-transmission cycle in the farrowing environment and wider livestock industry, it is conceivable that over an extended period there has been frequent long-range indirect interspecies transmission through human exposure to contaminated retail meat but more likely contaminated piggery by-products such as manure and compost in the community setting (**Figure 3**). Indeed, genomic studies from the USA and Europe have shown that the household environment and pet dogs are colonized with *C. difficile* RT014/ST2 representing reservoirs of RT014 in the community (124, 125).

The *C. difficile* Pan-Genome: Insights Into the Ecology of a Complex Pathogen

A bacterial pan-genome describes the full complement of genes in a species or individual phylogenetic lineage. It comprises a core component (those genes present in all strains) and an accessory or adaptive component (genes absent from one or more strain or unique to a particular strain) (189). Early microarray-based studies estimated the *C. difficile* pan-genome to be comprised of 9,640 coding sequences (CDS) with a core genome component many orders of magnitude lower at 600–3,000 CDS (190–192).

A

C. difficile RT014 clonal group ST13

Origin	ST	TP	State	Site	Date	CDI
Human	13	A+B+CDT-	NSW	HCF 1	Nov-12	HA-CDI
Human	13	A+B+CDT-	SA	HCF 2	Dec-12	HA-CDI
Human	13	A+B+CDT-	NSW	HCF 3	Mar-13	HA-CDI
Pig	13	A+B+CDT-	NSW	Farm 1	Apr-12	
Pig	13	A+B+CDT-	NSW	Farm 1	Apr-12	
Pig	13	A+B+CDT-	NSW	Farm 1	Apr-12	
Pig	13	A+B+CDT-	SA	Farm 2	Jun-12	
Pig	13	A+B+CDT-	NSW	Farm 3	Apr-12	

C. difficile RT014 clonal group ST2

Origin	ST	TP	State	Site	Date	CDI
Human	2	A+B+CDT-	SA	HCF 2	Apr-13	CA-CDI
Human	2	A+B+CDT-	VIC	HCF 4	Nov-12	CA-CDI
Human	2	A+B+CDT-	VIC	HCF 4	Nov-12	HA-CDI
Human	2	A+B+CDT-	VIC	HCF 4	Nov-12	HA-CDI
Human	2	A+B+CDT-	NSW	HCF 5	Nov-12	CA-CDI
Human	2	A+B+CDT-	QLD	HCF 6	Jan-13	CA-CDI
Human	2	A+B+CDT-	NSW	HCF 7	Dec-12	INDET
Human	2	A+B+CDT-	QLD	HCF 8	Jan-13	CA-CDI
Human	2	A+B+CDT-	QLD	HCF 9	Nov-12	CA-CDI
Pig	2	A+B+CDT-	SA	Farm 2	Jun-12	

B

67% of neonatal pigs across 5 states in Australia were positive for *C. difficile* with **RT014 comprising 23% of isolates**[3]

~25% of CDI cases in Australia are caused by *C. difficile* **RT014**[1,2]

30% of root vegetables from markets in metro area of WA were positive for *C. difficile* with **RT014 comprising 7% of isolates**[5]

59% of new lawns in metro areas of Western Australia were positive for *C. difficile* with **RT014 comprising 39% of isolates**[4]

27% of retail compost in metro areas of Western Australia were positive for *C. difficile* with **RT014 comprising 10% of isolates**[6]

FIGURE 3 | Transmission networks and community reservoirs of *C. difficile* RT014 in Australia. **(A)** summary of ST13 ($n = 8$) and ST2 (n = 10) RT014 clonal groups found in pigs and humans with CDI in Australia, adapted from Knight et al. (128). A clonal group is defined as two or more strains differing by <2 SNVs in their core genome. HCF, healthcare facility; NSW, New South Wales; SA, South Australia; VIC, Victoria; QLD, Queensland; INDET, indeterminate. **(B)** summary of RT014 ecological niches in Australia. [1]Knight et al. (186); [2]Collins et al. (188); [3]Knight et al. (52); [4]Moono et al. (151); [5]Lim et al. (84); [6]Lim et al. (122).

More recent WGS based studies of RT014 (128), RT078 (174), and ST11 (175) from humans and animals have provided further insights into the genetic diversity, plasticity and ecology of zoonotic *C. difficile* lineages.

Analysis of 44 Australian RT014 genomes (STs 2, 13, and 49) revealed a large pangenome (7,587 genes) comprising a core genome of 2,296 genes (30.3% of the total gene repertoire) and an accessory genome of 5,291 genes (128). Moreover, the human and porcine populations shared near identical proteomes (128). The global RT078 population (248 genomes from four continents) possessed a large pangenome of 6,239 genes with a core genome of 3,368 genes (53.9% of the total gene repertoire) and an accessory genome of 2,871 genes (174). Finally, the global ST11 population (207 genomes from four continents including RTs 078, 126, 127, 033, and 288) was defined by a massive pan-genome (10,378 genes), a remarkably small core genome of 2,058 genes (only 19.8% of the gene pool) and an accessory genome of 8,320 genes (175). In the case of RT014 and ST11, power-law regression analysis determined the pangenomes to be "open," that is, size increases indefinitely when adding new genomes. For example, in the ST11 analysis, after sequencing over 200 genomes there is an average of 16 new genes contributed to the gene pool with each additional sequenced strain (175).

The size and openness of a pan-genome is also a very useful proxy for characterizing the lifestyle of a bacterial species (193). The pan-genome data derived from these zoonotic and agricultural-associated *C. difficile* lineages predict a species with a sympatric lifestyle, occupying niches in extremely diverse environments that are enriched with mixed microbial communities of prokarya and archea (193). This is true of *C. difficile*, a versatile species which shows extraordinary adaption to multiple ecosystems including the gastrointestinal tract of multiple mammalian hosts, and several secondary habitats such as water, soil, and composts and invertebrate species (179). In contrast to allopatric and intracellular species such as *Rickettsia rickettsii* and *Chlamydia trachomatis*, which have small closed pan-genomes and live in isolated niches with limited exchange with the global microbial gene pool, sympatric

species like *C. difficile* (and *C. botulinum*) have larger, more complex open pan-genomes. Sympatry also means a higher frequency of gene acquisition events and a higher probability of acquiring parasitic DNA i.e., transposons and bacteriophages, both contributing to an increase in pan-genome size (193, 194). Indeed, underscoring the substantial genetic diversity in these zoonotic *C. difficile* lineages were large and diverse collections of clinically important prophages of the *Siphoviridae* and *Myoviridae* (128, 175) and AMR genetic elements (128, 174, 175). As corroborated by Dingle et al. in RT078 (181), many of these underlying AMR elements show evolutionary origins in commensal species residing within the gut of farm animals. Examples being macrolide resistance genes from *Campylobacter coli* (cryptic), aminoglycoside, and streptothricin genes cassettes from *E. rhusiopathiae,* and a plethora of tetracycline resistance genes from *S. suis, E. faecalis, Megasphaera elsdenii*, *C. jejuni, and C. perfringens* (128, 161, 174, 175). Moreover, AMR elements Tn*6194* (*ermB*$^+$) and Tn*5397* (*tetM*$^+$) are capable of intra-species transfer to different *C. difficile* RTs and even inter-species transfer to other genera (16, 191, 195).

Together, the phylogenetic, pangenome, and AMR data show that these zoonotic *C. difficile* lineages have the capability and propensity to move between humans, production animals, and their shared environment. By occupying niches within multiple host species, these *C. difficile* lineages are able to access and exchange DNA with an enormously diverse metagenome, particularly the ruminant gut and soil microbiota. Such promiscuous behavior provides *C. difficile* with a potential selective advantage over taxa inhabiting the same gut ecosystem, be it the pig, cow or human intestinal tract, therefore greatly enhancing their ability to adapt to fluctuating environmental factors and their likelihood of success.

Finally, in the case of ST11, it is remarkable that even after sampling >200 ST11 strains from over a dozen unique RT sub-lineages spread over four different continents; the complete gene complement of this lineage was not captured (175). With over 420 STs and >600 RTs currently recognized, it is likely that the complete species pan-genome for *C. difficile* could be astonishingly high. Such enormous diversity is more typical for phylogenetic distances between genera within a family, rather than strains within a species (179). In light of recent calls for taxonomic revisions (196–199), it is possible that *C. difficile* may, in fact, be a complex of sub-species divided along the major evolutionary clades.

FUTURE DIRECTIONS AND CHALLENGES

The One Health paradigm is a philosophical approach to improving and safeguarding the health of humans, animals and the environment and, importantly, recognizes that these three areas are inter-related (200). Specifically, improved treatment of disease common to humans and animals can be achieved through the application of interdisciplinary approaches between human and veterinary medicine, and the analysis of environment-derived isolate datasets. In this regard, CDI is the quintessential One Health issue (141). As we have highlighted here, the application of high-resolution microbial genomics in a One Health framework is the optimal paradigm for advancing our understanding of CDI in humans and animals. Together, this body of evidence challenges the existing paradigm and long-held conception that CDI is primarily a healthcare-associated infection and provides compelling evidence that CDI has a significant zoonotic component. More important, these findings should stimulate new discussions about One Health focused interventions for CDI.

Collaboration between human and veterinary medicine will be essential if we are to safeguard the health of humans and production animals (141). First and foremost, measures which reduce the levels of *C. difficile* spores in the piggery environment are of paramount importance, not only for mitigating the risk of community acquisition but also for improving animal health (141). In human medicine, these measures comprise stringent infection control policies such as case isolation, reduced use of late-generation cephalosporins, hand hygiene and deep environmental cleaning (201, 202). Analogous interventions have been employed in the veterinary hospital setting with a significant reduction in CDI cases (203); however, the vast scale of modern production animal systems may hinder successful implementation. Also, the frequent disagreement between clinicians, veterinarians and the livestock industry regarding appropriate risk management of *C. difficile* in animal populations remains an additional, significant hurdle to overcome (141, 204).

With several candidate *C. difficile* vaccines in development (205), immunization of livestock could be a highly effective way to reduce the overall prevalence of *C. difficile* and is a good example of an integrative One Health approach to tackling CDI (141). Finally, continued genetic and phenotypic surveillance of *C. difficile* is critical to an enhanced understanding of epidemiological and genetic factors contributing to the emergence, evolution, and spread of CDI (152, 179). Crucially, if we are to identify improved infection prevention and control strategies, and public health interventions designed to mitigate the risk of *C. difficile* transmission, it is imperative that such studies should have a strong One Health focus by including analysis of *C. difficile* strains derived from humans, animals and food, and their shared environment. As much of the focus to date has been on the ST11 group and RT014, future studies should examine the potential for clonal relationships between other lineages circulating in clinical and animal/environmental settings. As illustrated by the studies highlighted in this review, WGS will play a central role in this, providing a level of discrimination far beyond that achievable by conventional molecular typing methodologies.

AUTHOR CONTRIBUTIONS

All authors listed have made equal intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2019.00164/full#supplementary-material

REFERENCES

1. Larson HE, Price AB, Honour P, Borriello SP. *Clostridium difficile* and the aetiology of pseudomembranous colitis. *Lancet*. (1978) 311:1063–6. doi: 10.1016/S0140-6736(78)90912-1
2. George RH, Symonds JM, Dimock F, Brown JD, Arabi Y, Shinagawa N, et al. Identification of *Clostridium difficile* as a cause of pseudomembranous colitis. *Br Med J*. (1978) 1:695. doi: 10.1136/bmj.1.6114.695
3. Chang T-W, Bartlett JG, Gorbach SL, Onderdonk AB. Clindamycin-induced enterocolitis in hamsters as a model of pseudomembranous colitis in patients. *Infect Immun*. (1978) 20:526–9.
4. George WL, Goldstein EC, Sutter V, Ludwig S, Finegold S. Aetiology of antimicrobial-agent-associated colitis. *Lancet*. (1978) 311:802–3. doi: 10.1016/S0140-6736(78)93001-5
5. Leffler DA, Lamont JT. Editorial: not so nosocomial anymore: the growing threat of community-acquired *Clostridium difficile*. *Am J Gastroenterol*. (2012) 107:96–8. doi: 10.1038/ajg.2011.404
6. Sheridan PP, Freeman KH, Brenchley JE. Estimated minimal divergence times of the major bacterial and archaeal phyla. *Geomicrobiol J*. (2003) 20:1–14. doi: 10.1080/01490450303891
7. Lawson PA, Citron DM, Tyrrell KL, Finegold SM. Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prevot 1938. *Anaerobe*. (2016) 40:95–9. doi: 10.1016/j.anaerobe.2016.06.008
8. Oren A, Rupnik M. *Clostridium difficile* and *Clostridioides difficile*: two validly published and correct names. *Anaerobe*. (2018) 52:125–6. doi: 10.1016/j.anaerobe.2018.07.005
9. Heymann DL, Dar OA. Prevention is better than cure for emerging infectious diseases. *BMJ*. (2014) 348:g1499. doi: 10.1136/bmj.g1499
10. Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, Dunn JR, et al. Burden of *Clostridium difficile* infection in the United States. *N Engl J Med*. (2015) 372:825–34. doi: 10.1056/NEJMoa1408913
11. Rodriguez Diaz C, Seyboldt C, Rupnik M. Non-human *C. difficile* reservoirs and sources: animals, food, environment. *Adv Exp Med Biol*. (2018) 1050:227–43. doi: 10.1007/978-3-319-72799-8_13
12. Britton RA, Young VB. Role of the intestinal microbiota in resistance to colonization by *Clostridium difficile*. *Gastroenterol*. (2014) 146:1547–53. doi: 10.1053/j.gastro.2014.01.059
13. Carter GP, Rood JI, Lyras D. The role of toxin A and toxin B in the virulence of *Clostridium difficile*. *Trends Microbiol*. (2012) 20:21–9. doi: 10.1016/j.tim.2011.11.003
14. Chandrasekaran R, Lacy DB. The role of toxins in *Clostridium difficile* infection. *FEMS Microbiol Rev*. (2017) 41:723–50. doi: 10.1093/femsre/fux048
15. Rodriguez-Palacios A, Stampfli HR, Duffield T, Peregrine AS, Trotz-Williams LA, Arroyo LG, et al. *Clostridium difficile* PCR ribotypes in calves, Canada. *Emerg Infect Dis*. (2006) 12:1730–6. doi: 10.3201/eid1211.051581
16. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, et al. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet*. (2013) 45:109–13. doi: 10.1038/ng.2478
17. McDonald LC, Coignard B, Dubberke E, Song X, Horan T, Kutty PK. Recommendations for surveillance of *Clostridium difficile*–associated disease. *Infect Control Hosp Epidemiol*. (2007) 28:140–5. doi: 10.1086/511798
18. Kuntz JL, Chrischilles EA, Pendergast JF, Herwaldt LA, Polgreen PM. Incidence of and risk factors for community-associated *Clostridium difficile* infection: a nested case-control study. *BMC Infect Dis*. (2011) 11:194. doi: 10.1186/1471-2334-11-194
19. Khanna S, Pardi DS, Aronson SL, Kammer PP, Baddour LM. Outcomes in community-acquired *Clostridium difficile* infection. *Aliment Pharmacol Ther*. (2012) 35:613–8. doi: 10.1111/j.1365-2036.2011.04984.x
20. Bloomfield LE, Riley TV. Epidemiology and risk factors for community-associated *Clostridium difficile* infection: a narrative review. *Infect Dis Ther*. (2016) 5:231–51. doi: 10.1007/s40121-016-0117-y
21. Khanna S, Pardi DS. Community-acquired *Clostridium difficile* infection: an emerging entity. *Clin Infect Dis*. (2012) 55:1741–2. doi: 10.1093/cid/cis722
22. Naggie S, Frederick J, Pien BC, Miller BA, Provenzale DT, Goldberg KC, et al. Community-associated *Clostridium difficile* infection: experience of a veteran affairs medical center in southeastern USA. *Infection*. (2010) 38:297–300. doi: 10.1007/s15010-010-0025-0
23. Kutty PK, Woods CW, Sena AC, Benoit SR, Naggie S, Frederick J, et al. Risk factors for and estimated incidence of community-associated *Clostridium difficile* infection, North Carolina, USA. *Emerg Infect Dis*. (2010) 16:197–204. doi: 10.3201/eid1602.090953
24. Khanna S, Pardi DS, Aronson SL, Kammer PP, Orenstein R, St Sauver JL, et al. The epidemiology of community-acquired *Clostridium difficile* infection: a population-based study. *Am J Gastroenterol*. (2012) 107:89–95. doi: 10.1038/ajg.2011.398
25. Bauer MP, Notermans DW, van Benthem BH, Brazier JS, Wilcox MH, Rupnik M, et al. *Clostridium difficile* infection in Europe: a hospital-based survey. *Lancet*. (2011) 377:63–73. doi: 10.1016/S0140-6736(10)61266-4
26. Slimings C, Armstrong P, Beckingham WD, Bull AL, Hall L, Kennedy KJ, et al. Increasing incidence of *Clostridium difficile* infection, Australia, 2011–2012. *Med J Aust*. (2014) 200:272–6. doi: 10.5694/mja13.11153
27. Mitchell B, Wilson F, McGregor A. An increase in community onset *Clostridium difficile* infection: a population based study. *Healthcare Infect*. (2012) 17:127–32. doi: 10.1071/HI12029
28. Eyre DW, Tracey L, Elliott B, Slimings C, Huntington PG, Stuart RL, et al. Emergence and spread of predominantly community-onset *Clostridium difficile* PCR ribotype 244 infection in Australia, 2010 to 2012. *Euro Surveill*. (2015) 20:21059. doi: 10.2807/1560-7917.ES2015.20.10.21059
29. Furuya-Kanamori L, Stone JC, Clark J, McKenzie SJ, Yakob L, Paterson DL, et al. Comorbidities, exposure to medications, and the risk of community-acquired *Clostridium difficile* infection: a systematic review and meta-analysis. *Infect Control Hosp Epidemiol*. (2015) 36:132–41. doi: 10.1017/ice.2014.39
30. Sandora TJ, Fung M, Flaherty K, Helsing L, Scanlon P, Potter-Bynoe G, et al. Epidemiology and risk factors for *Clostridium difficile* infection in children. *Pediatr Infect Dis J*. (2011) 30:580–4. doi: 10.1097/INF.0b013e31820bfb29
31. Bignardi GE, Settle C. Different ribotypes in community-acquired *Clostridium difficile*. *J Hosp Infect*. (2008) 70:96–8. doi: 10.1016/j.jhin.2008.04.003
32. Goorhuis A, Bakker D, Corver J, Debast SB, Harmanus C, Notermans DW, et al. Emergence of *Clostridium difficile* infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. *Clin Infect Dis*. (2008) 47:1162–70. doi: 10.1086/592257
33. Patterson L, Wilcox MH, Fawley WN, Verlander NQ, Geoghegan L, Patel BC, et al. Morbidity and mortality associated with *Clostridium difficile* ribotype 078: a case–case study. *J Hosp Infect*. (2012) 82:125–8. doi: 10.1016/j.jhin.2012.07.011
34. Dumyati G, Stevens V, Hannett GE, Thompson AD, Long C, Maccannell D, et al. Community-associated *Clostridium difficile* infections, Monroe County, New York, USA. *Clin Infect Dis*. (2012) 18:392–400. doi: 10.3201/eid1803.102023
35. Gerding DN, Lessa FC. The epidemiology of *Clostridium difficile* infection inside and outside health care institutions. *Infect Dis Clin North Am*. (2015) 29:37–50. doi: 10.1016/j.idc.2014.11.004
36. Moono P, Foster NF, Hampson DJ, Knight DR, Bloomfield LE, Riley TV. *Clostridium difficile* infection in production animals and avian species: a review. *Foodborne Pathog Dis*. (2016) 13:647–55. doi: 10.1089/fpd.2016.2181
37. Candel-Perez C, Ros-Berruezo G, Martinez-Gracia C. A review of Clostridioides [Clostridium] difficile occurrence through the food chain. *Food Microbiol*. (2019) 77:118–29. doi: 10.1016/j.fm.2018.08.012
38. Knight DR, Thean S, Putsathit P, Fenwick S, Riley TV. Cross-sectional study reveals high prevalence of *Clostridium difficile* non-PCR ribotype 078 strains in Australian veal calves at slaughter. *Appl Environ Microbiol*. (2013) 79:2630–5. doi: 10.1128/AEM.03951-12
39. Zidaric V, Pardon B, Dos Vultos T, Deprez P, Brouwer MS, Roberts AP, et al. Multiclonal presence of *Clostridium difficile* PCR ribotypes 078, 126 and 033 within a single calf farm is associated with differences in antibiotic resistance and sporulation properties. *Appl Environ Microbiol*. (2012) 78:8515–22. doi: 10.1128/AEM.02185-12
40. Costa MC, Stampfli HR, Arroyo LG, Pearl DL, Weese JS. Epidemiology of *Clostridium difficile* on a veal farm: prevalence, molecular characterization and tetracycline resistance. *Vet Microbiol*. (2011) 152:379–84. doi: 10.1016/j.vetmic.2011.05.014
41. Costa MC, Reid-Smith R, Gow S, Hannon SJ, Booker C, Rousseau J, et al. Prevalence and molecular characterization of *Clostridium difficile* isolated

from feedlot beef cattle upon arrival and mid-feeding period. *BMC Vet Res*. (2012) 8:38. doi: 10.1186/1746-6148-8-38
42. Schneeberg A, Neubauer H, Schmoock G, Grossmann E, Seyboldt C. Presence of *Clostridium difficile* PCR ribotype clusters related to 033, 078 and 045 in diarrhoeic calves in Germany. *J Med Microbiol*. (2013) 62:1190–8. doi: 10.1099/jmm.0.056473-0
43. Magistrali CF, Maresca C, Cucco L, Bano L, Drigo I, Filippini G, et al. Prevalence and risk factors associated with *Clostridium difficile* shedding in veal calves in Italy. *Anaerobe*. (2015) 33:42–7. doi: 10.1016/j.anaerobe.2015.01.010
44. Bandelj P, Blagus R, Briski F, Frlic O, Vergles Rataj A, Rupnik M, et al. Identification of risk factors influencing *Clostridium difficile* prevalence in middle-size dairy farms. *Vet Res.* (2016) 47:41. doi: 10.1186/s13567-016-0326-0
45. Romano V, Albanese F, Dumontet S, Krovacek K, Petrini O, Pasquale V. Prevalence and genotypic characterization of *Clostridium difficile* from ruminants in Switzerland. *Zoonoses Public Health*. (2012) 59:545–8. doi: 10.1111/j.1863-2378.2012.01540.x
46. Hammitt MC, Bueschel DM, Keel MK, Glock RD, Cuneo P, DeYoung DW, et al. A possible role for *Clostridium difficile* in the etiology of calf enteritis. *Vet Microbiol*. (2008) 127:343–52. doi: 10.1016/j.vetmic.2007.09.002
47. Houser BA, Soehnlen MK, Wolfgang DR, Lysczek HR, Burns CM, Jayarao BM. Prevalence of *Clostridium difficile* toxin genes in the feces of veal calves and incidence of ground veal contamination. *Foodborne Pathog Dis*. (2012) 9:32–6. doi: 10.1089/fpd.2011.0955
48. Rodriguez-Palacios A, Pickworth C, Loerch S, LeJeune JT. Transient fecal shedding and limited animal-to-animal transmission of *Clostridium difficile* by naturally infected finishing feedlot cattle. *Appl Environ Microbiol.* (2011) 77:3391–7. doi: 10.1128/AEM.02736-10
49. Rodriguez-Palacios A, Koohmaraie M, LeJeune JT. Prevalence, enumeration, and antimicrobial agent resistance of *Clostridium difficile* in cattle at harvest in the United States. *J Food Prot.* (2011) 74:1618–24. doi: 10.4315/0362-028X.JFP-11-141
50. Kalchayanand N, Arthur TM, Bosilevac JM, Brichta-Harhay DM, Shackelford SD, Wells JE, et al. Isolation and characterization of *Clostridium difficile* associated with beef cattle and commercially produced ground beef. *J Food Prot*. (2013) 76:256–64. doi: 10.4315/0362-028X.JFP-12-261
51. Squire MM, Carter GP, Mackin KE, Chakravorty A, Noren T, Elliott B, et al. Novel molecular type of *Clostridium difficile* in neonatal pigs, Western Australia. *Emerg Infect Dis*. (2013) 19:790–2. doi: 10.3201/eid1905.121062
52. Knight DR, Squire MM, Riley TV. Nationwide surveillance study of *Clostridium difficile* in Australian neonatal pigs shows high prevalence and heterogeneity of PCR ribotypes. *Appl Environ Microbiol*. (2014) 81:119–23. doi: 10.1128/AEM.03032-14
53. Rodriguez C, Avesani V, Van Broeck J, Taminiau B, Delmee M, Daube G. Presence of *Clostridium difficile* in pigs and cattle intestinal contents and carcass contamination at the slaughterhouse in Belgium. *Int J Food Microbiol.* (2013) 166:256–62. doi: 10.1016/j.ijfoodmicro.2013.07.017
54. Weese JS, Wakeford T, Reid-Smith R, Rousseau J, Friendship R. Longitudinal investigation of *Clostridium difficile* shedding in piglets. *Anaerobe*. (2010) 16:501–4. doi: 10.1016/j.anaerobe.2010.08.001
55. Goldova J, Malinova A, Indra A, Vitek L, Branny P, Jiraskova A. *Clostridium difficile* in piglets in the Czech Republic. *Folia Microbiol.* (2012) 57:159–61. doi: 10.1007/s12223-012-0102-0
56. Schneeberg A, Neubauer H, Schmoock G, Baier S, Harlizius J, Nienhoff H, et al. *Clostridium difficile* genotypes in German piglet populations. *J Clin Microbiol.* (2013) 51:3796–803. doi: 10.1128/JCM.01440-13
57. Doosti A, Mokhtari-Farsani A. Study of the frequency of *Clostridium difficile tcdA*, *tcdB*, *cdtA* and *cdtB* genes in feces of calves in south west of Iran. *Ann Clin Microbiol Antimicrob.* (2014) 13:21. doi: 10.1186/1476-0711-13-21
58. Asai T, Usui M, Hiki M, Kawanishi M, Nagai H, Sasaki Y. *Clostridium difficile* isolated from the fecal contents of swine in Japan. *J Vet Med Sci*. (2013) 75:539–41. doi: 10.1292/jvms.12-0353
59. Cho A, Byun JW, Kim JW, Oh SI, Lee MH, Kim HY. Low prevalence of *Clostridium difficile* in Slaughter pigs in Korea. *J Food Prot.* (2015) 78:1034–6. doi: 10.4315/0362-028X.JFP-14-493
60. Pirs T, Ocepek M, Rupnik M. Isolation of *Clostridium difficile* from food animals in Slovenia. *J Med Microbiol.* (2008) 57:790–2. doi: 10.1099/jmm.0.47669-0
61. Avbersek J, Janezic S, Pate M, Rupnik M, Zidaric V, Logar K, et al. Diversity of *Clostridium difficile* in pigs and other animals in Slovenia. *Anaerobe*. (2009) 15:252–5. doi: 10.1016/j.anaerobe.2009.07.004
62. Pelaez T, Alcala L, Blanco JL, Alvarez-Perez S, Marin M, Martin-Lopez A, et al. Characterization of swine isolates of *Clostridium difficile* in Spain: a potential source of epidemic multidrug resistant strains? *Anaerobe*. (2013) 22:45–9. doi: 10.1016/j.anaerobe.2013.05.009
63. Norén T, Johansson K, Unemo M. *Clostridium difficile* PCR ribotype 046 is common among neonatal pigs and humans in Sweden. *Clin Microbiol Infect*. (2014) 20:O2–O6. doi: 10.1111/1469-0691.12296
64. Tsai BY, Ko WC, Chen TH, Wu YC, Lan PH, Chen YH, et al. Zoonotic potential of the *Clostridium difficile* RT078 family in Taiwan. *Anaerobe*. (2016) 41:125–30. doi: 10.1016/j.anaerobe.2016.06.002
65. Debast SB, van Leengoed LA, Goorhuis A, Harmanus C, Kuijper EJ, Bergwerff AA. *Clostridium difficile* PCR ribotype 078 toxinotype V found in diarrhoeal pigs identical to isolates from affected humans. *Environ Microbiol.* (2009) 11:505–11. doi: 10.1111/j.1462-2920.2008.01790.x
66. Keessen EC, Leengoed LA, Bakker D, van den Brink KM, Kuijper EJ, Lipman LJA. Prevalence of *Clostridium difficile* in swine thought to have *Clostridium difficile* infections (CDI) in eleven swine operations in the Netherlands. *Tijdschr Diergeneeskd*. (2010) 135:134–7.
67. Hopman NEM, Oorburg D, Sanders I, Kuijper EJ, Lipman LJA. High occurrence of various *Clostridium difficile* PCR ribotypes in pigs arriving at the slaughterhouse. *Vet Q.* (2011) 31:179–81. doi: 10.1080/01652176.2011.649370
68. Hopman NE, Keessen EC, Harmanus C, Sanders IM, van Leengoed LAMG, Kuijper EJ, et al. Acquisition of *Clostridium difficile* by piglets. *Vet Microbiol.* (2011) 149:186–92. doi: 10.1016/j.vetmic.2010.10.013
69. Keessen EC, van den Berkt AJ, Haasjes NH, Hermanus C, Kuijper EJ, Lipman LJA. The relationship between farm specific factors and prevalence of *Clostridium difficile* in slaughter pigs. *Vet Microbiol.* (2011) 154:130–4. doi: 10.1016/j.vetmic.2011.06.032
70. Norman KN, Harvey RB, Scott HM, Hume ME, Andrews K, Brawley AD. Varied prevalence of *Clostridium difficile* in an integrated swine operation. *Anaerobe.* (2009) 15:256–60. doi: 10.1016/j.anaerobe.2009.09.006
71. Baker AA, Davis E, Rehberger T, Rosener D. Prevalence and diversity of toxigenic *Clostridium perfringens* and *Clostridium difficile* among swine herds in the midwest. *Appl Environ Microbiol.* (2010) 76:2961–7. doi: 10.1128/AEM.02459-09
72. Thakur S, Putnam M, Fry PR, Abley M, Gebreyes WA. Prevalence of antimicrobial resistance and association with toxin genes in *Clostridium difficile* in commercial swine. *Am J Vet Res.* (2010) 71:1189–94. doi: 10.2460/ajvr.71.10.1189
73. Norman KN, Scott HM, Harvey RB, Norby B, Hume ME, Andrews K. Prevalence and genotypic characteristics of *Clostridium difficile* in a closed and integrated human and swine population. *Appl Environ Microbiol.* (2011) 77:5755–60. doi: 10.1128/AEM.05007-11
74. Thitaram SN, Frank J, Lyon S, Siragusa G, Bailey J, Lombard J, et al. *Clostridium difficile* from healthy food animals: optimized isolation and prevalence. *J Food Prot.* (2011) 74:130–3. doi: 10.4315/0362-028X.JFP-10-229
75. Fry PR, Thakur S, Abley M, Gebreyes WA. Antimicrobial resistance, toxinotype, and genotypic profiling of *Clostridium difficile* isolates of swine origin. *J Clin Microbiol.* (2012) 50:2366–72. doi: 10.1128/JCM.06581-11
76. Susick EK, Putnam M, Bermudez DM, Thakur S. Longitudinal study comparing the dynamics of *Clostridium difficile* in conventional and antimicrobial free pigs at farm and slaughter. *Vet Microbiol*. (2012) 157:172–8. doi: 10.1016/j.vetmic.2011.12.017
77. Knight DR, Riley TV. Prevalence of gastrointestinal *Clostridium difficile* carriage in Australian sheep and lambs. *Appl Environ Microbiol.* (2013) 79:5689–92. doi: 10.1128/AEM.01888-13
78. Esfandiari Z, Weese JS, Ezzatpanah H, Chamani M, Shoaei P, Yaran M, et al. Isolation and characterization of *Clostridium difficile* in farm animals from slaughterhouse to retail stage in Isfahan, Iran. *Foodborne Pathog Dis*. (2015) 12:864–6. doi: 10.1089/fpd.2014.1910

79. Avbersek J, Pirs T, Pate M, Rupnik M, Ocepek M. *Clostridium difficile* in goats and sheep in Slovenia: characterisation of strains and evidence of age-related shedding. *Anaerobe*. (2014) 28:163–7. doi: 10.1016/j.anaerobe.2014.06.009
80. Zidaric V, Zemljic M, Janezic S, Kocuvan A, Rupnik M. High diversity of *Clostridium difficile* genotypes isolated from a single poultry farm producing replacement laying hens. *Anaerobe*. (2008) 14:325–7. doi: 10.1016/j.anaerobe.2008.10.001
81. Koene MGJ, Mevius D, Wagenaar JA, Harmanus C, Hensgens MPM, Meetsma AM, et al. *Clostridium difficile* in Dutch animals: their presence, characteristics and similarities with human isolates. *Clin Microbiol Infect*. (2012) 18:778–84. doi: 10.1111/j.1469-0691.2011.03651.x
82. Rodriguez-Palacios A, Barman T, LeJeune JT. Three-week summer period prevalence of *Clostridium difficile* in farm animals in a temperate region of the United States (Ohio). *Can Vet J*. (2014) 55:786–9.
83. Simango C, Mwakurudza S. *Clostridium difficile* in broiler chickens sold at market places in Zimbabwe and their antimicrobial susceptibility. *Int J Food Microbiol*. (2008) 124:268–70. doi: 10.1016/j.ijfoodmicro.2008.03.020
84. Lim SC, Foster NF, Elliott B, Riley TV. High prevalence of *Clostridium difficile* on retail root vegetables, Western Australia. *J Appl Microbiol*. (2017) 124:585–90. doi: 10.1111/jam.13653
85. Tkalec V, Janezic S, Skok B, Simonic T, Mesaric S, Vrabic T, et al. High *Clostridium difficile* contamination rates of domestic and imported potatoes compared to some other vegetables in Slovenia. *Food Microbiol*. (2019) 78:194–200. doi: 10.1016/j.fm.2018.10.017
86. Usui M, Nanbu Y, Oka K, Takahashi M, Inamatsu T, Asai T, et al. Genetic relatedness between Japanese and European isolates of *Clostridium difficile* originating from piglets and their risk associated with human health. *Front Microbiol*. (2014) 5:513. doi: 10.3389/fmicb.2014.00513
87. Rodriguez-Palacios A, Staempfli HR, Duffield T, Weese JS. *Clostridium difficile* in retail ground meat, Canada. *Emerg Infect Dis*. (2007) 13:485–7. doi: 10.3201/eid1303.060988
88. De Boer E, Zwartkruis-Nahuis A, Heuvelink AE, Harmanus C, Kuijper EJ. Prevalence of *Clostridium difficile* in retailed meat in the Netherlands. *Int J Food Microbiol*. (2011) 144:561–4. doi: 10.1016/j.ijfoodmicro.2010.11.007
89. Weese JS, Avery BP, Rousseau J, Reid-Smith RJ. Detection and enumeration of *Clostridium difficile* spores in retail beef and pork. *Appl Environ Microbiol*. (2009) 75:5009–11. doi: 10.1128/AEM.00480-09
90. Rahimi E, Jalali M, Weese JS. Prevalence of *Clostridium difficile* in raw beef, cow, sheep, goat, camel and buffalo meat in Iran. *BMC Pub Health*. (2014) 14:119. doi: 10.1186/1471-2458-14-119
91. Weese JS, Reid-Smith RJ, Avery BP, Rousseau J. Detection and characterization of *Clostridium difficile* in retail chicken. *Lett Appl Microbiol*. (2010) 50:362–5. doi: 10.1111/j.1472-765X.2010.02802.x
92. Harvey RB, Norman KN, Andrews K, Hume ME, Scanlan CM, Callaway TR, et al. *Clostridium difficile* in poultry and poultry meat. *Foodborne Pathog Dis*. (2011) 8:1321–3. doi: 10.1089/fpd.2011.0936
93. Harvey RB, Norman KN, Andrews K, Norby B, Hume ME, Scanlan CM, et al. *Clostridium difficile* in retail meat and processing plants in Texas. *J Vet Diagn Invest*. (2011) 23:807–11. doi: 10.1177/104063871107893
94. Curry SR, Marsh JW, Schlackman JL, Harrison LH. Clostridium difficile prevalence in uncooked ground meat products from Pittsburgh, PA. *Appl Environ Microbiol*. (2012) 78:4183–6. doi: 10.1128/AEM.00842-12
95. Pasquale V, Romano V, Rupnik M, Capuano F, Bove D, Aliberti F, et al. Occurrence of toxigenic *Clostridium difficile* in edible bivalve molluscs. *Food Microbiol*. (2012) 31:309–12. doi: 10.1016/j.fm.2012.03.001
96. Quesada-Gómez C, Mulvey MR, Vargas P, del Mar Gamboa-Coronado M, Rodríguez C, Rodríguez-Cavillini E. Isolation of a toxigenic and clinical genotype of *Clostridium difficile* in retail meats in Costa Rica. *J Food Prot*. (2013) 76:348–51. doi: 10.4315/0362-028X.JFP-12-169
97. Rodriguez-Palacios A, Reid-Smith RJ, Staempfli HR, Daignault D, Janecko N, Avery BP, et al. Possible seasonality of *Clostridium difficile* in retail meat, Canada. *Emerg Infect Dis*. (2009) 15:802–5. doi: 10.3201/eid1505.081084
98. Hoffer E, Haechler H, Frei R, Stephan R. Low occurrence of *Clostridium difficile* in fecal samples of healthy calves and pigs at slaughter and in minced meat in Switzerland. *J Food Prot*. (2010) 73:973–5. doi: 10.4315/0362-028X-73.5.973
99. Metcalf D, Avery BP, Janecko N, Matic N, Reid-Smith R, Weese JS. *Clostridium difficile* in seafood and fish. *Anaerobe*. (2011) 17:85–6. doi: 10.1016/j.anaerobe.2011.02.008
100. Metcalf D, Reid-Smith RJ, Avery BP, Weese JS. Prevalence of *Clostridium difficile* in retail pork. *Can Vet J*. (2010) 51:873–6.
101. Metcalf DS, Costa MC, Dew WM, Weese JS. *Clostridium difficile* in vegetables, Canada. *Lett Appl Microbiol*. (2010) 51:600–2. doi: 10.1111/j.1472-765X.2010.02933.x
102. Wu YC, Chen CM, Kuo CJ, Lee JJ, Chen PC, Chang YC, et al. Prevalence and molecular characterization of *Clostridium difficile* isolates from a pig slaughterhouse, pork, and humans in Taiwan. *Int J Food Microbiol*. (2017) 242:37–44. doi: 10.1016/j.ijfoodmicro.2016.11.010
103. Rodriguez C, Taminiau B, Van Broeck J, Avesani V, Delmee M, Daube G. *Clostridium difficile* in young farm animals and slaughter animals in Belgium. *Anaerobe*. (2012) 18:621–5. doi: 10.1016/j.anaerobe.2012.09.008
104. Bakri M. Prevalence of *Clostridium difficile* in raw cow, sheep, and goat meat in Jazan, Saudi Arabia. *Saudi J Biol Sci*. (2018) 25:783–5. doi: 10.1016/j.sjbs.2016.07.002
105. Kouassi KA, Dadie AT, N'Guessan KF, Dje KM, Loukou YG. *Clostridium perfringens* and *Clostridium difficile* in cooked beef sold in Cote d'Ivoire and their antimicrobial susceptibility. *Anaerobe*. (2014) 28:90–4. doi: 10.1016/j.anaerobe.2014.05.012
106. Troiano T, Harmanus C, Sanders IMJG, Pasquale V, Dumontet S, Capuano F, et al. Toxigenic *Clostridium difficile* PCR ribotypes in edible marine bivalve molluscs in Italy. *Int J Food Microbiol*. (2015) 208:30–4. doi: 10.1016/j.ijfoodmicro.2015.05.002
107. Eckert C, Burghoffer B, Barbut F. Contamination of ready-to-eat raw vegetables with *Clostridium difficile* in France. *J Med Microbiol*. (2013) 62 (Pt 9):1435–8. doi: 10.1099/jmm.0.056358-0
108. Norman KN, Harvey RB, Andrews K, Hume ME, Callaway TR, Anderson RC, et al. Survey of *Clostridium difficile* in retail seafood in College Station, Texas. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess*. (2014) 31:1127–9. doi: 10.1080/19440049.2014.888785
109. Rodriguez-Palacios A, Ilic S, LeJeune JT. *Clostridium difficile* with moxifloxacin/clindamycin resistance in vegetables in Ohio, USA, and prevalence meta-analysis. *J Pathog*. (2014) 2014:158601. doi: 10.1155/2014/158601
110. Yamoudy M, Mirlohi M, Isfahani BN, Jalali M, Esfandiari Z, Hosseini NS. Isolation of toxigenic *Clostridium difficile* from ready-to-eat salads by multiplex polymerase chain reaction in Isfahan, Iran. *Adv Biomed Res*. (2015) 4:87. doi: 10.4103/2277-9175.156650
111. Bakri MM, Brown DJ, Butcher JP, Sutherland AD. *Clostridium difficile* in ready-to-eat salads, Scotland. *Emerg Infect Dis*. (2009) 15:817–8. doi: 10.3201/eid1505.081186
112. Alam MJ, Anu A, Walk ST, Garey KW. Investigation of potentially pathogenic *Clostridium difficile* contamination in household environs. *Anaerobe*. (2014) 27:31–3. doi: 10.1016/j.anaerobe.2014.03.002
113. Shaughnessy MK, Bobr A, Kuskowski MA, Johnston BD, Sadowsky MJ, Khoruts A, et al. Environmental contamination in households of patients with recurrent clostridium difficile infection. *Appl Environ Microbiol*. (2016) 82:2686–92. doi: 10.1128/AEM.03888-15
114. Orden C, Neila C, Blanco JL, Alvarez-Perez S, Harmanus C, Kuijper EJ, et al. Recreational sandboxes for children and dogs can be a source of epidemic ribotypes of *Clostridium difficile*. *Zoonoses Public Health*. (2018) 65:88–95. doi: 10.1111/zph.12374
115. Janezic S, Potocnik M, Zidaric V, Rupnik M. Highly divergent *Clostridium difficile* strains isolated from the environment. *PLoS ONE*. (2016) 11:e0167101. doi: 10.1371/journal.pone.0167101
116. Zidaric V, Beigot S, Lapajne S, Rupnik M. The occurrence and high diversity of *Clostridium difficile* genotypes in rivers. *Anaerobe*. (2010) 16:371–5. doi: 10.1016/j.anaerobe.2010.06.001
117. Chukwu EE, Ogunsola FT, Nwaokorie FO, Coker AO. Characterization of clostridium species from food commodities and faecal specimens in Lagos state, Nigeria. *West Afr J Med*. (2015) 34:167–73.
118. Al Saif N, Brazier JS. The distribution of *Clostridium difficile* in the environment of South Wales. *J Med Microbiol*. (1996) 45:133–7. doi: 10.1099/00222615-45-2-133

119. Simango C. Prevalence of *Clostridium difficile* in the environment in a rural community in Zimbabwe. *Trans R Soc Trop Med Hyg*. (2006) 100:1146–50. doi: 10.1016/j.trstmh.2006.01.009
120. Pasquale V, Romano VJ, Rupnik M, Dumontet S, CiŽnár I, Aliberti F, et al. Isolation and characterization of *Clostridium difficile* from shellfish and marine environments. *Folia Microbiol*. (2011) 56:431–7. doi: 10.1007/s12223-011-0068-3
121. Xu C, Weese JS, Flemming C, Odumeru J, Warriner K. Fate of *Clostridium difficile* during wastewater treatment and incidence in Southern Ontario watersheds. *J Appl Microbiol*. (2014) 117:891–904. doi: 10.1111/jam.12575
122. Lim S-C, Moono P, Riley TV. *Clostridium difficile* found in gardening products: innocent bystander or the cause of community-acquired C. difficile infection through contamination of foods and environments? In: *Proceeding 44nd The Australian Society for Microbiology (ASM)*. Perth, WA: ASM (2016).
123. Keel MK, Songer JG. The comparative pathology of *Clostridium difficile*-associated disease. *Vet Pathol*. (2006) 43:225–40. doi: 10.1354/vp.43-3-225
124. Nagy J, Bilkei G. Neonatal piglet losses associated with *Escherichia coli* and *Clostridium difficile* infection in a Slovakian outdoor production unit. *Vet J*. (2003) 166:98–100. doi: 10.1016/S1090-0233(02)00252-6
125. Songer JG, Anderson MA. *Clostridium difficile*: an important pathogen of food animals. *Anaerobe*. (2006) 12:1–4. doi: 10.1016/j.anaerobe.2005.09.001
126. Weese JS, Salgado-Bierman F, Rupnik M, Smith DA, van Coeverden de Groot P. Clostridium (Clostridioides) difficile shedding by polar bears (*Ursus maritimus*) in the Canadian Arctic. *Anaerobe*. (2019) 57:35–8. doi: 10.1016/j.anaerobe.2019.03.013
127. Rodriguez C, Taminiau B, Van Broeck J, Delmee M, Daube G. *Clostridium difficile* in food and animals: a comprehensive review. *Adv Exp Med Biol*. (2016) 932:65–92. doi: 10.1007/5584_2016_27
128. Knight DR, Squire MM, Collins DA, Riley TV. Genome analysis of *Clostridium difficile* PCR ribotype 014 lineage in Australian pigs and humans reveals a diverse genetic repertoire and signatures of long-range interspecies transmission. *Front Microbiol*. (2017) 7:2138. doi: 10.3389/fmicb.2016.02138
129. Weese JS. *Clostridium difficile* in food-innocent bystander or serious threat? *Clin Microbiol Infect*. (2010) 16:3–10. doi: 10.1111/j.1469-0691.2009.03108.x
130. Songer JG, Trinh HT, Killgore GE, Thompson AD, McDonald LC, Limbago BM. *Clostridium difficile* in retail meat products, USA, 2007. *Emerg Infect Dis*. (2009) 15:819–21. doi: 10.3201/eid1505.081071
131. Janezic S, Mlakar S, Rupnik M. Dissemination of *Clostridium difficile* spores between environment and households: dog paws and shoes. *Zoonoses Public Health*. (2018) 65: 669–74 doi: 10.1111/zph.12475
132. Gerding DN, Muto CA, Owens RC Jr. Measures to control and prevent *Clostridium difficile* infection. *Clin Infect Dis*. (2008) 46:43–9. doi: 10.1086/521861
133. Edwards AN, Karim ST, Pascual RA, Jowhar LM, Anderson SE, McBride SM. Chemical and stress resistances of *Clostridium difficile* spores and vegetative cells. *Front Microbiol*. (2016) 7:1698. doi: 10.3389/fmicb.2016.01698
134. Rupnik M, Songer JG. *Clostridium difficile*: its potential as a source of foodborne disease. *Adv Food Nutr Res*. (2010) 60:53–66. doi: 10.1016/S1043-4526(10)60003-4
135. Squire MM, Riley TV. *Clostridium difficile* infection: the next big thing! *Microbiol Aus*. (2012) 33:163–4. Available online at: http://microbiology.publish.csiro.au/?paper=MA12163
136. Burt SA, Siemeling L, Kuijper EJ, Lipman LJ. Vermin on pig farms are vectors for *Clostridium difficile* PCR ribotypes 078 and 045. *Vet Microbiol*. (2012) 160:256–8. doi: 10.1016/j.vetmic.2012.05.014
137. Keessen EC, Donswijk CJ, Hol SP, Hermanus C, Kuijper EJ, Lipman LJ. Aerial dissemination of *Clostridium difficile* on a pig farm and its environment. *Environ Res*. (2011) 111:1027–32. doi: 10.1016/j.envres.2011.09.014
138. Knight DR, Putsathit P, Elliott B, Riley TV. Contamination of Australian newborn calf carcasses at slaughter with *Clostridium difficile*. *Clin Microbiol Infect*. (2016) 22:266.e1–7. doi: 10.1016/j.cmi.2015.11.017
139. Xu C, Wang D, Huber A, Weese SJ, Warriner K. Persistence of *Clostridium difficile* in wastewater treatment-derived biosolids during land application or windrow composting. *J Appl Microbiol*. (2016) 120:312–20. doi: 10.1111/jam.13018
140. Squire MM, Knight DR, Riley TV. Community-acquired *Clostridium difficile* infection and Australian food animals. *Microbiol Aus*. (2015) 36:111–3. doi: 10.1071/MA15040
141. Squire MM, Riley TV. *Clostridium difficile* infection in humans and piglets: a 'One Health' opportunity. *Curr Top Microbiol Immunol*. (2013) 365:299–314. doi: 10.1007/82_2012_237
142. Rodriguez-Palacios A, Ilic S, LeJeune JT. Subboiling moist heat favors the selection of enteric pathogen *Clostridium difficile* PCR ribotype 078 spores in food. *Can J Infect Dis Med Microbiol*. (2016) 2016:1462405. doi: 10.1155/2016/1462405
143. Deng K, Plaza-Garrido A, Torres JA, Paredes-Sabja D. Survival of *Clostridium difficile* spores at low temperatures. *Food Microbiol*. (2015) 46:218–21. doi: 10.1016/j.fm.2014.07.022
144. Rodriguez-Palacios A, LeJeune JT. Moist heat resistance, spore aging, and superdormancy in *Clostridium difficile*. *Appl Environ Microbiol*. (2011) 77:3085–91. doi: 10.1128/AEM.01589-10
145. Rodriguez-Palacios A, Reid-Smith RJ, Staempfli HR, Weese JS. *Clostridium difficile* survives minimal temperature recommended for cooking ground meats. *Anaerobe*. (2010) 16:540–2. doi: 10.1016/j.anaerobe.2010.05.004
146. MLA. *Market Information & Industry Insights – Australian Cattle Industry Projections 2015*. Meat and Livestock Australia (MLA) (2015). Available online at: http://www.mla.com.au/About-MLA/News-and-media/Media-releases/2015-cattle-industry-projections-released (accessed March 1, 2019).
147. Wu YC, Lee JJ, Tsai BY, Liu YF, Chen CM, Tien N, et al. Potentially hypervirulent *Clostridium difficile* PCR ribotype 078 lineage isolates in pigs and possible implications for humans in Taiwan. *Int J Med Microbiol*. (2016) 306:115–22. doi: 10.1016/j.ijmm.2016.02.002
148. Viau E, Peccia J. Survey of wastewater indicators and human pathogen genomes in biosolids produced by class A and class B stabilization treatments. *Appl Environ Microbiol*. (2009) 75:164–74. doi: 10.1128/AEM.01331-08
149. Romano V, Pasquale V, Krovacek K, Mauri F, Demarta A, Dumontet S. Toxigenic *Clostridium difficile* PCR ribotypes from wastewater treatment plants in southern Switzerland. *Appl Environ Microbiol*. (2012) 78:6643–6. doi: 10.1128/AEM.01379-12
150. Squire MM, Lim SC, Foster NF, Riley TV. Detection of *Clostridium difficile* after treatment in a two-stage pond system. In: van Barneveld RJ, editor. *Manipulating Pig Production, Vol 8: Proceedings 13th Biannual Conference of the Australian Pig Science Association* (APSA). Adelaide, SA. (2011). p. 215.
151. Moono P, Lim SC, Riley TV. High prevalence of toxigenic *Clostridium difficile* in public space lawns in Western Australia. *Sci Rep*. (2017) 7:41196. doi: 10.1038/srep41196
152. Eyre DW, Walker AS. *Clostridium difficile* surveillance: harnessing new technologies to control transmission. *Expert Rev Anti Infect Ther*. (2013) 11:1193–205. doi: 10.1586/14787210.2013.845987
153. Olson ND, Lund SP, Colman RE, Foster JT, Sahl JW, Schupp JM, et al. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Front Genet*. (2015) 6:235. doi: 10.3389/fgene.2015.00235
154. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, et al. Diverse sources of *Clostridium difficile* infection identified on whole-genome sequencing. *N Engl J Med*. (2013) 369:1195–205. doi: 10.1056/NEJMoa1216064
155. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol*. (2012) 13:R118. doi: 10.1186/gb-2012-13-12-r118
156. Eyre DW, Fawley WN, Rajgopal A, Settle C, Mortimer K, Goldenberg SD, et al. Comparison of control of *Clostridium difficile* infection in six english hospitals using whole-genome sequencing. *Clin Infect Dis*. (2017) 65:433–41. doi: 10.1093/cid/cix338
157. Eyre DW, Fawley WN, Best EL, Griffiths D, Stoesser NE, Crook DW, et al. Comparison of multilocus variable-number tandem-repeat analysis and whole-genome sequencing for investigation of *Clostridium difficile* transmission. *J Clin Microbiol*. (2013) 51:4141–9. doi: 10.1128/JCM.01095-13

158. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. *J Clin Microbiol.* (2018) 56:e01987-17. doi: 10.1128/JCM.01987-17
159. Cairns MD, Preston MD, Hall CL, Gerding DN, Hawkey PM, Kato H, et al. Comparative genome analysis and global phylogeny of the toxin variant *Clostridium difficile* PCR ribotype 017 reveals the evolution of two independent sub-lineages. *J Clin Microbiol.* (2016) 55:865–76. doi: 10.1128/JCM.00487-17
160. Jhung MA, Thompson AD, Killgore GE, Zukowski WE, Songer G, Warny M, et al. Toxinotype V *Clostridium difficile* in humans and food animals. *Emerg Infect Dis.* (2008) 14:1039–45. doi: 10.3201/eid1407.071641
161. Knetsch CW, Connor TR, Mutreja A, van Dorp SM, Sanders IM, Browne HP, et al. Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011. *Euro Surveill.* (2014) 19:30–41. doi: 10.2807/1560-7917.ES2014.19.45.20954
162. Stoesser N, Eyre DW, Quan TP, Godwin H, Pill G, Mbuvi E, et al. Epidemiology of *Clostridium difficile* in infants in Oxfordshire, UK: risk factors for colonization and carriage, and genetic overlap with regional *C. difficile* infection strains. *PLoS ONE*. (2017) 12:e0182307. doi: 10.1371/journal.pone.0182307
163. Barbut F, Decre D, Lalande V, Burghoffer A, Noussair L, Gigandon A, et al. Clinical features of *Clostridium difficile*-associated diarrhoea due to binary toxin (actin-specific ADP-ribosyltransferase)-producing strains. *J Med Microbiol.* (2005) 54:181–5. doi: 10.1099/jmm.0.45804-0
164. Stewart DB, Berg AS, Hegarty JP. Single nucleotide polymorphisms of the *tcdC* gene and presence of the binary toxin gene predict recurrent episodes of *Clostridium difficile* infection. *Ann Surg.* (2013) 260:299–304. doi: 10.1097/SLA.0000000000000469
165. Knetsch CW, Hensgens MPM, Harmanus C, van der Bijl MW, Savelkoul PH, Kuijper EJ, et al. Genetic markers for *Clostridium difficile* lineages linked to hypervirulence. *Microbiol.* (2011) 157:3113–23. doi: 10.1099/mic.0.051953-0
166. Carter GP, Douce GR, Govind R, Howarth PM, Mackin KE, Spencer J, et al. The anti-sigma factor TcdC modulates hypervirulence in an epidemic BI/NAP1/027 clinical isolate of *Clostridium difficile*. *PLoS Pathog.* (2011) 7:e1002317. doi: 10.1371/journal.ppat.1002317
167. Curry SR, Marsh JW, Muto CA, O'Leary MM, Pasculle AW, Harrison LH. *tcdC* genotypes associated with severe TcdC truncation in an epidemic clone and other strains of *Clostridium difficile*. *J Clin Microbiol.* (2007) 45:215–21. doi: 10.1128/JCM.01599-06
168. Spigaglia P, Barbanti F, Mastrantonio P. Multidrug resistance in European *Clostridium difficile* clinical isolates. *J Antimicrob Chemother*. (2011) 66:2227–34. doi: 10.1093/jac/dkr292
169. Collins J, Danhof H, Britton RA. The role of trehalose in the global spread of epidemic *Clostridium difficile*. *Gut Microbes*. (2018) 10:204–9. doi: 10.1080/19490976.2018.1491266
170. Bakker D, Corver J, Harmanus C, Goorhuis A, Keessen EC, Fawley WN, et al. Relatedness of human and animal *Clostridium difficile* PCR ribotype 078 isolates determined on the basis of multilocus variable-number tandem-repeat analysis and tetracycline resistance. *J Clin Microbiol.* (2010) 48:3744–9. doi: 10.1128/JCM.01171-10
171. Stabler RA, Dawson LF, Valiente E, Cairns MD, Martin MJ, Donahue EH, et al. Macro and micro diversity of *Clostridium difficile* isolates from diverse sources and geographical locations. *PLoS ONE*. (2012) 7:e31559. doi: 10.1371/journal.pone.0031559
172. Álvarez-Pérez S, Blanco J, Harmanus C, Kuijper E, García M. Subtyping and antimicrobial susceptibility of *Clostridium difficile* PCR ribotype 078/126 isolates of human and animal origin. *Vet Microbiol*. (2017) 199:15–22. doi: 10.1016/j.vetmic.2016.12.001
173. Niwa H, Kato H, Hobo S, Kinoshita Y, Ueno T, Katayama Y, et al. Postoperative Clostridium difficile infection with PCR ribotype 078 strain identified at necropsy in five Thoroughbred racehorses. *Vet Record*. (2013) 173:607. doi: 10.1136/vr.101960
174. Knetsch CW, Kumar N, Forster SC, Connor TR, Browne HP, Harmanus C, et al. Zoonotic transfer of *Clostridium difficile* harboring antimicrobial resistance between farm animals and humans. *J Clin Microbiol.* (2018) 56:e01384–17. doi: 10.1128/JCM.01384-17
175. Knight DR, Kullin B, Androga GO, Barbut F, Eckert C, Johnson S, et al. Evolutionary and genomic insights into *Clostridioides difficile* sequence type 11: a diverse, zoonotic and antimicrobial resistant lineage of global One Health importance. *MBio*. (2019) 10:e00446–19. doi: 10.1128/mBio.00446-19
176. Baines SD, Wilcox MH. Antimicrobial resistance and reduced susceptibility in *Clostridium difficile*: potential consequences for induction, treatment, and recurrence of *Clostridium difficile* infection. *Antibiotics*. (2015) 4:267–98. doi: 10.3390/antibiotics4030267
177. Corver J, Bakker D, Brouwer MS, Harmanus C, Hensgens MP, Roberts AP, et al. Analysis of a *Clostridium difficile* PCR ribotype 078 100 kilobase island reveals the presence of a novel transposon, Tn*6164*. *BMC Microbiol*. (2012) 12:130. doi: 10.1186/1471-2180-12-130
178. Spigaglia P. Recent advances in the understanding of antibiotic resistance in *Clostridium difficile* infection. *Ther Adv Infect Dis.* (2016) 3:23–42. doi: 10.1177/2049936115622891
179. Knight DR, Elliott B, Chang BJ, Perkins TT, Riley TV. Diversity and evolution in the genome of *Clostridium difficile*. *Clin Microbiol Rev*. (2015) 28:721–41. doi: 10.1128/CMR.00127-14
180. Knight DR, Androga GO, Ballard SA, Howden BP, Riley TV. A phenotypically silent vanB2 operon carried on a Tn1549-like element in *Clostridium difficile*. *mSphere*. (2016) 1:e00177–16. doi: 10.1128/mSphere.00177-16
181. Dingle K, Didelot X, Quan P, Eyre DW, Stoesser N, Marwick C, et al. A role for tetracycline selection in recent evolution of the agriculture-associated *Clostridium difficile* PCR-ribotype 078. *mBIO*. (2019) 10:e02790–18. doi: 10.1128/mBio.02790-18
182. Weinert LA, Chaudhuri RR, Wang J, Peters SE, Corander J, Jombart T, et al. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nat Commun*. (2015) 6:6740. doi: 10.1038/ncomms7740
183. Wertheim HF, Nghia HD, Taylor W, Schultsz C. *Streptococcus suis*: an emerging human pathogen. *Clin Infect Dis.* (2009) 48:617–25. doi: 10.1086/596763
184. Robinson TP, Wertheim HF, Kakkar M, Kariuki S, Bu D, Price LB. Animal production and antimicrobial resistance in the clinic. *Lancet.* (2016) 387:e1–3. doi: 10.1016/S0140-6736(15)00730-8
185. Freeman J, Vernon J, Morris K, Nicholson S, Todhunter S, Longshaw C, et al. Pan-European longitudinal surveillance of antibiotic resistance among prevalent *Clostridium difficile* ribotypes. *Clin Microbiol Infect.* (2014) 21:248 e9-e16. doi: 10.1016/j.cmi.2014.09.017
186. Knight DR, Giglio S, Huntington PG, Korman TM, Kotsanas D, Moore CV, et al. Surveillance for antimicrobial resistance in Australian isolates of *Clostridium difficile*, 2013–14. *J Antimicrob Chemother.* (2015) 70:2992–9. doi: 10.1093/jac/dkv220
187. Foster NF, Collins DA, Ditchburn SL, Duncan CN, van Schalkwyk JW, Golledge CL, et al. Epidemiology of *Clostridium difficile* infection in two tertiary-care hospitals in Perth, Western Australia: a cross-sectional study. *N Microb N Infect*. (2014) 2:64–71. doi: 10.1002/nmi2.43
188. Collins DA, Putsathit P, Elliott B, Riley TV. Laboratory-based surveillance of *Clostridium difficile* strains circulating in the Australian healthcare setting in 2012. *Pathology*. (2017) 49:309–13. doi: 10.1016/j.pathol.2016.10.013
189. Tettelin H, Riley D, Cattuto C, Medini D. Comparative genomics: the bacterial pan-genome. *Curr Opin Microbiol.* (2008) 11:472–7. doi: 10.1016/j.mib.2008.09.006
190. Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang YF. Analysis of ultra low genome conservation in *Clostridium difficile*. *PLoS ONE.* (2010) 5:e15147. doi: 10.1371/journal.pone.0015147
191. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, et al. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci USA*. (2010) 107:7527–32. doi: 10.1073/pnas.0914322107
192. Forgetta V, Oughton MT, Marquis P, Brukner I, Blanchette R, Haub K, et al. Fourteen-genome comparison identifies DNA markers for severe-disease-associated strains of *Clostridium difficile*. *J Clin Microbiol.* (2011) 49:2230–8. doi: 10.1128/JCM.00391-11
193. Rouli L, Merhej V, Fournier P-E, Raoult D. The bacterial pangenome as a new tool for analysing pathogenic bacteria. *N Microb New Infect.* (2015) 7:72–85. doi: 10.1016/j.nmni.2015.06.005

194. Georgiades K, Raoult D. Defining pathogenic bacterial species in the genomic era. *Front Microbiol.* (2010) 1:151. doi: 10.3389/fmicb.2010.00151
195. Wasels F, Monot M, Spigaglia P, Barbanti F, Ma L, Bouchier C, et al. Inter- and intraspecies transfer of a *Clostridium difficile* conjugative transposon conferring resistance to MLSB. *Microb Drug Resist.* (2014) 20:555–60. doi: 10.1089/mdr.2014.0015
196. Kunin V, Ahren D, Goldovsky L, Janssen P, Ouzounis CA. Measuring genome conservation across taxa: divided strains and united kingdoms. *Nucleic Acids Res.* (2005) 33:616–21. doi: 10.1093/nar/gki181
197. Yutin N, Galperin MY. A genomic update on clostridial phylogeny: gram-negative spore formers and other misplaced Clostridia. *Environ Microbiol.* (2013) 15:2631–41. doi: 10.1111/1462-2920.12173
198. Collins MD, Lawson PA, Willems A, Cordoba JJ, Fernandez-Garayzabal J, Garcia P, et al. The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *Int J Syst Bacteriol.* (1994) 44:812–26. doi: 10.1099/00207713-44-4-812
199. Ludwig W, Schleifer KH, Whitman WB. Revised road map to the phylum Firmicutes. In: De Vos P, Garrity GM, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer K-H and Whitman WB, editors. *Bergey's Manual of Systematic Bacteriology, 2nd ed, Vol 3: The Firmicutes*. New York, NY: Springer (2009). p. 1–14. doi: 10.1007/978-0-387-68489-5_1
200. Lerner H, Berg C. The concept of health in One Health and some practical implications for research and education: what is One Health? *Infect Ecol Epidemiol.* (2015) 5:25300. doi: 10.3402/iee.v5.25300
201. Thomas C, Stevenson M, Williamson DJ, Riley TV. *Clostridium difficile*-associated diarrhea: epidemiological data from Western Australia associated with a modified antibiotic policy. *Clin Infect Dis.* (2002) 35:1457–62. doi: 10.1086/342691
202. Price J, Cheek E, Lippett S, Cubbon M, Gerding DN, Sambol SP, et al. Impact of an intervention to control *Clostridium difficile* infection on hospital-and community-onset disease; an interrupted time series analysis. *Clin Microbiol Infect.* (2010) 16:1297–302. doi: 10.1111/j.1469-0691.2009.03077.x
203. Weese JS, Armstrong J. Outbreak of *Clostridium difficile*-associated disease in a small animal veterinary teaching hospital. *J Vet Intern Med.* (2003) 17:813–6. doi: 10.1111/j.1939-1676.2003.tb02519.x
204. Riley TV. Is *Clostridium difficile* a threat to Australia's biosecurity? *Med J Aust*. (2009) 190:661–2. doi: 10.5694/j.1326-5377.2009.tb02630.x
205. Ghose C, Kelly CP. The prospect for vaccines to prevent *Clostridium difficile* infection. *Infect Dis Clin North Am.* (2015) 29:145–62. doi: 10.1016/j.idc.2014.11.013

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

REVIEW
published: 27 June 2019
doi: 10.3389/fpubh.2019.00172

Whole Genome Sequencing: Bridging One-Health Surveillance of Foodborne Diseases

Peter Gerner-Smidt[1], John Besser[1], Jeniffer Concepción-Acevedo[1], Jason P. Folster[1], Jasmine Huffman[1], Lavin A. Joseph[1], Zuzana Kucerova[1], Megin C. Nichols[2], Colin A. Schwensohn[2] and Beth Tolar[1]*

[1] The Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, United States,
[2] The Outbreak Response and Prevention Branch, Centers for Disease Control and Prevention, Atlanta, GA, United States

Infections caused by pathogens commonly acquired from consumption of food are not always transmitted by that route. They may also be transmitted through contact to animals, other humans or the environment. Additionally, many outbreaks are associated with food contaminated from these non-food sources. For this reason, such presumed foodborne outbreaks are best investigated through a One Health approach working across human, animal and environmental sectors and disciplines. Outbreak strains or clones that have propagated and continue to evolve in non-human sources and environments often show more sequence variation than observed in typical monoclonal point-source outbreaks. This represents a challenge when using whole genome sequencing (WGS), the new gold standard for molecular surveillance of foodborne pathogens, for outbreak detection and investigation. In this review, using recent examples from outbreaks investigated in the United States (US) some aspects of One Health approaches that have been used successfully to solve such outbreaks are presented. These include using different combinations of flexible WGS based case definition, efficient epidemiological follow-up, traceback, surveillance, and testing of potential food and environmental sources and animal hosts.

Keywords: whole genome sequencing (WGS), outbreak, one health, zoonotic, food, environment, animals, investigation

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control, Sweden

Reviewed by:
Arie Hendrik Havelaar, University of Florida, United States
Eelco Franz, Centre for Infectious Disease Control (RIVM), Netherlands

****Correspondence:***
Peter Gerner-Smidt
plg5@cdc.gov

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *08 April 2019*
Accepted: *10 June 2019*
Published: *27 June 2019*

Citation:
Gerner-Smidt P, Besser J, Concepción-Acevedo J, Folster JP, Huffman J, Joseph LA, Kucerova Z, Nichols MC, Schwensohn CA and Tolar B (2019) Whole Genome Sequencing: Bridging One-Health Surveillance of Foodborne Diseases. Front. Public Health 7:172. doi: 10.3389/fpubh.2019.00172

INTRODUCTION

Infections caused by pathogens commonly transmitted by food are common, potentially all preventable and therefore of major public health importance. They are a problem all over the world affecting all parts of society in developing and developed countries (1). Although mostly presenting as a self-limiting diarrheal illness, more severe illness requiring hospitalization is frequently seen. Foodborne illness caused by certain pathogens, e.g., *Listeria monocytogenes,* carry a significant mortality. Outbreaks are common with ∼1,000 outbreaks being investigated in the US every year (2). Foodborne pathogens can be any infectious agent, e.g., bacteria, parasites, virus, and prions, even though this review focuses on bacterial pathogens.

In the US approximately one in six persons acquires a foodborne illness every year (3). However, it needs to be kept in mind that not all infections caused by pathogens commonly transmitted through food are actually foodborne. Although illness is often caused through ingestion of contaminated food, the primary reservoir of these pathogens is rarely food but rather animals,

water or the environment. The reservoir of pathogens like non-typhoidal *Salmonella*, *E. coli*, *Campylobacter*, and *Yersinia* is primarily zoonotic, i.e., wildlife, pets, or food production animals. *Listeria monocytogenes* is ubiquitous and may be found in the environment, animals and food. A classical example of a waterborne pathogen is *Vibrio* spp. but many foodborne enteric pathogens may also be transmitted through contaminated recreational or drinking water. Ill humans can also infect each other. Thus, infection caused by pathogens commonly transmitted by food is a classic example of a One-Health challenge. The One Health concept includes the health of humans, animals and the environment. In this paper, the focus is on human infections. If public health investigators only focus their attention to food sources and vehicles when investigating potential foodborne outbreaks, they will miss opportunities to identify primary sources and prevent further illness and outbreaks from animal or environmental sources. Even when the vehicle is foodborne, e.g., meat from a specific supplier, a proper conducted investigation should include a root cause analysis. For example, in addition to removing a vehicle from the market, a thorough trace-back of the vehicle to the primary production should be performed, e.g., to the farm and the suppliers of that farm, even if the ultimate source is in a different country or on a different continent. This can best be achieved through a One Health approach to investigation working across human, animal, and environmental sectors and disciplines.

With the introduction of affordable and fast next generation sequencers in the early 2000s, WGS has revolutionized molecular epidemiology and laboratory surveillance of infections caused by pathogens commonly transmitted though food providing public health researchers with a tool of unprecedented precision and discrimination for subtyping. Additionally, WGS may provide a wealth of information at the push of a button that exceeds what in the past was typically gathered using traditional phenotypic and genotypic tests in public health laboratories e.g., species identification, serotype, pathotype, virulence profile, antimicrobial resistance, and plasmid content to name a few. A description of the analytical tools is beyond the scope of this paper and may be found elsewhere (4–6). Using WGS, public health scientists typically detect outbreaks by looking for tight clusters of infections caused by a specific pathogens in time and space typically differing by <10 single nucleotide polymorphisms (SNPs) or 10 alleles by core genome multi-locus sequence typing (cgMLST) analysis (7). This is the typical scenario of a monoclonal outbreak from a point source that has been contaminated because of a single event (8–12). However, in many outbreaks with a zoonotic or environmental source, the outbreak strains have persisted in their hosts and reservoirs and therefore have time to diversify beyond what is expected in a point source outbreak (13, 14). In such outbreaks, the source may also be contaminated with more strains leading to polyclonal, possibly multi-species outbreaks. Detecting and investigating such outbreaks pose specific challenges. In this paper, a number of such outbreaks that recently have been investigated in the US will be reviewed with an emphasis on their characteristics as experienced with WGS using the cgMLST subtyping approach used by PulseNet, the US molecular subtyping network for foodborne disease surveillance (15), and how the challenges of their interpretation was overcome.

A PERSISTENT POLYCLONAL OUTBREAK OF LISTERIOSIS ASSOCIATED WITH CONTAMINATION OF ICE CREAM PRODUCTION PREMISES

In 2015, *Listeria monocytogenes* was isolated from a number of samples of ice cream from a distribution center (https://www.cdc.gov/listeria/outbreaks/ice-cream-03-15/index.html). Some of these isolates matched four clinical isolates from a single hospital in Kansas collected during the past year by PFGE and WGS; a fifth clinical case in the hospital was infected with an unrelated strain. The particular brand of ice cream was regularly served in milkshakes at the hospital and all cases were considered nosocomially acquired. This led to the inspection of the company's production facilities in three states by local authorities and the Food and Drug Administration (FDA) over the next months. Numerous samples from the production facilities and products from two states were positive for *Listeria* in low numbers (16). All of the new isolates were compared against the PulseNet database using PFGE and WGS. Five clinical isolates from patients in three states matched product or production environment isolates by WGS spanning the years 2010–2014. Researchers from FDA compared the sequences of 137 food and environmental and nine clinical isolates (17). This analysis included the four clinical isolates from the hospital outbreak that matched any food or environmental isolates. The isolates represented 13 PFGE patterns but were clustered in only two groups by SNP analysis, one corresponding to the hospital cluster in Kansas and the other containing the historical clinical isolates from three states. All isolates belonged to sequence type (ST) 5 of clonal complex 5 (CC5) of lineage I, molecular serogroup IIb (serotypes 1/2b, 3b, or 7). The isolates within the clusters differed from each other by up to 29 SNPs and between each cluster by 40–52 SNPs. A summary of the Centers for Disease Control and Prevention's (CDC) cgMLST analysis with representative isolate sequences of the outbreak isolates using the PulseNet customized version of the Pasteur scheme (18) is shown in **Figure 1**.

The fifth Kansas hospital isolate is not included in the figure since it belongs to a different lineage, ST and serotype, lineage II, ST573 and serotype 1/2a, and differs by 1,290–1,377 alleles from any other outbreak isolate. This analysis are generally consistent with the FDA SNP analysis (17). Two clusters are seen, one containing the four Kansas hospital patient isolates and food and environment isolates from one plant and the other the five historical clinical isolates and the non-human isolates from the other facility. Isolates in each cluster differ from each other by up to 16 and 10 alleles, respectively. The two clusters differed by up to 123 alleles. The genetic differences observed within each cluster is slightly higher than typically is observed for point source outbreaks. However, the allele differences between the two clusters were twice as high than observed in the FDA SNP analysis and an average number of allele differences of more than

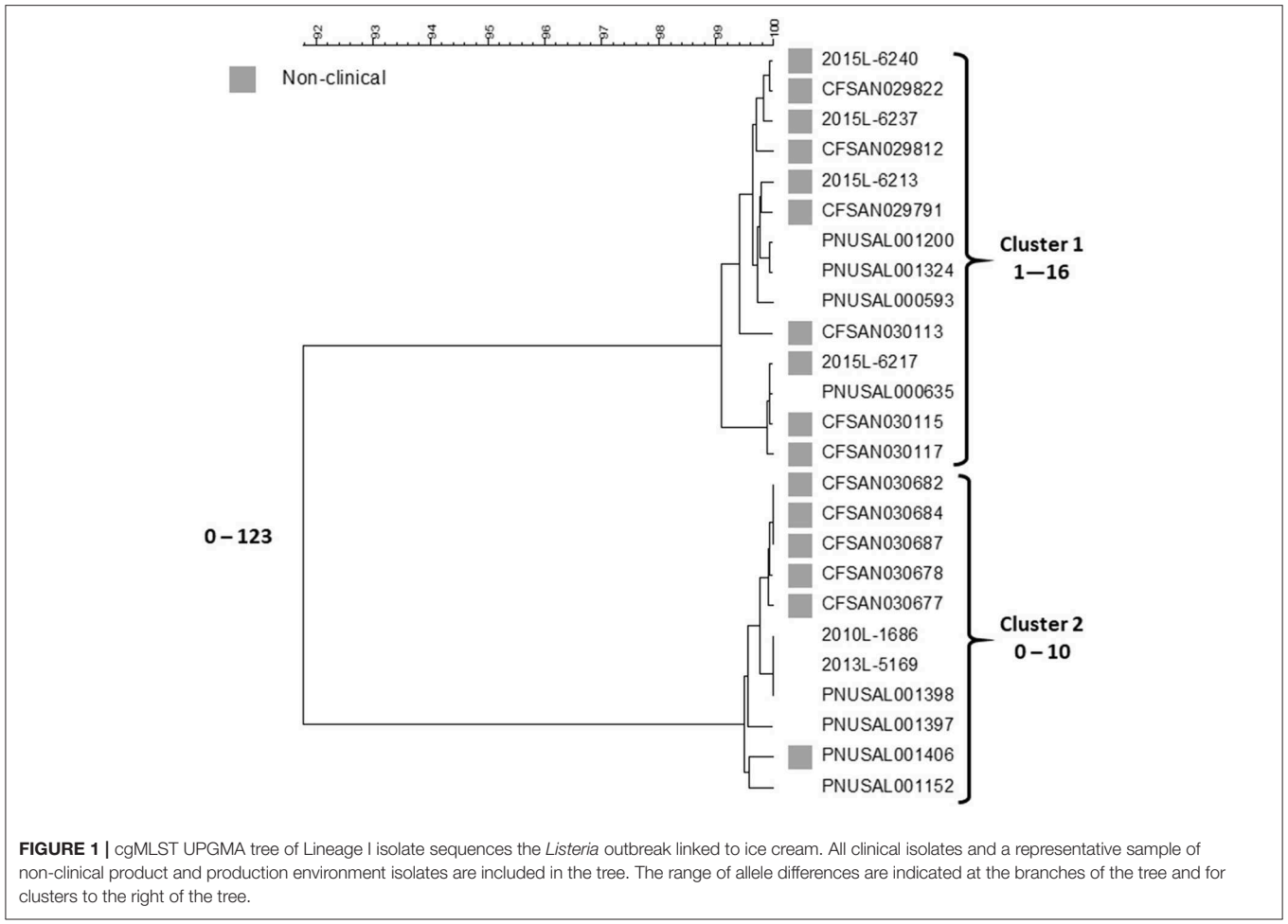


FIGURE 1 | cgMLST UPGMA tree of Lineage I isolate sequences the *Listeria* outbreak linked to ice cream. All clinical isolates and a representative sample of non-clinical product and production environment isolates are included in the tree. The range of allele differences are indicated at the branches of the tree and for clusters to the right of the tree.

100 alleles are higher than typically observed between isolates that could be related epidemiologically (7). This speaks against the hypothesis of a recent common origin of the two clones. By phenotypic serotyping the isolates in the Kansas hospital cluster was 1/2b whereas those in the "historical cluster" was 3b. Sequence types (from 7 house-keeping gene MLST) (19) associated with serotype 1/2b strains often also contain serotype 3b strains (CDC, unpublished observation) so although we have never observed two *Listeria* serotypes in a tight monoclonal outbreak it is possible that strains of serotype 1/2b may evolve to serotype 3b or *vice versa*. Although no attempt has been made to use the data as a "molecular clock" to characterize the divergence of the two clusters in this clone, it is not impossible that the clusters could have originated from the same strain at some point in the fairly recent past. It could have been introduced in the two plants at the same time or first in one facility and then shortly thereafter from the first facility to the second. The strains may then have diversified further in each plant. Even though the products seem to have been almost uniformly contaminated (16), the contamination levels in the products were so low to rarely cause disease. Such "low and slow" outbreaks, i.e., outbreaks that go on for a long time with clinical cases occurring within long intervals, could not be detected or were not further pursued in the past because of the poorer resolution of PFGE. With the superior resolving power of WGS, this has now changed. This challenges the time aspect of a typical outbreak investigation, i.e., a cluster of clinical illness in space and time. A typical monoclonal point source outbreak evolves quickly over days to a few months. However, this outbreak shows that the time aspect of the clustering may be much longer, i.e., years. This outbreak is also noteworthy for two other aspects: (1) both clusters were detected by matching food/environmental isolates to clinical cases, and (2) the diversity by PFGE was higher than observed by WGS; at least 16 different PFGE profiles were observed by PulseNet, whereas WGS indicated that two possibly related clones caused it with one case patient harboring a third unrelated strain. It is well-known that PFGE diversity is driven by loss or acquisition of mobile genetic elements and not by mutations. In their study of this outbreak, Chen et al. (17) observed that loss or gain of prophages could explain some of the PFGE variations. Such gains and losses typically occur during long term *in vivo* propagation of a strain and therefore supports the notion that the outbreak strain evolved over the years and it likely was present in the production plants. Gains and losses of mobile genetic elements are usually not reflected in a SNP or cgMLST analysis since such sequences are often filtered out before analysis because they distort the phylogenetic signal.

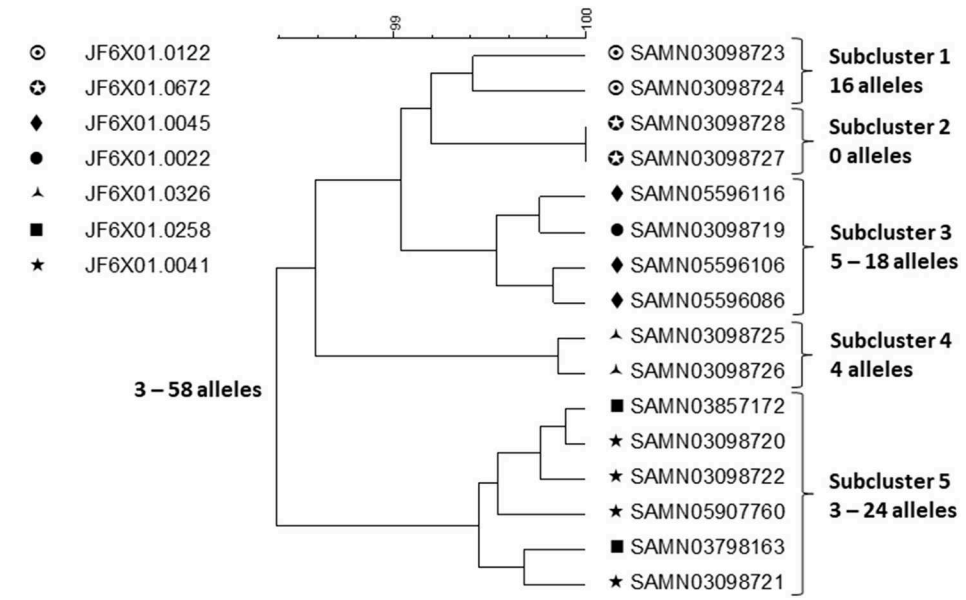


FIGURE 2 | cgMLST UPGMA tree of a representative sample of sequences of *Salmonella* ser. Heidelberg isolates displaying the full WGS diversity and representing the seven PFGE patterns from the outbreak associated with chicken produced by Company A. The range of allele differences are indicated at the branches of the tree and subclusters to the right of the tree.

A PERSISTENT POLYCLONAL MULTI DRUG RESISTANT OUTBREAK OF *Salmonella* ser. HEIDELBERG LINKED TO CHICKEN FROM SINGLE PRODUCTION COMPANY

This outbreak was investigated using PFGE, the PulseNet primary subtyping method at the time it happened. After the outbreak was over, WGS was conducted on a small sample of 30 isolates representing all PFGE patterns and sources, and representative antimicrobial susceptibilities (https://www.cdc.gov/salmonella/heidelberg-10-13/index.html). The investigation began after a cluster of infections caused by *Salmonella* ser. Heidelberg of a rare PFGE pattern (PulseNet pattern JF6X01.0258) was detected by PulseNet in 2013 (20). At the same time, a chicken breast retail sample from a production company A cultured positive for the same strain. During a few months following the detection of the outbreak, six additional clusters of clinical isolates were identified. Some of the PFGE patterns in these clusters were similar to the original outbreak pattern (differing by up to three bands) and since the patients clustered in time, geographic distribution and food history with patients from the first cluster, all seven clusters were merged into the investigation. Six out of seven outbreak strains were found in left-over raw chicken from patient homes and from products from three production establishments of company A. A total of 634 outbreak related patients were identified in 29 states and Puerto Rico. Antimicrobial susceptibility testing of patient and product isolates showed numerous profiles with isolates being pan-susceptible, resistant to one, two, three, or more classes of antimicrobials with weak correlation between resistance profile and PFGE pattern. Following recalls and operational adjustments at company A, the outbreak was declared over a year later.

A small sample of outbreak related isolates from patients with exposure to chicken from company A and product samples were sequenced to shed further light on the outbreak strains (**Figure 2**).

Clustering was performed by cgMLST using the PulseNet scheme, which contains the same loci as the Enterobase scheme (21). Isolates of the same PFGE pattern clustered together but two subclusters (subclusters 3 and 5) contained isolates with two PFGE patterns intermingled. Food isolates intermingled with patient isolates of the same PFGE pattern by WGS. Considering the whole outbreak cluster, isolates differed by up to 58 alleles and within each subcluster by up to 24 alleles. Using ResFinder (22) and PlasmidFinder (23), 14 different resistance determinants, conferring resistance to seven different drug classes, were identified; eight different plasmid types were identified including common multi drug resistance plasmids, e.g., IncHI2, IncI1, and IncA/C2 confirming the diversity observed by the other method. Due to the small sample of isolates that were sequenced, the sequence variation was likely underestimated. *Salmonella* ser. Heidelberg is commonly associated with chicken. In this outbreak, the outbreak strains had probably been present in the production system for long time, likely years, giving them ample time to diversify and acquire/lose plasmids and with them resistance determinants. It is likely that fewer case-patients would have been recognized in this outbreak if cgMLST alone had been used to detect and delineate it because of the high sequence diversity among subclusters displaying the same

PFGE pattern. Thus, this is an example of an outbreak where a subtyping method with poorer discrimination than WGS, i.e., PFGE, better identifies its full extent. It is likely that by WGS small subclusters of highly similar isolates e.g., associated with restaurant or other local events, would have been identified and perhaps linked to products from company A. However, linking them all together and identifying other outbreak related ser. Heidelberg isolates among the large background of sporadic infections caused by this serotype, would be a daunting if not an impossible task.

A *Salmonella* OUTBREAK INVOLVING SIX SEROTYPES ASSOCIATED WITH CONSUMPTION OF A HERBAL SUPPLEMENT FROM SOUTH EAST ASIA

In early 2018, a tight cluster of *Salmonella* ser. I 4,[5],12:b:- (monophasic Paratyphi B var. L(+) tartrate+ [formerly Java]) was identified by PulseNet (https://www.cdc.gov/salmonella/kratom-02-18/index.html). The investigation soon confirmed it as an outbreak and pointed to an unusual vehicle, an opiod agonistic herbal supplement, kratom (*Mitragyna speciose* also known as thang, kakuam, thom, ketum, and biak) sold as powder, capsules, or tea. Leftover and unopened kratom products were tested by local authorities and FDA for *Salmonella* contamination and a number of different serotypes were identified. The outbreak strain was confirmed in the product by PFGE and WGS. A search of the PulseNet national database identified potential case patients infected with some of these additional serotypes, including *Salmonella* ser. Heidelberg, Javiana, Okatie, Thompson, and Weltevreden dating back to the beginning of 2017. Among these serotypes, Javiana and Heidelberg are among the 20 most common among clinical cases in the US, I 4,[5],12:b:-, Thompson and Weltevreden less common but still among the 100 most common serotypes, whereas Okatie is rare with 0–6 annual clinical cases typically reported (https://www.cdc.gov/nationalsurveillance/pdfs/2016-Salmonella-report-508.pdf). The outbreak investigation was expanded to include these serotypes. No particular brand of the product could be implicated but the product was recalled from the market by several distributors and retailers, including on-line businesses. In total, 199 cases were identified from 41 states (https://www.cdc.gov/salmonella/kratom-02-18/index.html). **Figures 3A,B** shows cgMLST trees of representative isolates from patients and kratom of ser. I 4,[5],12:b:- and Okatie. Most of the I 4,[5],12:b:- isolates (**Figure 3A**) formed a tight subcluster with no more than one allele difference and this cluster lead to the identification of the source. However, another subcluster contained isolates differing by up to 25 alleles and an additional two isolates differing from the clustered isolates by up to 552 alleles. The ser. Okatie isolates were loosely clonal (**Figure 3B**). Of this serotype, 10 clinical isolates and 10 product isolates were sequenced differing from each other by up to 78 alleles; four subclusters were identified each containing isolates that differed by up to 2, 8, 9, and 13 alleles, respectively. Of note, each subcluster contained both clinical and product isolates.

The cgMLST results of the four other serotypes showed loose clustering in between what was seen with the two serotypes in the figures with more than 10 allele differences typically seen in monoclonal *Salmonella* outbreaks.

Kratom is grown and harvested in several countries in South East Asia and the sale and distribution systems are not transparent. Thus, it is possible that product for sale in the US originated from multiple producers in different countries and that the same product could contain kratom from more than one source. This likely explains why so many serotypes were involved. It may be speculated that the cluster caused by *Salmonella ser.* I 4,[5],12:b:- have recently contaminated kratom from one producer since it was tightly clonal, whereas the other serotypes may have been present in the production or distribution systems longer giving them time to diversify or have resulted in different contamination events at multiple producers. Because of the observed strain diversity with all serotypes it is unlikely that all clinical case-patients could have been identified without the availability of product isolates. However, the cluster associated with ser. Okatie could have been and actually was detected before the ser. I 4,[5],12:b:- cluster by serotype-based laboratory surveillance without considering WGS since it is so rare in the US. However, the association to kratom was not established before the serotype was detected in the product and the patients interviewed about that exposure. This serotype has scarcely been reported in the scientific literature but could have a focus in South East Asia.

A POLYCLONAL OUTBREAK OF MULTIDRUG RESISTANT *Campylobacter* LINKED TO CONTACT WITH PUPPIES SOLD IN A SPECIFIC PET STORE CHAIN IN THE US

This outbreak was investigated and included illnesses reported over 2 years from 2016 to 2018 (https://www.cdc.gov/campylobacter/outbreaks/puppies-9-17/index.html). One-hundred and eighteen cases of illness caused by *Campylobacter jejuni* were identified in 18 states. The isolates were resistant to 7–9 antimicrobials including the drugs commonly used to treat patients with severe illness, e.g., azithromycin, ciprofloxacin and tetracycline. This particular multidrug resistant pattern was very rare in the US when compared to data from the National Antimicrobial Resistance Monitoring System (https://www.cdc.gov/narms/index.html). Infection was associated with contact to puppies sold in a specific pet store chain. Fifty-six clinical and puppy isolates were sequenced and analyzed by cgMLST using the PulseNet customized version of the Oxford scheme (24). A sample representing the full diversity observed in the outbreak is shown in **Figure 4**. At least three outbreak clusters were identified among the patient isolates. Two of the clusters (cluster 2 and 3) also contained puppy isolates.

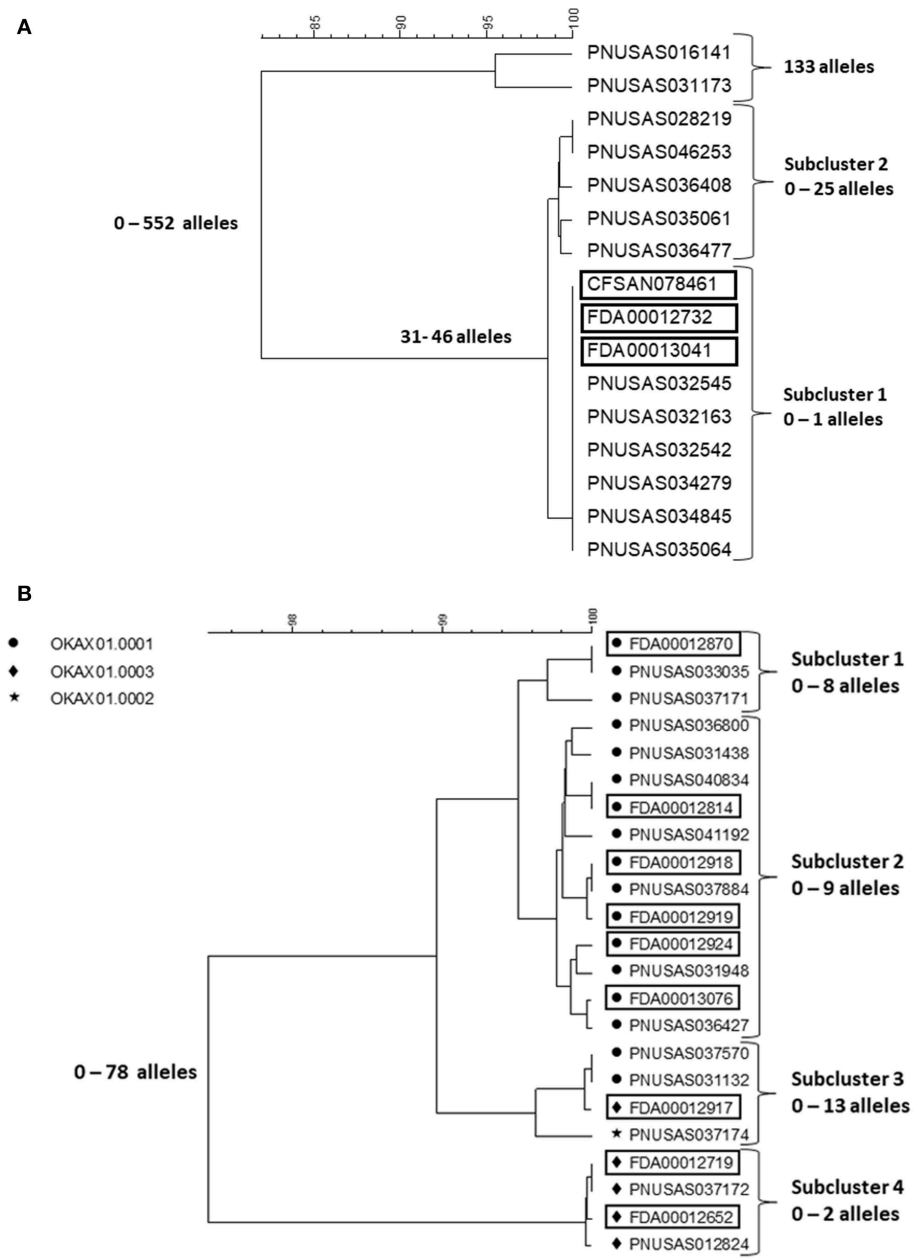


FIGURE 3 | cgMLST UPGMA trees of representative samples of sequences of *Salmonella* ser. I 4,[5],12:b:- **(A)** and Okatie **(B)**. Isolates of ser. I 4,[5],12:b:- all displayed the same PFGE pattern. Isolates of ser. Okatie displayed three PFGE as indicated on the figure. Product isolate identifiers are indicated with rectangles around them. The range of allele differences are indicated at the branches of the tree and subclusters to the right of the tree.

In the cgMLST analysis for this paper, cluster 1 contained clinical isolates that differed by up to 23 alleles; the second cluster contained clinical and puppy isolates that differed by up to 8 alleles, and the last cluster also consisted of clinical and puppy isolates differing from each other by up to 28 alleles. All isolates were multidrug resistant as determined by WGS using ResFinder, which produced similar resistance profiles by phenotypic antimicrobial resistance testing examined on select isolates. For a small subset of isolates, long read sequencing was used to determine the genetic context of resistance determinants. These determinants were found to be located on the chromosome, or on a plasmid, or on both, or missing altogether. While some determinant's location, for example the *tetO* gene, tended to sort according to clonal group (plasmid for cluster 1 and 2, plasmid and chromosome for cluster 3), other genes' location, including several aminoglycoside resistance genes, did not sort by cluster. Moreover, at least one isolate had no plasmids but had all of the resistance determinants seen in this outbreak on its chromosome. Thus, there was no apparent correlation between plasmid content and resistance,

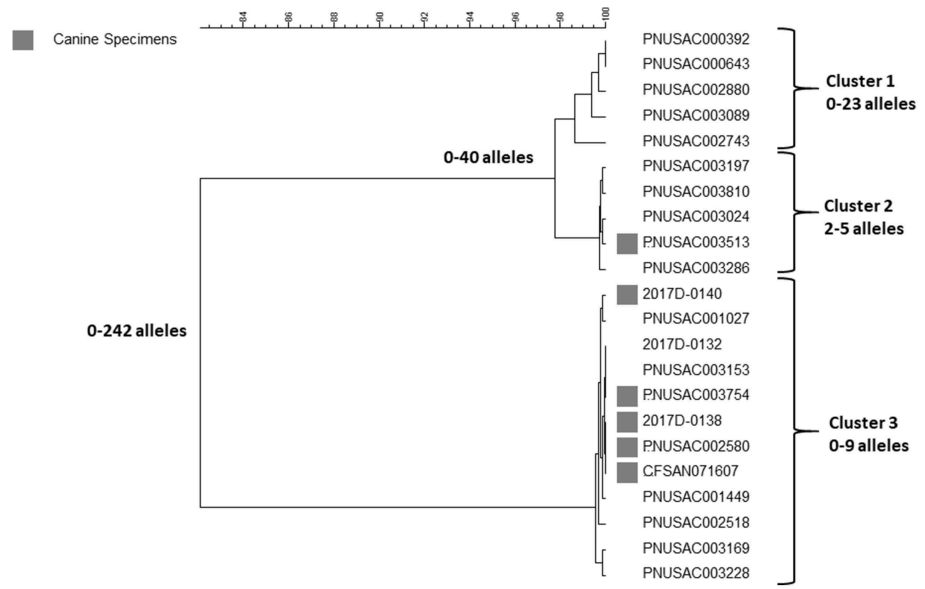


FIGURE 4 | cgMLST UPGMA tree of a representative sample of sequences of clinical and animal isolates of *Campylobacter jejuni* from an outbreak associated with contact to puppies sold in a specific pet store chain. Puppy isolates are marked with gray squares. The range of allele differences are indicated at the branches of the tree and clusters to the right of the tree.

but the resistance pattern itself was relatively stable among outbreak isolates.

MULTIPLE OUTBREAKS OF SALMONELLOSIS LINKED TO SMALL PET TURTLES, 2015–2016

Contact to reptiles is a well-known risk factor for salmonellosis. Outbreaks associated with contact to small pet turtles are common [(25, 26), https://www.cdc.gov/salmonella/agbeni-08-17/index.html]. Their characteristics are similar and here we focus on four multi-state outbreaks caused by *Salmonella* in 2015–16 (26) and in particular, the WGS results in one of them, a polyclonal outbreak, caused by ser. Pomona and Poona. The investigation began as a follow-up on a consumer complaint about a child who had acquired a *Salmonella* infection from a small turtle acquired at a flea market of a serotype involved in turtle associated outbreaks years earlier (25). The PulseNet national database was checked for PFGE clusters the past year of serotypes previously linked to turtles. This way, four multistate outbreaks with 143 case patients from 25 states of three serotypes, Sandiego, Pomona and Poona, representing six PFGE patterns were identified. This outbreak investigation included testing of human, animal and environmental isolates. Nineteen *Salmonella* isolates were cultured from the pond water of four turtle production farms in Louisiana and from turtles and water tanks from eight cases. Since turtles from the US are exported all over the world, international inquiries and literature review were conducted resulting in the identification of one potential PFGE matching patient isolates in Chile and four in Luxembourg. The patients from Chile and two from Luxembourg had confirmed exposure to turtles. Of the 116 US patients with information available, 56 (48%) reported exposure to turtles. PFGE could not separate isolates from patients reporting contact to turtles from isolates from patients with no turtle contact. WGS was then used to test if isolates associated with different sources could be differentiated by this method. cgMLST results of isolates of ser. Poona and ser. Pomona from the biggest of the outbreaks [outbreak 2 in (26)] are shown in the **Figures 5A,B**.

The Poona isolates were loosely clustered in two tighter subclusters. Overall isolates differed by up to 12 alleles whereas isolates in the two tighter subclusters differed up to seven and five alleles, respectively. All isolates except one had the same PFGE pattern JL6X01.0104; one isolate, in the first subcluster, had a different PFGE pattern JL6X01.0554. Subcluster 1 contained isolates from Luxembourg, clinical isolates from the US and a turtle tank water isolate from a patient's home. The second subcluster also contained patient isolates from the US and one turtle tank water isolate. Whereas, the allele variation in each subcluster was <10 alleles typically observed in point source outbreaks, each of them could have been detected by WGS. Because they were less related between clusters, an association between them to the same source might not have been suspected without additional information, i.e., exposure information and/or a non-human isolate linking them to turtles. The isolates from Luxembourg were obtained in 2012 and 2013 indicating that the ser. Poona isolates from 2015 to 2016 had hardly evolved.

The Pomona isolates all displayed the same PFGE profile and formed two tight subclusters and three isolate that

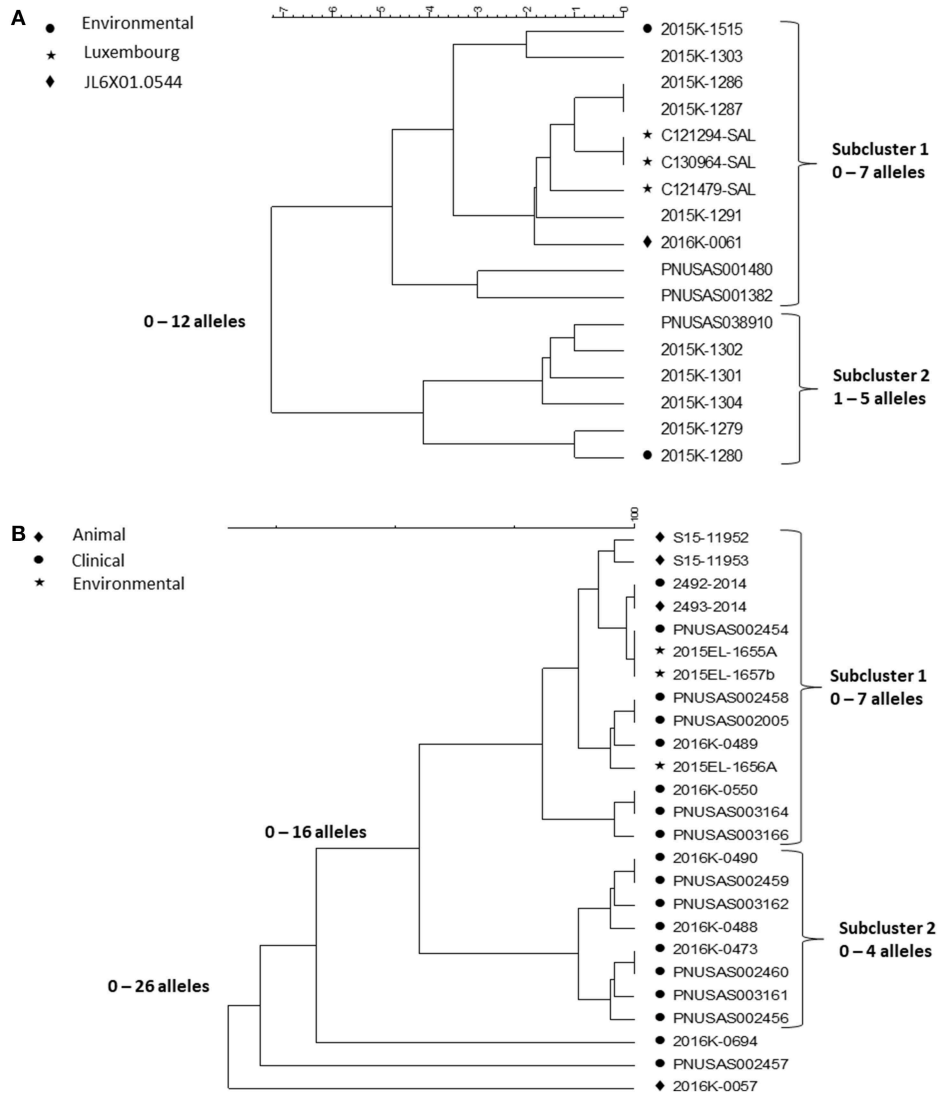


FIGURE 5 | cgMLST UPGMA trees of representative samples of sequences from one of the *Salmonella* outbreaks associated with turtle exposure. **(A)** shows the relationships between isolates of ser. Poona; isolates from the environment, from Luxemburg or of a PFGE pattern different from that displayed by all other isolates in the figure are indicated next to the strain identifier. **(B)** shows the relationship between isolates of ser. Pomona; the source of the isolates is indicated next to the strain identifier. Isolates 2493-2014 and 2492-2014 are from Chile. The range of allele differences are indicated at the branches of the tree and subclusters to the right of the tree.

appeared unrelated to the two subclusters. The first subcluster contained clinical isolates from patients with turtle exposure and isolates from turtles, pond, and tank water; the subcluster also contained the patient and associated turtle isolate from Chile from 2014 (2493-2014 and 2492-2014 in the figure). The sequences in this subcluster differed from each other by up to seven alleles. Isolates in the second subcluster differed by up to four alleles. It only contained clinical isolates and all patients reported no turtle exposure. A common exposure between patients in this subcluster was never identified. The three non-clustered isolates were a turtle isolate (2016K-0057) from an earlier outbreak in 2012 and two current clinical isolates.

DISCUSSION

A well-functioning surveillance system that integrates elements from public and animal health and the food production is optimal to detect, investigate, and solve infections commonly transmitted through food (27). The examples provided in this paper illustrate that zoonotic outbreaks and outbreaks with a persistent environmental focus, which is typical for outbreaks in the One-Health context, are often not tightly monoclonal and may therefore be difficult to recognize through laboratory based surveillance by whole genome sequencing (WGS). This technology provides so much resolution that outbreaks that are caused by strains that have had time to evolve in the

environment or in their natural hosts can be seen to have more variation than observed in typical point-source outbreaks. Using a One-Health approach in an integrated surveillance system, epidemiologic information, and isolates from animal and environmental sources, can greatly add to the ability to discriminate relatedness to clinical outbreak isolates. A number of different approaches may be used to detect, delineate, and investigate these outbreaks.

Considering additional information extracted from the sequencing information may help identify outbreaks, e.g., serotype information about an outbreak strain for a rare serotype such as *Salmonella* ser. Okatie that was associated with the outbreak linked to kratom described before. However, additional information may also cause confusion. For example, detailed information about resistance markers and plasmids can be confusing since these markers often not stable traits. However, despite such diversity multidrug resistance was helpful in recognizing and investigating two of the outbreaks described before: the *Salmonella* ser. Heidelberg associated with chicken from one production company, and the *Campylobacter* outbreak linked to pet store puppies. Similarly, PFGE may be used the same way. During the past 5 years, PFGE has remained the primary subtyping method in PulseNet with WGS used as a secondary confirmatory method except for *Listeria* where both methods have been used concomitantly for real-time surveillance. *Campylobacter* isolates are rarely subtyped in PulseNet unless an outbreak is suspected by other methods, e.g., like a cluster of multidrug resistant cases in the puppy outbreak. In the examples provided in this paper, PFGE mostly provided too much discrimination between isolates or the opposite, failed to differentiate isolates that were unrelated: multiple PFGE patterns were identified in the *Campylobacter* outbreak but only three clones were observed by WGS with so much variation in two of them that it would have been difficult to recognize them without additional resistance and exposure information. The outbreak was eventually confirmed by isolating the outbreak clones in pet store puppies and puppies owned by ill people. In the *Listeria* outbreak, enormous diversity was observed by PFGE, whereas WGS easily defined three outbreak clones/strains. In the turtle Poona outbreak subcluster described before, WGS helped differentiate PFGE clustered isolates from patients without contact to turtles from patients who had this exposure.

If a persistent or zoonotic focus for foodborne pathogens is suspected, the sequencing cluster definition may be relaxed. This may be done by initially looking for tight monoclonal clusters, e.g., differing by up to 10 alleles/SNPs, spanning a short time span since logically isolates from patients getting ill at the same time has a higher likelihood of originating from a point source, which could be a sub cluster of a larger zoonotic outbreak. Once the outbreak is recognized, and the initial patient interviews indicate that exposure to animals or an environmental source could be the vehicle, the case definition may be expanded in increments to include isolates that differ from the index cluster by for instance 25, 50, and 100 alleles or SNPs. Without associated epidemiological information, this approach may result in the inclusion of too many epidemiologically unrelated isolates during the outbreak investigation diluting any epidemiological signal that may be present. Therefore, foodborne, zoonotic, and environmental exposure information and isolates from food, zoonotic, and environmental sources should be used to determine different allele or SNP cutoffs choosing the values that provide the strongest epidemiological association. The utility of having access to sequencing information from potential sources is also extremely useful when working on an outbreak with a zoonotic or environmental focus. However, the ability to gather this information from animal isolates can be limited, as there often are few animal isolates available for comparison purposes during outbreaks, unless additional efforts are undertaken to collect them. This is at variance with clinical isolates, which are routinely collected by public health laboratories and sequenced to obtain additional information. Representative enteric bacterial isolates collected from animals are not routinely sequenced in the US. As shown in all the outbreaks described before, obtaining isolates from the potential sources was helpful to confirm the vehicle and also to facilitate recognition of the outbreak (the *Listeria* outbreak) or define its full scope (the ser. Heidelberg outbreak and the *Salmonella* outbreak linked to kratom). Thus, the importance of using a One-Health or farm to a table approach with efficient trace back when investigating outbreaks caused by pathogens commonly transmitted through food cannot be over emphasized.

International outbreaks caused by foodborne pathogens are common and WGS has the potential of bringing their recognition to the next level as more laboratories implement WGS in their routine surveillance. Until now most international outbreaks have been recognized by linking national outbreaks to each other when one country is investigating an outbreak with possible international spread and contacts other countries. Public health authorities in another country or countries may be contacted directly if there is a strong suspicion that the source of the outbreak is present in that country/those countries. Alternatively, the country may send out an inquiry through international rapid alert systems, e.g., the European RASSF system (https://ec.europa.eu/food/safety/rasff_en), or alert WHO through the IHR system (28). However, the information is more commonly shared broadly through listservs or data sharing boards, e.g., the European Center for Disease Prevention & Control (ECDC) EPIS system (29), the WHO INFOSAN (https://www.who.int/foodsafety/areas_work/infosan/en/) or the PulseNet International forum (30). The countries who receive this information are expected to report whether they are investigating a similar outbreak or see the frequency of the outbreak strain at a higher than usual rate in their surveillance of clinical and non-human surveillance isolates. If a country routinely uses a low discriminatory subtyping method for laboratory surveillance, e.g., species or serotype, this kind of comparison is insensitive and countries with one or a few outbreak related isolates are likely to overlook them. For instance, two of the outbreaks described here, the kratom and turtle-associated *Salmonella* outbreaks, were linked to globally distributed vehicles and yet, only two countries reported cases associated with turtle outbreak and no cases linked to kratom were detected outside the US. Another weakness of the international inquiry approach is that the comparison is not performed until an investigation is well under way in one

country thereby delaying the investigation. Any country should ideally be able to access subtyping information on isolates from other countries in order to recognize international outbreaks fast. Except for the US and Canada who since 2005 have had access to each other's PulseNet databases, no other countries shared molecular surveillance data this way in real-time until the advent of WGS.

The potential of WGS to transform detection and investigation of international outbreaks was realized already in 2011 when scientists from what was later established as the Global Microbial Identifier (GMI) initiative met with the European commission in Brussels. The outcome of the meeting was published as a white paper (31). GMI envisions a global system of DNA genome databases for microbial and infectious disease identification and diagnostics fully embracing the One-Health concept. Sharing of surveillance sequence data with the global scientific community supports the mission of public health institutions and the One Health concept by facilitating early recognition and investigation of international outbreaks that a country is impacted by and therefore need to know about in order to act to protect its citizens. A global system for sharing of genomic data will benefit those tackling individual problems at the frontline, clinicians, veterinarians, environmental scientists, as well as policy-makers, regulators, and industry. By enabling access to this global resource, a professional response on health threats will be within reach of all countries with basic laboratory infrastructure (http://www.globalmicrobialidentifier.org/). PulseNet expanded on that vision in 2017 (32) suggesting a global system of databases containing data extracted from raw sequences of foodborne pathogens using standard analytical pipelines including the cgMLST pipelines used by PulseNet USA in this paper. The advantage of storing data extracted using standardized methods is 2-fold, (i) the data volume is greatly reduced enabling its exchange over slow internet connections, which is still the standard in many developing countries, and (ii) the data are standardized and can be used with minimal additional processing by any PulseNet participant ensuring fast comparison of data from databases in different regions of the world. Also, similar to PulseNet practices, realizing this global vision would likely be aided by additional laboratories submitting raw sequence files of all isolates obtained as part of routine surveillance in real-time to public repositories, e.g., the European Nucleotide Archive (ENA), the DNA Data Bank of Japan (DDBJ) or the GenomeTrakr databases in the Sequence Read Archive (SRA) at the National Center for Bioinformatics Information (NCBI). However, this is currently not possible for institutions in many countries for different reasons, e.g., protection of personal identifiable information (PII), intellectual property rights, or protection against scientific parasitism, i.e., publication of analyzed data generated by others without permission. The federal agencies in the US including CDC, FDA, and the US Department of Agriculture's (USDA's) Food Safety Inspection Service (FSIS) have uploaded all their raw sequences to the SRA in real time for the last 5 years without any noticeable adverse effects. An increasing number of agencies and institutions in other countries are now following suit, but there is still a long way to go before this is done by all countries.

CONCLUSIONS

Outbreaks linked to animals and environmental sources can be challenging to recognize by laboratory surveillance by WGS because they are often polyclonal and more diverse than observed in typical point source outbreaks. The availability and use of supporting epidemiological information and microbiological information from non-clinical sources may be critical for their recognition and successful investigation. In the future, linking public health and food regulatory databases that include patient and food/feed/ingredient demographics, interview data, and microbiological data to national and international databases containing diverse types of other information, e.g., trade and distribution of different commodities, including live animals, raw agricultural products, processed foods, and international travel information, to name a few, could be used in a "big data" approach to detect and investigate outbreaks sometimes even before they become apparent by traditional syndromic or laboratory surveillance. However, critical first steps toward this vision include collection and sequencing of isolates from animal and environmental sources and all countries agree to make all their WGS surveillance data available to the others as they are generated before an outbreak is suspected.

AUTHOR CONTRIBUTIONS

PG-S conceived the paper and was the responsible writer. JC-A, JF, JH, LJ, ZK, and BT analyzed the sequences, interpreted the data, and contributed to the writing of the manuscript. JB, MN, and CS contributed with interpretation of the data and the writing of the manuscript.

FUNDING

All authors are employees of the CDC and the investigations presented here were conducted as part their daily jobs.

ACKNOWLEDGMENTS

Microbiologists, epidemiologists, and environmental scientists in state and local public health departments in the US, Chile and Luxembourg, and FDA and USDA/FSIS are thanked for their contributions to the outbreak investigations.

REFERENCES

1. Havelaar AH, Kirk MD, Torgerson PR, Gibb HJ, Hald T, Lake RJ, et al. World Health Organization global estimates and regional comparisons of the burden of foodborne disease in 2010. *PLoS Med.* (2015) 12:e1001923. doi: 10.1371/journal.pmed.1001923
2. Jones TF, Yackley J. Foodborne disease outbreaks in the United States: a historical overview. *Foodborne Pathog Dis.* (2018) 15:11–5. doi: 10.1089/fpd.2017.2388

3. Scallan E, Griffin PM, Angulo FJ, Tauxe RV, Hoekstra RM. Foodborne illness acquired in the United States–unspecified agents. *Emerg Infect Dis*. (2011) 17:16–22. doi: 10.3201/eid1701.P21101
4. Schurch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clin Microbiol Infect*. (2018) 24:350–4. doi: 10.1016/j.cmi.2017.12.016
5. Carrico JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect*. (2018) 24:342–9. doi: 10.1016/j.cmi.2017.12.015
6. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect*. (2018) 24:335–41. doi: 10.1016/j.cmi.2017.10.013
7. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, et al. The use of next generation sequencing for improving food safety: translation into practice. *Food Microbiol*. (2019) 79:96–115. doi: 10.1016/j.fm.2018.11.005
8. Waldram A, Dolan G, Ashton PM, Jenkins C, Dallman TJ. Epidemiological analysis of *Salmonella* clusters identified by whole genome sequencing, England and Wales 2014. *Food Microbiol*. (2018) 71:39–45. doi: 10.1016/j.fm.2017.02.012
9. Reimer AR, Domselaar GV, Stroika S, Walker M, Kent H, Tarr C, et al. Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa. *Emerg Infect Dis*. (2011) 17:2113–21. doi: 10.3201/eid1711.110794
10. Moura A, Tourdjman M, Leclercq A, Hamelin E, Laurent E, Fredriksen N, et al. Real-time whole-genome sequencing for surveillance of *Listeria monocytogenes*, France. *Emerg Infect Dis*. (2017) 23:1462–70. doi: 10.3201/eid2309.170336
11. Holmes A, Allison L, Ward M, Dallman TJ, Clark R, Fawkes A, et al. Utility of whole-genome sequencing of *Escherichia coli* O157 for outbreak detection and epidemiological surveillance. *J Clin Microbiol*. (2015) 53:3565–73. doi: 10.1128/JCM.01066-15
12. Grad YH, Lipsitch M, Feldgarden M, Arachchi HM, Cerqueira GC, Fitzgerald M, et al. Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci USA*. (2012) 109:3065–70. doi: 10.1073/pnas.1121491109
13. Chattaway MA, Chandra N, Painset A, Shah V, Lamb P, Acheampong E, et al. Genomic approaches used to investigate an atypical outbreak of *Salmonella* Adjame. *Microb Genom*. (2019) 5:e000248. doi: 10.1099/mgen.0.000248
14. Wang YU, Pettengill JB, Pightling A, Timme R, Allard M, Strain E, et al. Genetic diversity of *Salmonella* and *Listeria* isolates from food facilities. *J Food Prot*. (2018) 81:2082–9. doi: 10.4315/0362-028X.JFP-18-093
15. Ribot EM, Freeman M, Hise KB, Gerner-Smidt P. PulseNet: entering the age of next generation sequencing. *Foodborne Pathog Dis*. (In press).
16. Chen YI, Burall LS, Macarisin D, Pouillot R, Strain EDE, et al. Prevalence and level of *Listeria monocytogenes* in ice cream linked to a listeriosis outbreak in the United States. *J Food Prot*. (2016) 79:1828–32. doi: 10.4315/0362-028X.JFP-16-208
17. Chen Y, Luo Y, Curry P, Timme R, Melka D, Doyle M, et al. 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*. (2017) 12:e0171389. doi: 10.1371/journal.pone.0171389
18. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, et al. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. *Nat Microbiol*. (2016) 2:16185. doi: 10.1038/nmicrobiol.2016.185
19. Ragon M, Wirth T, Hollandt F, Lavenir R, Lecuit M, Le Monnier A, et al. A new perspective on *Listeria monocytogenes* evolution. *PLoS Pathog*. (2008) 4:e1000146. doi: 10.1371/journal.ppat.1000146
20. Gieraltowski L, Higa J, Peralta V, Green A, Schwensohn C, Rosen H, et al. National outbreak of multidrug resistant *Salmonella* Heidelberg infections linked to a single poultry company. *PLoS ONE*. (2016) 11:e0162369. doi: 10.1371/journal.pone.0162369
21. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of *Salmonella*. *PLoS Genet*. (2018) 14:e1007261. doi: 10.1371/journal.pgen.1007261
22. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother*. (2012) 67:2640–4. doi: 10.1093/jac/dks261
23. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, et al. *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrob Agents Chemother*. (2014) 58:3895–903. doi: 10.1128/AAC.02412-14
24. Cody AJ, Bray JE, Jolley KA, McCarthy ND, Maiden MCJ. Core genome multilocus sequence typing scheme for stable, comparative analyses of *Campylobacter jejuni* and *C. coli* human disease isolates. *J Clin Microbiol*. (2017) 55:2086–97. doi: 10.1128/JCM.00080-17
25. Bosch S, Tauxe RV, Behravesh CB. Turtle-associated salmonellosis, United States, 2006–2014. *Emerg Infect Dis*. (2016) 22:1149–55. doi: 10.3201/eid2207.150685
26. Gambino-Shirley K, Stevenson L, Concepcion-Acevedo J, Trees E, Wagner D, Whitlock L, et al. Flea market finds and global exports: four multistate outbreaks of human *Salmonella* infections linked to small turtles, United States-2015. *Zoonoses Public Health*. (2018) 65:560–8. doi: 10.1111/zph.12466
27. Ford L, Miller M, Cawthorne A, Fearnley E, Kirk M. Approaches to the surveillance of foodborne disease: a review of the evidence. *Foodborne Pathog Dis*. (2015) 12:927–36. doi: 10.1089/fpd.2015.2013
28. Tucker JB. Updating the International Health Regulations. *Biosecur Bioterror*. (2005) 3:338–47. doi: 10.1089/bsp.2005.3.338
29. Gossner CM. New version of the Epidemic Intelligence Information System for food- and waterborne diseases and zoonoses (EPIS-FWD) launched. *Euro Surveill*. (2016) 21:30422. doi: 10.2807/1560-7917.ES.2016.21.49.30422
30. Swaminathan B, Gerner-Smidt P, Ng LK, Lukinmaa S, Kam KM, Rolando S, et al. Building PulseNet International: an interconnected system of laboratory networks to facilitate timely public health recognition and response to foodborne disease outbreaks and emerging foodborne diseases. *Foodborne Pathog Dis*. (2006) 3:36–50. doi: 10.1089/fpd.2006.3.36
31. Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour MW, Harmsen D, et al. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. *Emerg Infect Dis*. (2012) 18:e1. doi: 10.3201/eid/1811.120453
32. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global foodborne disease surveillance. *Eurosurveillance*. (2017) 22:30544. doi: 10.2807/1560-7917.ES.2017.22.23.30544

Disclaimer: The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

CORRECTION
published: 06 December 2019
doi: 10.3389/fpubh.2019.00365

Corrigendum: Whole Genome Sequencing: Bridging One-Health Surveillance of Foodborne Diseases

Peter Gerner-Smidt[1], John Besser[1], Jeniffer Concepción-Acevedo[1], Jason P. Folster[1], Jasmine Huffman[1], Lavin A. Joseph[1], Zuzana Kucerova[1], Megin C. Nichols[2], Colin A. Schwensohn[2] and Beth Tolar[1]*

[1] The Enteric Diseases Laboratory Branch, Centers for Disease Control and Prevention, Atlanta, GA, United States,
[2] The Outbreak Response and Prevention Branch, Centers for Disease Control and Prevention, Atlanta, GA, United States

Keywords: whole genome sequencing (WGS), outbreak, one health, zoonotic, food, environment, animals, investigation

OPEN ACCESS

Approved by:
Frontiers Editorial Office, Frontiers Media SA, Switzerland

****Correspondence:***
Peter Gerner-Smidt
plg5@cdc.gov

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *02 October 2019*
Accepted: *18 November 2019*
Published: *06 December 2019*

Citation:
Gerner-Smidt P, Besser J, Concepción-Acevedo J, Folster JP, Huffman J, Joseph LA, Kucerova Z, Nichols MC, Schwensohn CA and Tolar B (2019) Corrigendum: Whole Genome Sequencing: Bridging One-Health Surveillance of Foodborne Diseases. Front. Public Health 7:365. doi: 10.3389/fpubh.2019.00365

A Corrigendum on

Whole Genome Sequencing: Bridging One-Health Surveillance of Foodborne Diseases
by Gerner-Smidt, P., Besser, J., Concepción-Acevedo, J., Folster, J. P., Huffman, J., Joseph, L. A., et al. (2019). Front. Public Health 7:172. doi: 10.3389/fpubh.2019.00172

In the original article, there was a mistake in **Figure 1** and **Figure 2** as published. The graphics used are different than those originally submitted. The corrected **Figure 1** and **Figure 2** appear below.

The authors apologize for this error and state that this does not change the scientific conclusions of the article in any way. The original article has been updated.

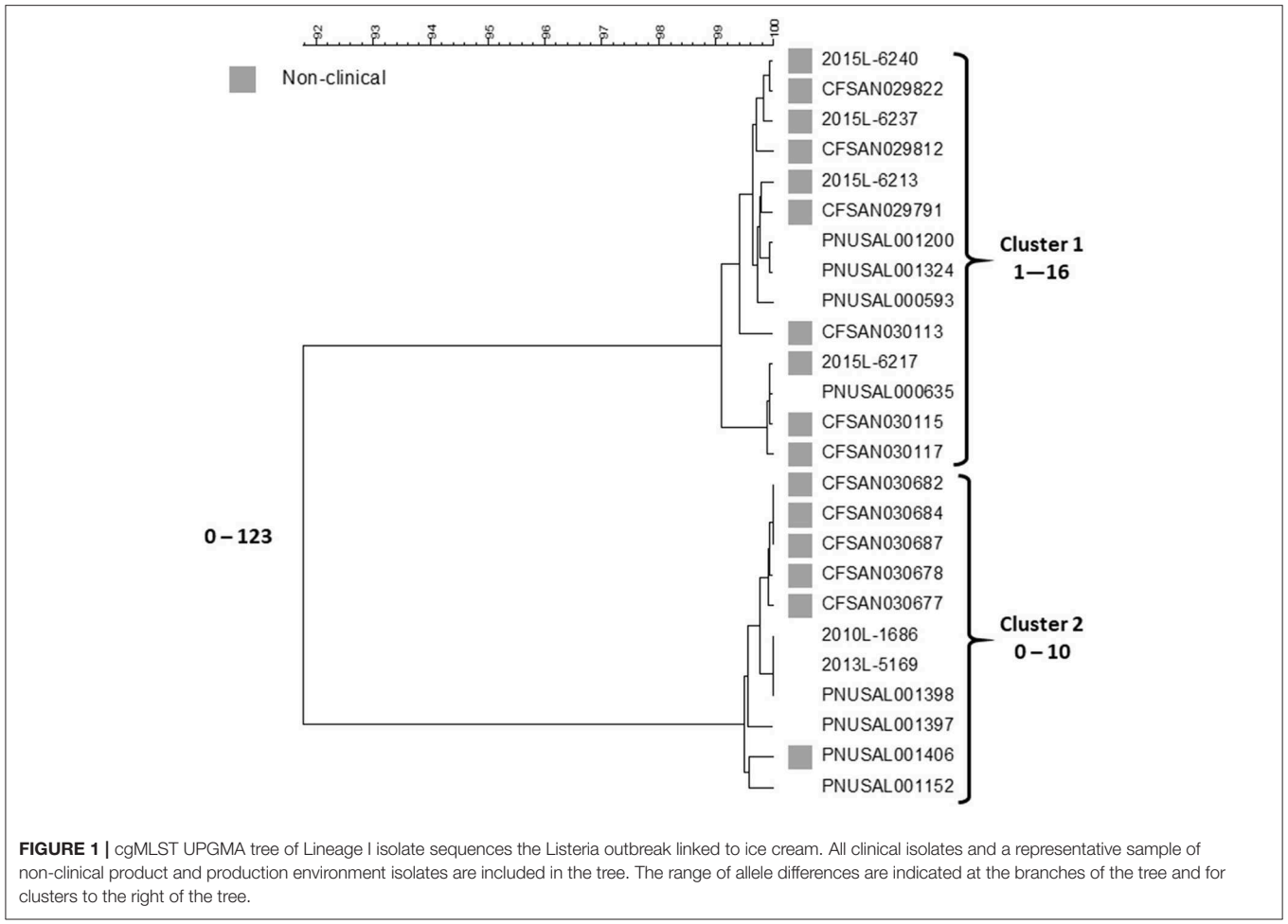


FIGURE 1 | cgMLST UPGMA tree of Lineage I isolate sequences the Listeria outbreak linked to ice cream. All clinical isolates and a representative sample of non-clinical product and production environment isolates are included in the tree. The range of allele differences are indicated at the branches of the tree and for clusters to the right of the tree.

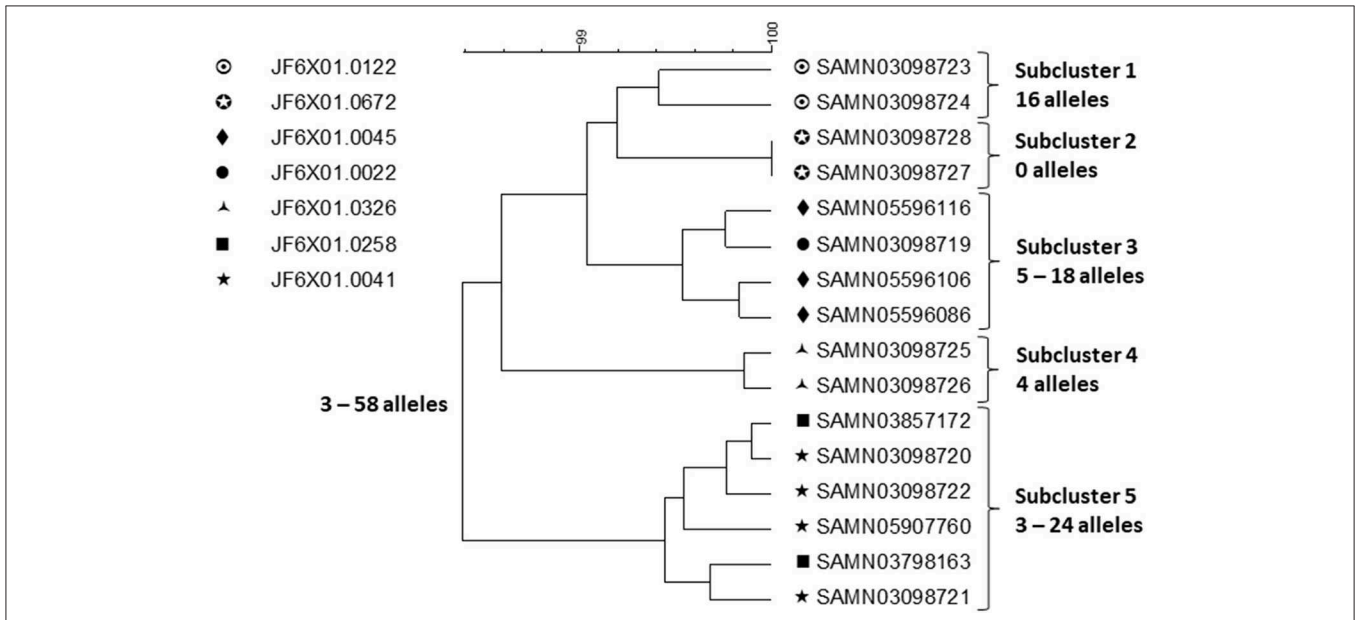


FIGURE 2 | cgMLST UPGMA tree of a representative sample of sequences of Salmonella ser. Heidelberg isolates displaying the full WGS diversity and representing the seven PFGE patterns from the outbreak associated with chicken produced by Company A. The range of allele differences are indicated at the branches of the tree and subclusters to the right of the tree.

frontiers in Public Health

REVIEW
published: 02 August 2019
doi: 10.3389/fpubh.2019.00208

Advances in Visualization Tools for Phylogenomic and Phylodynamic Studies of Viral Diseases

*Kristof Theys, Philippe Lemey, Anne-Mieke Vandamme and Guy Baele**

Department of Microbiology, Immunology and Transplantation, Rega Institute for Medical Research, Clinical and Epidemiological Virology, KU Leuven, Leuven, Belgium

Genomic and epidemiological monitoring have become an integral part of our response to emerging and ongoing epidemics of viral infectious diseases. Advances in high-throughput sequencing, including portable genomic sequencing at reduced costs and turnaround time, are paralleled by continuing developments in methodology to infer evolutionary histories (dynamics/patterns) and to identify factors driving viral spread in space and time. The traditionally static nature of visualizing phylogenetic trees that represent these evolutionary relationships/processes has also evolved, albeit perhaps at a slower rate. Advanced visualization tools with increased resolution assist in drawing conclusions from phylogenetic estimates and may even have potential to better inform public health and treatment decisions, but the design (and choice of what analyses are shown) is hindered by the complexity of information embedded within current phylogenetic models and the integration of available meta-data. In this review, we discuss visualization challenges for the interpretation and exploration of reconstructed histories of viral epidemics that arose from increasing volumes of sequence data and the wealth of additional data layers that can be integrated. We focus on solutions that address joint temporal and spatial visualization but also consider what the future may bring in terms of visualization and how this may become of value for the coming era of real-time digital pathogen surveillance, where actionable results and adequate intervention strategies need to be obtained within days.

Keywords: visualization, phylogenetics, phylogenomics, phylodynamics, infectious disease, epidemiology, evolution

OPEN ACCESS

Edited by:
Vitali Sintchenko, University of Sydney, Australia

Reviewed by:
Denis Baurain, University of Liège, Belgium
Sergey Eremin, World Health Organization, Switzerland
John-Sebastian Eden, University of Sydney, Australia

****Correspondence:***
Guy Baele
guy.baele@kuleuven.be

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *07 April 2019*
Accepted: *12 July 2019*
Published: *02 August 2019*

Citation:
Theys K, Lemey P, Vandamme A-M and Baele G (2019) Advances in Visualization Tools for Phylogenomic and Phylodynamic Studies of Viral Diseases. Front. Public Health 7:208. doi: 10.3389/fpubh.2019.00208

1. VIRUS EPIDEMIOLOGY AND EVOLUTION

Despite major advances in drug and vaccine design in recent decades, viral infectious diseases continue to pose serious threats to public health, both as globally well-established epidemics of e.g., Human Immunodeficiency Virus Type 1 (HIV-1), Dengue virus (DENV) or Hepatitis C virus (HCV), and as emerging or re-emerging epidemics of e.g., Zika virus (ZIKV), Middle East Respiratory Syndrome Coronavirus (MERS-CoV), Measles virus (MV), or Ebola virus (EBOV). Efforts to reconstruct the dynamics of viral epidemics have gained considerable attention as they may support the design of optimal disease control and treatment strategies (1, 2). These analyses are able to provide answers to questions on the diverse processes underlying disease epidemiology, including the (zoonotic) origin and timing of virus outbreaks, drivers of spatial spread, characteristics of transmission clusters and factors contributing to enhanced viral pathogenicity and adaptation (3–5).

Molecular epidemiological techniques have proven to be important and effective in informing public health and therapeutic decisions in the context of viral pathogens (6, 7), given that most of the viruses with a severe global disease burden are characterized by high rates of evolutionary change. These genetic changes are being accumulated in viral genomes on a time scale similar to the one where the dynamics of population genetic and epidemiological processes can be observed, which has lead to the definition of viral phylodynamics as the study of how epidemiological, immunological, and evolutionary processes act and potentially interact to shape viral phylogenies (8). As such, phylogenetic trees constitute a crucial instrument in studies of virus evolution and molecular epidemiology, elucidating evolutionary relationships between sampled virus variants based on the temporal resolution in the genetic data of these fast-evolving viruses that allows resolving their epidemiology in terms of months or years. Through the integration of population genetics theory, epidemiological data and mathematical modeling, insights into epidemiological, immunological, and evolutionary processes shaping genetic variation can be inferred from these phylogenies. The field of phylodynamics has generated new opportunities to obtain a more detailed understanding of evolutionary histories—through time as well as geographic space—and transmission dynamics of both well-established viral epidemics and emerging outbreaks (9, 10).

The ability of molecular epidemiological analyses, and phylodynamic analyses in particular, to fully exploit the information embedded in viral sequence data has significantly improved through a combination of technological innovations and advances in inference frameworks during the past decades. From a data perspective, genomic epidemiology is becoming a standard framework driven by high-throughput sequencing technologies that are associated with reduced costs and increasing turnover. Moreover, the portability and potential of rapid deployment on-site of these new technologies enable the generation of complete genome data from samples within hours of taking the samples (11). This rising availability of whole-genome sequences increases the resolution by which historical events and epidemic dynamics can be reconstructed. From a methodological perspective, new developments in statistical and computational methods along with advances in hardware infrastructure have allowed the analysis of ever-growing data sets, the incorporation of more complex models and the inclusion of information related to sample collection, infected host characteristics and clinical or experimental status (generally known as metadata) (9, 10, 12, 13).

In contrast to a marked increase in the number of software packages targetting increasingly efficient but complex approaches to infer annotated phylogenies by exploiting genomic data and the associated metadata, the intuitive and interactive visualization of their outcomes has not received the same degree of attention, despite being a key aspect in the interpretation and dissemination of the rich information that is inferred. Phylogenies are typically visualized in a rather simplistic manner, with the concept of depicting evolutionary relationships using a tree structure already illustrated in Charles Darwin's notebook (1837) and his seminal book "The Origin of Species" (14). Early phylogenetic tree visualization efforts constituted an integral part of phylogenetic inference software packages and as such were restricted to simply showing the inferred phylogenies on a command line or in a simple text file, often even without an accompanying graphical user interface. The longstanding use of phylogenies in molecular epidemiological analyses has however led to the emergence of increasingly feature-rich visualization tools over time. The advent of the new research disciplines such as phylogenomics and phylodynamics necessitated more complex visualizations in order to accommodate projections of pathogen dispersal onto a geographic map, ancestral reconstruction of various types of trait data and appealing animations of the reconstructed evolution and spread over time. Tree visualizations resulting from these analyses are also complemented by visual reconstructions of other important aspects of the model reconstructions, such as population size dynamics over time, transmission networks and estimates of ancestral states for traits of interest throughout the tree (15).

Across disciplines, adequate visualizations are pivotal to communicate, disseminate and translate research findings into meaningful information and actionable insights for clinical, research and public health officials. The aim to improve data-driven decision making fits within a broader scope to establish a universal data visualization literacy (16). To this end, enhancing collaborations and dissemination of visualizations is increasingly achieved through sharing of online resources for hosting annotated tree reconstructions (17), online workspaces (18) and continuously updated pipelines that accommodate increasing data flow during infectious disease outbreaks (19) (see further sections for more information and examples of these packages). Given the plethora of options for presenting and visualizing results, and its importance for effectively communicating with a wide audience, choosing the appropriate representation and visualization strategy can be challenging. Recent work on this topic focuses on navigating through all the available visualization options by offering clear guidelines on how to turn large datasets into compelling and aesthetically appealing figures (20).

2. A FRAMEWORK FOR VISUALIZATION

A large array of software packages for performing phylogenetic and phylodynamic analyses have emerged in the last decade, in particularly for fast-evolving RNA viruses [see (10) for a recent overview]. A more recent but similar trend can be seen for methodologies and applications aimed at visualization of the output of these frameworks. In addition to the need to communicate these outputs in a visual manner, an increasing recognition for the added value of adequate visualization for surveillance, prevention, control and treatment of viral infectious diseases has resulted into the merging of data analytics and visualization, with the visualization aspect being increasingly considered as an elementary component within all-round analysis platforms. This review illustrates the evolution in phylogenetically-informed visualization modalities for evolutionary inference and epidemic modeling based on viral sequence data, evolving from an initial purpose to serve

basic interpretation of the results to an in-depth translation of complex information into usable data for virologists, researchers and public health officials alike. Novel features and innovative approaches often stem from a community need, which can be translated into a specific challenge to be addressed by current and future software applications. Throughout this article, we discuss some of the major bottlenecks for interpretation and visualization of phylodynamic results, and subsequently solutions that have addressed or can address these challenges.

A closer inspection of how tools for manipulation, visualization and interpretation of evolutionary scenarios have steadily grown over time reveals different trends of interest. First, visualization needs for phylodynamic analyses are very heterogeneous in nature, driven by the intrinsic objective to better understand viral disease epidemiology. Due to the increasing complexity and interactivity of the various aspects that make up phylodynamic analyses, the gradual change in visualization tools has resulted in a wide but incomplete range of solutions provided (illustrated by the Wikipedia list of phylogenetic tree visualization software[1]). Software applications for phylodynamic analyses have extended into investigations of population dynamics over time, trait evolution and spatio-temporal dispersal, while still using a phylogenetic tree as their core concept. While we will focus predominantly on the concept of a phylogenetic tree as the backbone of phylodynamic visualization, these analyses also produce other types of output that go beyond visualizing phylogenies, especially when it comes to trait data reconstruction. Second, the continuing advances in visualization—which try to keep up with increasing complexities in the statistical models employed—not only result in more features being available for end users to exploit, they may also come at an increased cost in terms of usability and responsiveness. Formats for input and output files have increased in complexity, from simple text files to XML specifications and (Geo)JSON file formats for geographical features. Reading, understanding and editing such files may prove to be a challenging task for practitioners. However, most visualization tools do not expose these complexities to their users and offer an intuitive point-and-click interface and/or drag-and-drop functionality for customizing the visualization (18). Despite such intuitive interactivity, intricate knowledge and a certain amount of programming/scripting experience is often required for those users who want to customize and/or extend their visualization beyond what the application has to offer. Third, visualization goals tend to become context-dependent in that not all phylodynamic analyses deal with the same sense of urgency, with established epidemics requiring different prevention and treatment strategies than outbreak detection and surveillance. For example, in established epidemics (e.g., HIV-1) thefocus may be on identifying (important) clusters within a very large phylogeny (17), whereas analyses in ongoing outbreaks often determine whether newly generated sequences correspond to strains of the virus known to circulate in a certain region and try to establish spillover from animal reservoirs (21). Finally, despite the major achievements so far, visualization tools are reaching the limits of their capacity to comprehensibly present analysis results of large datasets. Promising developments and strategies are becoming available that move visualization beyond the goal of communicating and synthesizing results, and actively play an important role in providing analytics to better understand evolutionary and demographic processes fueling viral dispersal and pathogenicity.

[1]https://en.wikipedia.org/wiki/List_of_phylogenetic_tree_visualization_software

3. VISUALIZATION CHALLENGES AND SOLUTIONS

Phylogenetic tree visualizations have played a central role since the earliest evolutionary and molecular epidemiological analyses of fast-evolving viral pathogens. The first computer programs aimed at constructing phylogenies [e.g., PAUP*; (22, 23), and PHYLIP; (24)] were only equipped with minimal tree drawing and printing facilities, limited by the available operating systems and programming languages of that time. Standalone, phylogenetically-oriented programs [e.g., MUST; (25) and later on Treeview; (26)] were specifically developed to interact with tree reconstruction output and to ease tree editing and viewing. Even as phylogenetic inference became inherently more sophisticated, for example with the development of Bayesian phylogenetic inference and the release of initial versions of MrBayes (27) which contained sophisticated search strategies to ensure finding the optimal set of phylogenetic trees, these software packages still contained their own text-based tree visualization component(s).

However, over time a wide range tree visualization software has been released, offering a continuous increase of tree visualization and manipulation functionalities. These packages have been developed as either standalone software packages or have been integrated into larger data management and analysis platforms [e.g., MEGA (28)]. The numerous all-round programs available to date offer a range of similar basic tree editing capabilities including the coloring and formatting of tree nodes, edges and labels, the addition of numerical or textual annotations, searching for specific taxa as well as the re-rooting, rotation and collapsing of clades. Different tree formats can be imported and again exported to various textual and graphical formats (e.g., vector-based formats: portable document format (pdf), encapsulated postscript (eps), scalable vector graphics (svg), …). A limited set of applications provide more advanced visualization functionalities that enable interactive visualization and management of highly customized and annotated phylogenetic trees. Nevertheless, major hurdles still exist that hinder adequate communication and interpretation of phylodynamic analyses. These hurdles mainly relate to the scalability of the visualization, highlighting uncertainty associated with the results and the interactive rendering of available metadata. Recent innovative developments attempt to tackle these bottlenecks, although some tools are specifically directed toward addressing a single (visualization) challenge. We here provide an overview of such challenges, along with examples of figures generated by software packages that aim to tackle these

challenges. Note that all of our visualization examples are shown in the *Evolving visualization examples* section below.

First, a major challenge is the ever-increasing size of data sets being analyzed, leading to difficulties with navigating through the resulting phylogenetic trees and to problems with interpreting the inferred dynamics, not only from a computational perspective (e.g., to render large images in a timely manner) but also from the human capability to deal with high levels of detail. Software packages that mainly aim to visualize phylogenetic trees as well as those that target more broad analyses have implemented various solutions to accommodate systematic exploration of large phylogenies. Dendroscope (29) was one of the first visualization tools aimed at large phylogenies, with its own format to save and reopen trees that had been edited graphically, offering a magnifier functionality to focus on specific parts of the (large) phylogeny. Follow-up versions (30) focused on rooted phylogenetic trees and networks, and offered parallel implementations of demanding algorithms for computing consensus trees and consensus networks to increase responsiveness. Phylo.io (31) improves the legibility of large trees by automatically collapsing nodes so that an overview of the tree remains visible at any given time. iTOL [(18), but see below] and IcyTree (32) also provide intuitive panning and zooming utilities that make exploring large phylogenetic trees of many thousands of taxa feasible. The PhyloGeoTool [(17); also see **Figure 4**] eases navigation of large trees by performing an *a priori* iterative clustering of subtrees according to a predefined diversity ratio, as well as pre-rendering the visualization of those subtrees enabling fluent navigation. PastML (33) allows visualizing the tree annotated with reconstructed ancestral states as a zoomable HTML map based on the Cytoscape framework (34). PastView (35) offers synthetic views such as transition maps, integrates comparative analysis methods to highlight agreements or discrepancies between methods of ancestral annotations inference, and is also available as a webserver instance. Grapetree (36) initially collapses branches if there are more than 20,000 nodes in the tree and then uses a static layout that splits the tree layout task into a series of sequential node layout tasks. With the development of many packages targetting the visualization of large phylogenies in recent years, the question arises whether they will continue to be maintained and extended with novel features.

A second challenge lies with the fact that phylogenies represent hypotheses that encompass different sources of error, and the extent of uncertainty at different levels should be presented accordingly. Bootstrapping (37) and other procedures are often used to investigate the robustness of clustering in estimated tree topologies,. Numerical values that express the support of a cluster are generally shown on the internal nodes of a single consensus summary tree [e.g., FigTree; (38)] or by a customized symbol [e.g., iTOL; (18)]. Although conceptually different, posterior probabilities on a maximum clade credibility (MCC) tree, majority consensus tree or other condensed trees from the posterior set of trees resulting from Bayesian phylogenetic inference can be shown in a similar manner. An informative and qualitative approach to represent the complete distribution of rooted tree topologies is provided by DensiTree [(39); also see **Figure 10**], which draws all trees in a set simultaneously and transparently, and the different output visualizations highlight various aspects of tree uncertainty. For time-scaled phylogenetic trees, uncertainty in divergence time estimates of ancestral nodes (e.g., 95% highest posterior density (HPD) intervals) is usually displayed with a horizontal (node) bar (see **Figure 1** for an example). Additionally, ancestral reconstructions of discrete or continuous trait states at the inner nodes of a tree are increasingly facilitated by various probabilistic frameworks, and these inferences are also accompanied by posterior distributions describing uncertainty. To visualize this uncertainty, PastML (33) inserts pie charts at inner nodes to show likely states when reconstructing discrete traits such as the evolutionary history of drug resistance mutations, while SpreaD3 (40) is able to depict uncertainty of continuous traits, e.g., as polygon contours for (geographical) states at the inner nodes [see (40) for an example]. Much like the visualization packages that focus on large phylogenies (see above), the applications listed here have their own specific focus with sometimes limited overlap in functionality.

A third challenge consists of the visual integration of metadata with phylogenetic trees—often in the form of either a discrete and/or continuous trait associated with each sequence—which is in part related to the previous challenge concerning uncertainty of trait reconstructions. Incorporating virus trait information (e.g., drug resistance mutations, treatment activity scores) or host characteristics (e.g., gender, age, risk group) in phylogenetic inference can substantially facilitate the interpretation for end users and accelerate the identification of potential transmission patterns. Tree reconstruction and visualization software generally share a set of basic operations for coloring taxa, branches or clades according to partial or exact label matches. While these annotations can be performed manually using a graphical user interface, this can be time-consuming and is prone to errors. Hence, several software programs offer functionalities to automate the selection and annotation of clades of interest, for example through the use of JavaScript libraries [e.g., PhyD3; (41), SpreaD3; (40)]—also see **Figure 3**—or Python toolkits [e.g., ETE; (42), Baltic; (43)]. Alternatively, drag-and-drop functionality of plain text annotation files generated with user-friendly text editors facilitate this process, as is for example the case in iTOL (18). These scripting visualization frameworks also foster more intense tree editing through their functionalities to annotate inner nodes, clades and individual leaves with charts (pie, line, bar, heatmap, boxplot), popup information, images, colored strips and even multiple sequence alignments. Even more advanced integration efforts entail the superimposition of tree topology with layers of information on geographical maps, such as terrain elevation, type of landcover and human population density [e.g., R package seraphim; (44, 45)].

Finally, visualization and accompanying interpretation are a critical component of infectious disease epidemiological and evolutionary analyses. Indeed, many researchers use visualization software during analyses for data exploration, identifying inconsistencies, and refining their data set to ensure well-supported conclusions regarding an ongoing outbreak. As such, the visualizations themselves are gradually refined and

improved over the course of a research project, with the final figures accompanying a publication often being post-processed versions of the default output of a visualization package or customly designed to attract a wide audience, both through the journal's website and especially social media [see e.g., (5)]. On the other hand, the advent of one-stop platforms [MicroReact; (46) and Nextstrain; (19, 47), also see **Figure 5**] that seamlessly connect the different steps of increasingly complex analyses and visualization of genomic epidemiology and phylodynamics allows automating this process. Applications that are exclusively tailored toward tree manipulation and viewing are starting to offer management services and registration of user accounts [iTOL; (18)], while command-line tools (Gotree; https://github.com/evolbioinfo/gotree) aimed at manipulating phylogenetic trees and inference methods (PASTML; (33) increasingly enable exporting trees that can directly be uploaded to iTOL, supporting the automation of scripting and analysis pipelines.

4. EVOLVING VISUALIZATION EXAMPLES

In the previous sections, we have already covered a wide range of software packages for visualizing phylogenetic trees as well as their associated metadata, which may or may not be used in a joint estimation of sequence and trait data [for an overview of integrating these data types in various inference frameworks for pathogen phylodynamics, we refer to (9)]. We here organize our visualization examples into different broader categories: different approaches toward visualizing associated trait data with a focus on phylogeography (**Figures 1**–**3**), browser-based online applications (**Figures 4**, **5**), applications that use existing libraries such as those available in R, Python and JavaScript for example (**Figures 6**, **7**), non-phylogenetic visualizations typically associated with pathogen phylodynamics (**Figure 8**), and finally custom-written code or applications that focus on assessing phylogenetic uncertainty (**Figures 9**, **10**).

As a first example, we illustrate the development of innovative visualization software packages on the output of a Bayesian phylodynamic analysis of a rabies virus (RABV) data set consisting of time-stamped genetic data along with two discrete trait characteristics per sequence, i.e., the sampling location—in this case the state within the United States from which the sample originated—and the bat host type. This RABV data set comprises 372 nucleoprotein gene sequences from North American bat populations, with a total of 17 bat species sampled between 1997 and 2006 across 14 states in the United States (52). We used BEAST 1.10 (51) in combination with BEAGLE 3 (13) to estimate the time-scaled phylogenetic tree relating the sequences, along with inferring the ancestral locations of the virus using a Bayesian discrete phylogeographic approach (53) and, at the same time, infer the history of host jumping using the same model approach. Upon completion of the analysis, we constructed a maximum clade credibility (MCC) tree from the posterior tree distribution using TreeAnnotator, a software tool that is part of the BEAST distribution. This MCC tree contains at its internal nodes the age estimates of all of the internal nodes, along with discrete probability distributions for the inferred location and host traits at those internal nodes.

Figure 1 shows the visualization of the MCC tree in FigTree, with internal nodes annotated according to the posterior ancestral location state probabilities within the MCC tree file. As expected, one can observe that posterior support for the preferred ancestral location decreases from the observed tips toward the root, in other words the further we go back in time, the more uncertain the inferred location states become. All of the information required to make the FigTree visualization in **Figure 1** is contained within a NEXUS file, containing all of the ancestral trait annotations, which we use as the (only) input for the FigTree (38). The standard Newick file format itself does not contain such trait annotations but remains in popular use for storing phylogenetic trees and is hence supported by most (if not all) phylogenetic visualization packages. In general however, Newick and other older formats (e.g., NEXUS) offer limited expressiveness for storing and visualizing annotated phylogenetic trees and associated data, which has lead to extensions for this format being proposed [e.g., the extended Newick format; (54)]. FigTree allows users to upload annotation information for the sequences in the analyzed alignment in the form of a simple tab-delimited text file, and a parsimony approach can be used to infer the most parsimonious state reconstruction for the internal nodes from those provided for the tips. iTOL (18) is another application that can take an MCC tree as its input file and allows annotating branches and nodes of the phylogenetic tree using descriptions provided through the use of simple text files in which custom visualization options can easily be declared (**Figure 2**). iTOL is even suited for showing very large trees (with more than 10,000 leaves) with Webkit-based browsers—such as Chromium/Google Chrome, Opera and Safari—offering the best performance.

Newer input/output file formats for phylogenetic trees and their accompanying annotations, including the XML-based standards PhyloXML (55) and NeXML (56), have the benefit of being more robust for complex analyses and easier to process and extend. In particular, applications of phylodynamics aimed at reconstruction and interpretation of spatio-temporal histories have become broadly and increasingly applied in viral disease investigations. The incorporation of geographical and phylogenetic uncertainty into molecular epidemiology dynamics is now well-established (53, 57), and dedicated developments from a visualization perspective have soon followed to accommodate the outcomes of these models. Early attempts include the mapping of geo-referenced phylogenetic taxa to their geographical coordinates [e.g., GenGis; (58), Cartographer; (59)], while more recent efforts of joint ancestral reconstruction of geographical and evolutionary histories enable visual summaries of spatial-temporal diffusion through the interactive cartographic projection using GIS- and KML-based virtual globe software (60). The latest developments generalize toward interactive web-based visualization of any phylogenetic trait history and are based on data-driven documents (D3) JavaScript libraries and the JSON format to store geographic and other tree-related information (40). As an example, we have created a web-based visualization of our analyzed RABV data set by loading the obtained MCC tree into the SpreaD3

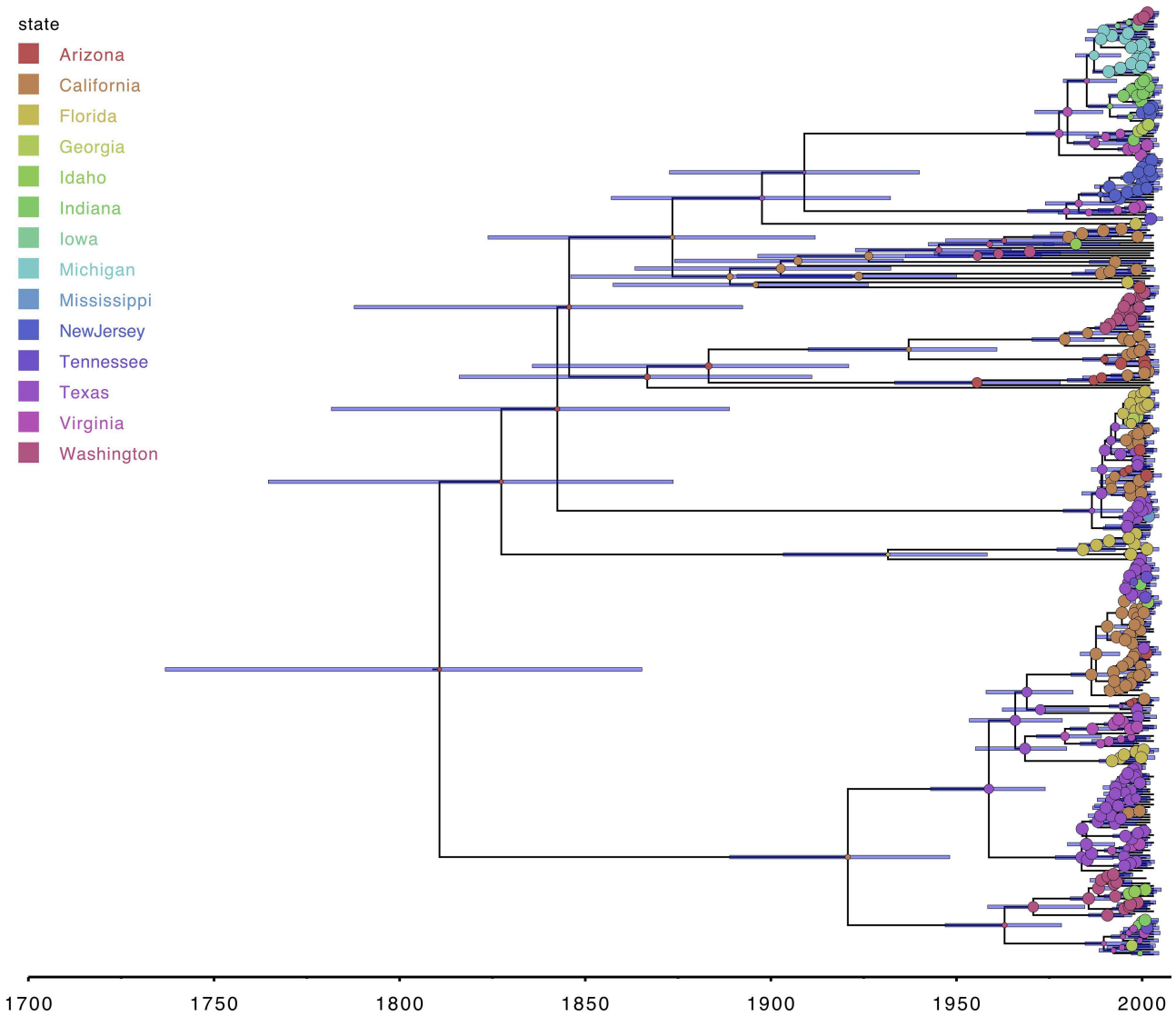


FIGURE 1 | FigTree allows visualizing various tree formats, including maximum clade credibility trees from Bayesian phylogenetic analyses (38). External and internal nodes can easily be annotated using the information in the source tree file, and the time information within the tree allows adding a time axis which facilitates interpretation. Annotations shown here for the RABV data set are the 95% highest posterior density (HPD) age intervals and the most probable ancestral location state at each internal node, with the circle width corresponding to the posterior support for the internal location state reconstruction.

(40) phylodynamic visualization software package (see **Figure 3**). SpreaD3 actually consists of a parsing and a rendering module, with the former obtaining the relevant information out of the MCC tree and the latter converting this information into a (Geo)JSON file format, potentially in combination with a geographic map, which can easily be downloaded from websites offering GeoJSON files of different regions of the world and with different levels of detail. The generated output consists of an in-browser animation that allows tracking a reconstructed epidemic over time using a simple slider bar, with the possibility to zoom into specific areas of the map. In **Figure 3**, we show the reconstructed spread of RABV across the United States at four different time points throughout the epidemic, starting with the estimated location of origin in the state of Arizona and tracking the RABV spread as it disperses to all of the 14 states in our data set.

The SpreaD3 visualization in **Figure 3** is an example of an increasing trend toward web-based software tools that can run in any modern browser, making them compatible with all major operation systems, without requiring any additional software packages to be installed by the user. A distinction can be made between browser-based tools that are able to work without internet access [Phylocanvas; (http://phylocanvas.org), phylotree.js; (61), IcyTree; (32), SpreaD3; (40), PhyloGeoTool; (17), see **Figure 4**] or that are only accessible online [iTOL; (18), phylo.io; (31)]. Web-based visualization platforms enhance

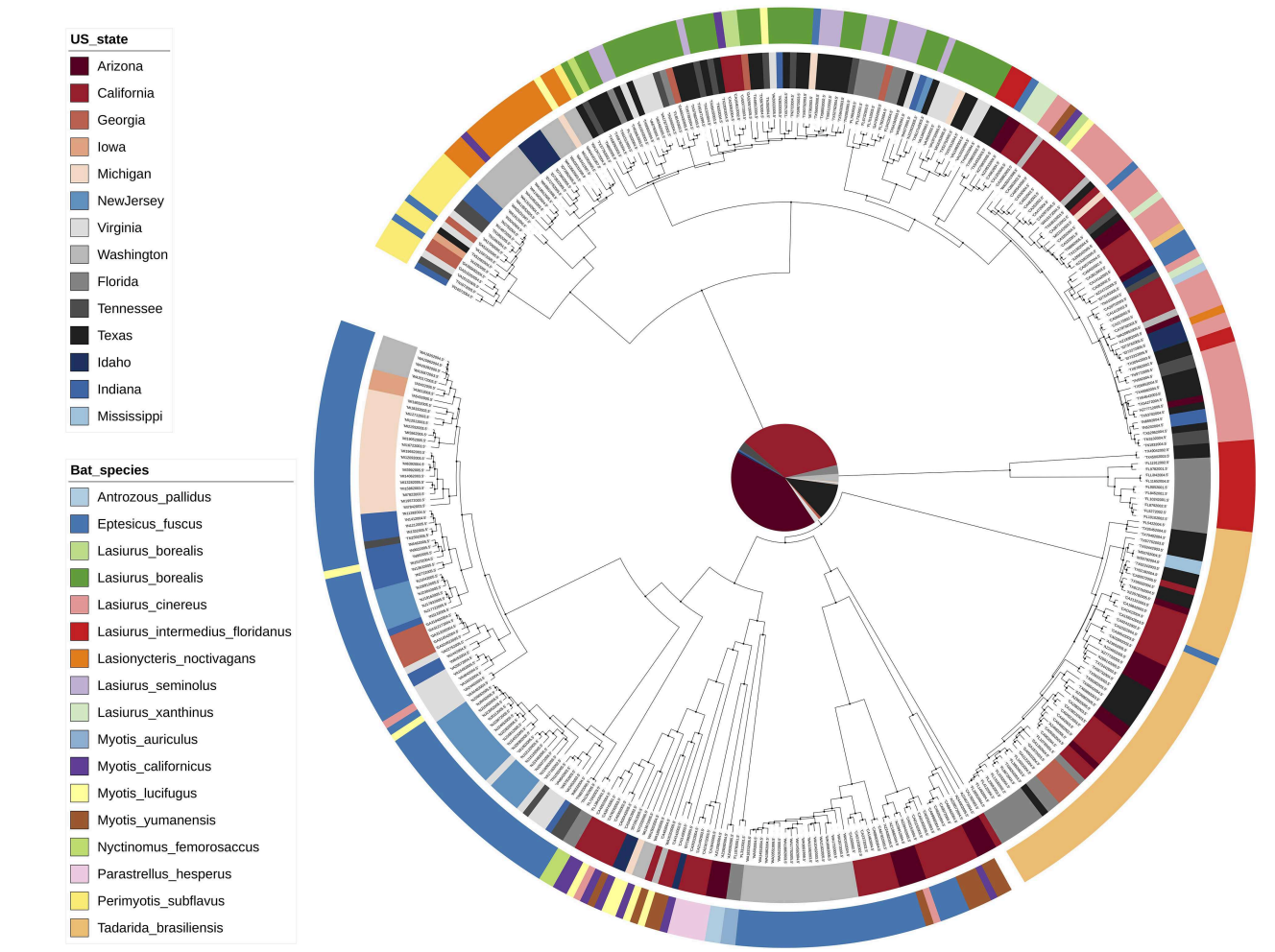


FIGURE 2 | Interactive tree of life [iTOL; (18)] visualization of the MCC tree for RABV. Rather than exploiting the annotations within an MCC tree, iTOL allows importing external text files with annotations through an easy drag-and-drop interface. We have here colored the tip nodes according to the bat host species (outer circle) as well as the sampling location (inner circle) corresponding to each sample. Many visual aspects can be set this way and an extensive online help page is available.

collaborations and output dissemination in a very efficient and simple manner through their ability to share web links of complex and pre-annotated tree visualizations. Transferring genomic data and associated data to an online service may invoke privacy issues which is not the case for tools that execute data processing purely on the client side. By contrast, online accessible visualization tools such as iTOL (18) offer tree management possibilities to organize and save different projects, annotated datasets and trees for their users. These online packages typically also provide export functionalities to facilitate the production of publication-quality and high-resolution illustrations [see also MrEnt; (62), Mesquite; (59)], directed toward end-users with minimal programming experience.

SpreaD3 also illustrates the growing movement toward animated visualizations over time and (geographic) space and as such focuses entirely on the visualization aspect of pathogen phylodynamics. Recently, entire pipelines have emerged that include data curation, analysis and visualization components, with Nextstrain as its most popular example (19). On the data side, Python scripts maintain a database of available sequences and related metadata, sourced from public repositories as well as GitHub repositories and other (more custom-made) sources of genomic data. Fast heuristic tools enable performing phylodynamic analysis including subsampling, aligning and phylogenetic inference, dating of ancestral nodes and discrete trait geographic reconstruction, capturing the most likely transmission events. The accompanying Nextstrain website (https://nextstrain.org/) provides a continually-updated view of publicly available data alongside visualization for a number of pathogens such as West Nile virus, Ebola virus, and Zika virus. For the latter virus, we provide the currently available data visualization in Nextstrain (at time of submission) in **Figure 5**, showing a color-coded time-scaled maximum-likelihood tree alongside an animation of Zika geographic transmissions over

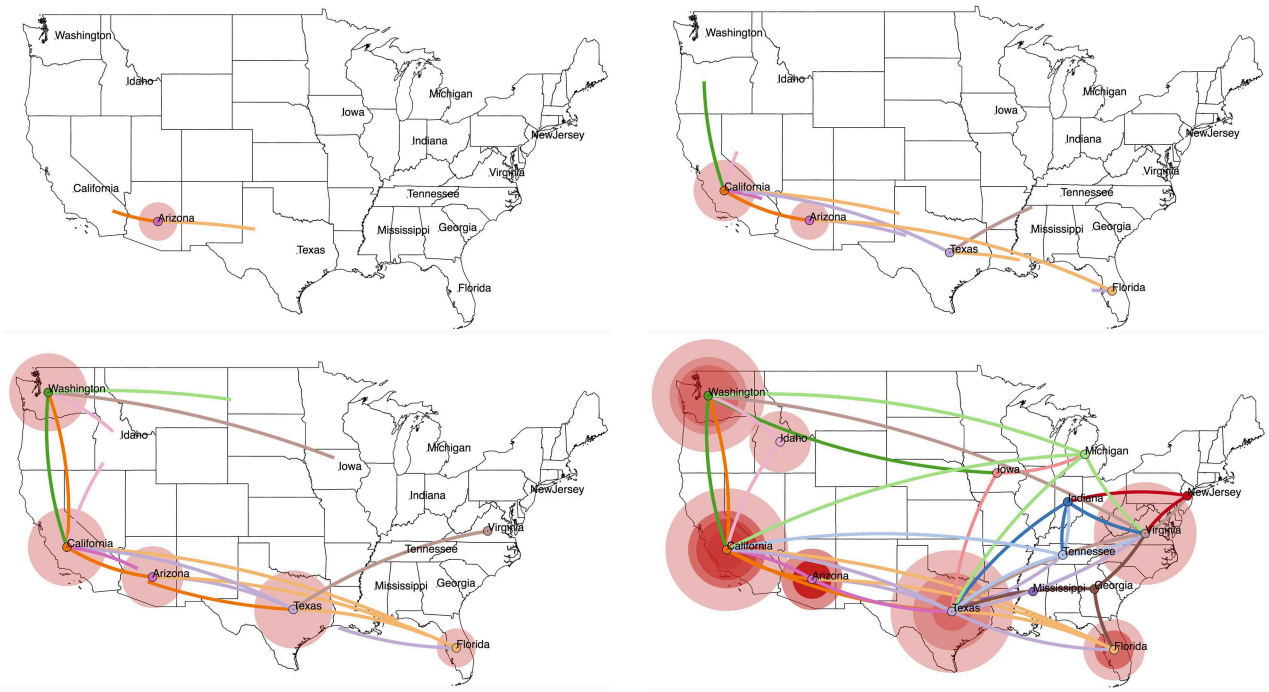


FIGURE 3 | Projecting an MCC tree onto a geographic map using SpreaD3 (40). In a discrete phylogeography setting, as is the case here, the ancestral location states are combined with coordinates corresponding to the states in the US from which the RABV samples were obtained. We use centroid coordinates for the US states to enable this visualization. SpreaD3 animates the reconstructed virus dispersal over time, and we here show four snapshots (starting from the estimated origin of the epidemic at the root node) that capture the reconstructed dispersal over time and geographic space, i.e., in 1860, 1940, 1980, and the "present" (mid 2005). The transitions between the US states are colored according to the US state of destination for that particular transition, whereas the size of the circles around a location is proportional to the number of lineages that maintain that location.

time as well as the genetic diversity across the genome. Analysis of such outbreaks relies on public sharing of data, and Nextstrain has taken the lead to address data sharing concerns by preventing access to the raw genome sequences, and by clearly indicating the source of each sequence, while allowing derived data—such as the inferred phylogenetic trees—to be made publicly available. We note that these animated visualizations by their very nature do not easily yield publication-ready figures, requiring alternative approaches to be devised. Animations resulting from software packages such as SPREAD, SpreaD3 and Nextstrain can be hosted on the authors' website or they can be captured into a video file format and uploaded as supplementary materials onto the journal website. Alternatively, screenshots of the animation can be taken at relevant time points throughout the visualization and subsequently post-processed to include in the main or supplementary publication text.

Finally, browser-based packages such as SpreaD3 employ JavaScript libraries (e.g., D3) to produce dynamic, interactive data visualizations in web browsers, known specifically for allowing great control over the final visualization. Custom programs are also typically written in R as a long list of popular R libraries are readily available, with ggplot2 quickly rising to popularity and finding use in both R and Python languages. A system for declaratively creating graphics based on *The Grammar of Graphics* (63), ggplot2 was built for making professional looking figures with minimal programming efforts. **Figure 6** shows an example of ggtree, which extends ggplot2 and is designed for visualization and annotation of phylogenetic trees with their covariates and other associated data (48). A recent software package that is implemented in JavaScript and Python is PastML (33), which uses the Cytoscape.js library (64) for visualizing phylogenetic trees (**Figure 7**). Given that these types of libraries contain many tried-and-tested functions that save substantial time when creating novel software packages, future visualization efforts are likely to see increased usage of readily available visualization libraries in programming languages such as R, Python and JavaScript.

5. OTHER COMMON VISUALIZATIONS IN PHYLODYNAMICS

Phylogenies reconstructed from viral sequence data and their corresponding annotated tree-like drawings and animations lie at the heart of many evolutionary and epidemiological studies that involve phylogenomics and phylodynamics applications. Additional graphical output can be generated using visualization packages that focus on other aspects than the estimated phylogeny, but that are however in some manner dependent on the phylogeny. Coalescent-based phylodynamic models that connect population genetics theory to genomic data can infer the demographic history of viral populations (65), and plots of

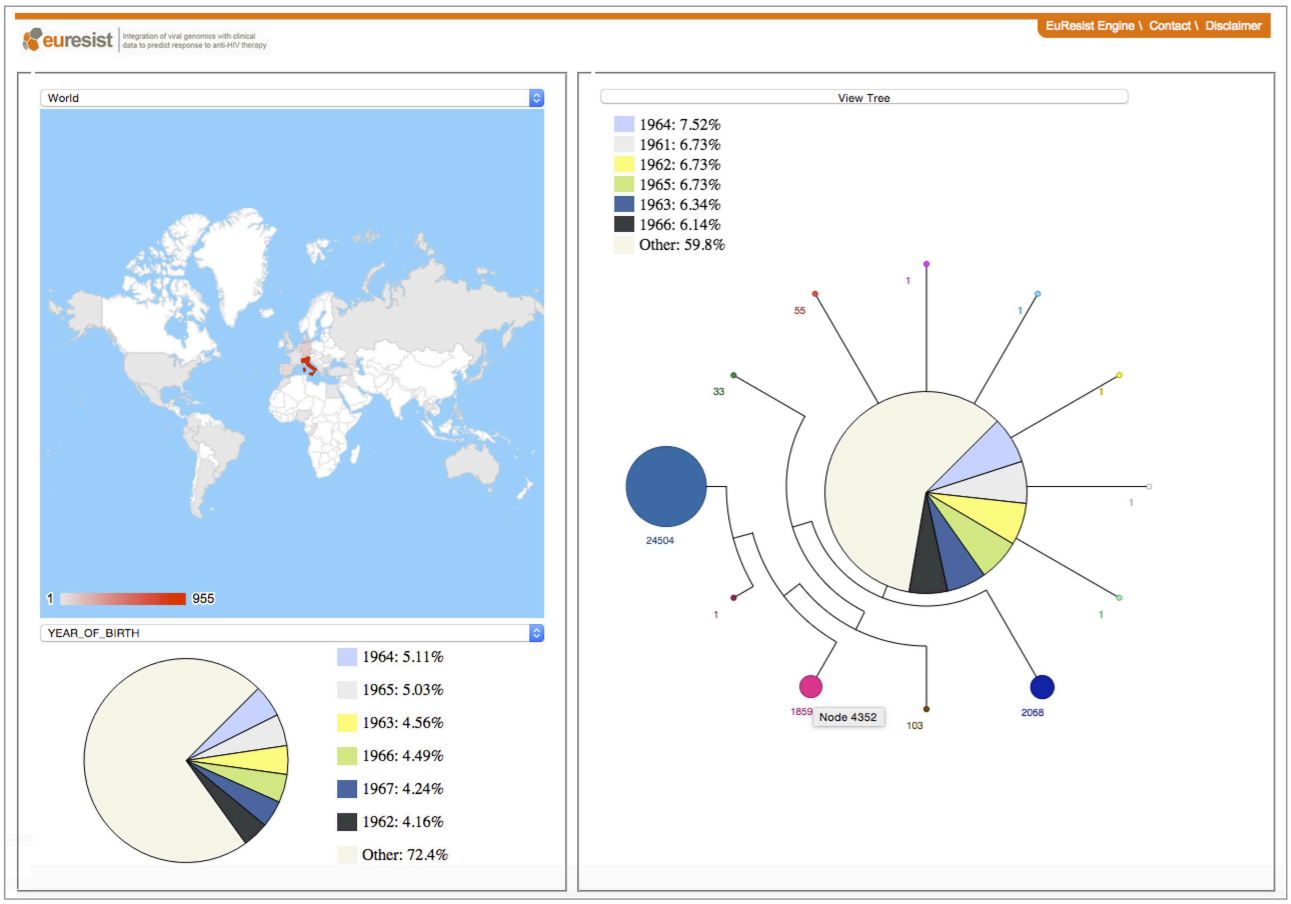


FIGURE 4 | The PhyloGeoTool offers a visual approach to explore large phylogenetic trees and to depict characteristics of strains and clades—including for example the geographic context and distribution of sampling dates—in an interactive way (17). A progressive zooming approach is used to ensure an efficient and interactive visual navigation of the entire phylogenetic tree.

the effective population sizes over time—such as the one shown in **Figure 8** for our RABV data set, which uses the Skygrid model (50) and its accompanying visualization in Tracer (49)—are commonly used to visualize the inferred past population size dynamics (50, 66, 67).

A variety of other summary statistics computed over the course of a phylogeny also benefit from visual representations, such as for the basic reproduction number and its rate of change as a function through time (68). Closely related are lineage-through-time plots (69) that allow exploring graphically the demographic signal in virus sequence data and revealing temporal changes of epidemic spread. Neher et al. (70) plotted cumulative antigenic changes over time by integrating viral phenotypic information into phylogenetic trees of influenza viruses, thereby providing additional insights into the rate of antigenic evolution compared to representations of neutralization titers that are traditionally transformed into a lower-dimensional space (71, 72). Another example relates to reconstructions of phylogeographic diffusion in discrete space, where patterns of migration links are typically projected into a cartographic context, but quantitive measures are additionally computed including the expected number of effective location state transitions (known as "Markov jumps"). Information on migrations in and out of a location state can be obtained by visualizations of the number of actual jumps between locations as well as the waiting times for each location, either as a total or proportionally over time (73–76).

The inference of transmission trees and networks ("who infected whom and when") using temporal, epidemiological and genetic information is an application of phylodynamics that has made substantial methodological progress in the last decade (77–79). Different from phylogenetic trees that represent evolutionary relationships between sampled viruses, transmission trees describe transmission events between hosts and require visualizations that are tailored to the analysis objectives (80–82). Consensus transmission trees, such as maximum parent credibility (MPC) trees (80) or Edmonds' trees (83), visually alert the user on putative infectors (indicated with arrows), corresponding infection times and potential super-spreaders. (80) use the Cytoscape framework (34) for drawing the transmission trees, and a similar adaptation of the original

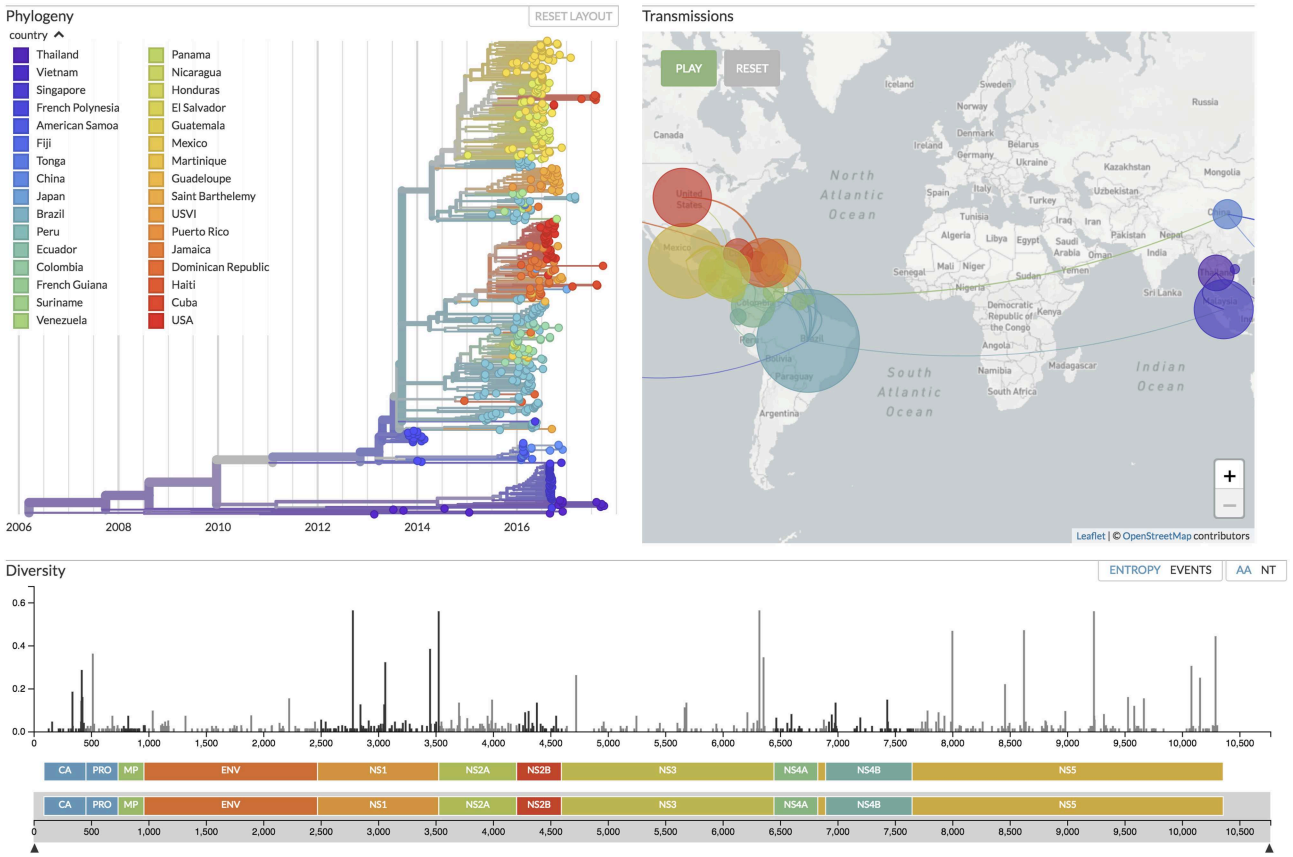


FIGURE 5 | Focusing on real-time tracking of several viruses, Nextstrain (19) provides up-to-date visualizations of phylogenetic and phylogeographic analyses, the latter with an animation over time similar to SpreaD3 (40). Shown here is the current situation for Zika virus evolution, based on analyses of 506 genomes sampled between February 2013 and September 2017 from 32 countries around the world (see the figure legend), corresponding to samples taken in 6 different regions of the world: China, Southeast Asia, Japan and Korea, Oceania, South America, and North America.

biological network-oriented framework has been done by PastML (33) (see above).

Finally, in order to compare two or more trees that are estimated from the same set of virus samples, but differ in the method used for tree construction or in the genomic region considered, tanglegrams provide insightful visualizations. The most popular use case is the comparison of two trees displayed leaf-by-leaf-wise with differences in clustering highlighted by lines connecting shared tips (84). Alternatively, tanglegrams allow mapping tree tip locations to mapped geographical locations using GenGis (58, 85). The Python toolkit Baltic (https://github.com/evogytis/baltic) provides functionalities to draw tangled chains, as shown in **Figure 9**, which are advanced sequential tanglegrams to compare a series of trees (43, 86). The use of phylogenetic networks, which are a generalization of phylogenetic trees, can also visualize phylogenetic incongruences, which could be due to reticulate evolutionary phenomena such as recombination (e.g., HIV-1) and hybridization (e.g., influenza virus) events (30, 32, 87). Tanglegrams and related visualization of sets of trees [e.g., DensiTree (39); see **Figure 10**] provide a qualitative and illustrative comparison of trees, buy this may prove to be less suited for the interpretation of extremely large trees or sets of trees. Recent quantitative approaches allow the exploration and visualization of the relationships between trees in a multi-dimensional space of tree similarities, based on different tree-to-tree distance metrics that identify a reduced tree space that maximally describe distinct patterns of observed evolution [Mesquite; (88), R package treespace; (89, 90)].

6. CONTEXT DEPENDENCE OF VISUALIZATION REQUIREMENTS

We have discussed a wide range of visualization packages for phylogenetic and phylodynamic analyses that allow improving our understanding of viral epidemiological and population dynamics. While these efforts may ultimately assist in informing public health or treatment decisions, visualization needs can differ according to the type of virus epidemics studied and questions that need to be answered. For example, the required level of visualization detail is high for (re-)emerging viral outbreaks when actionable insights should be obtained in a timely fashion in order to control further viral transmissions,

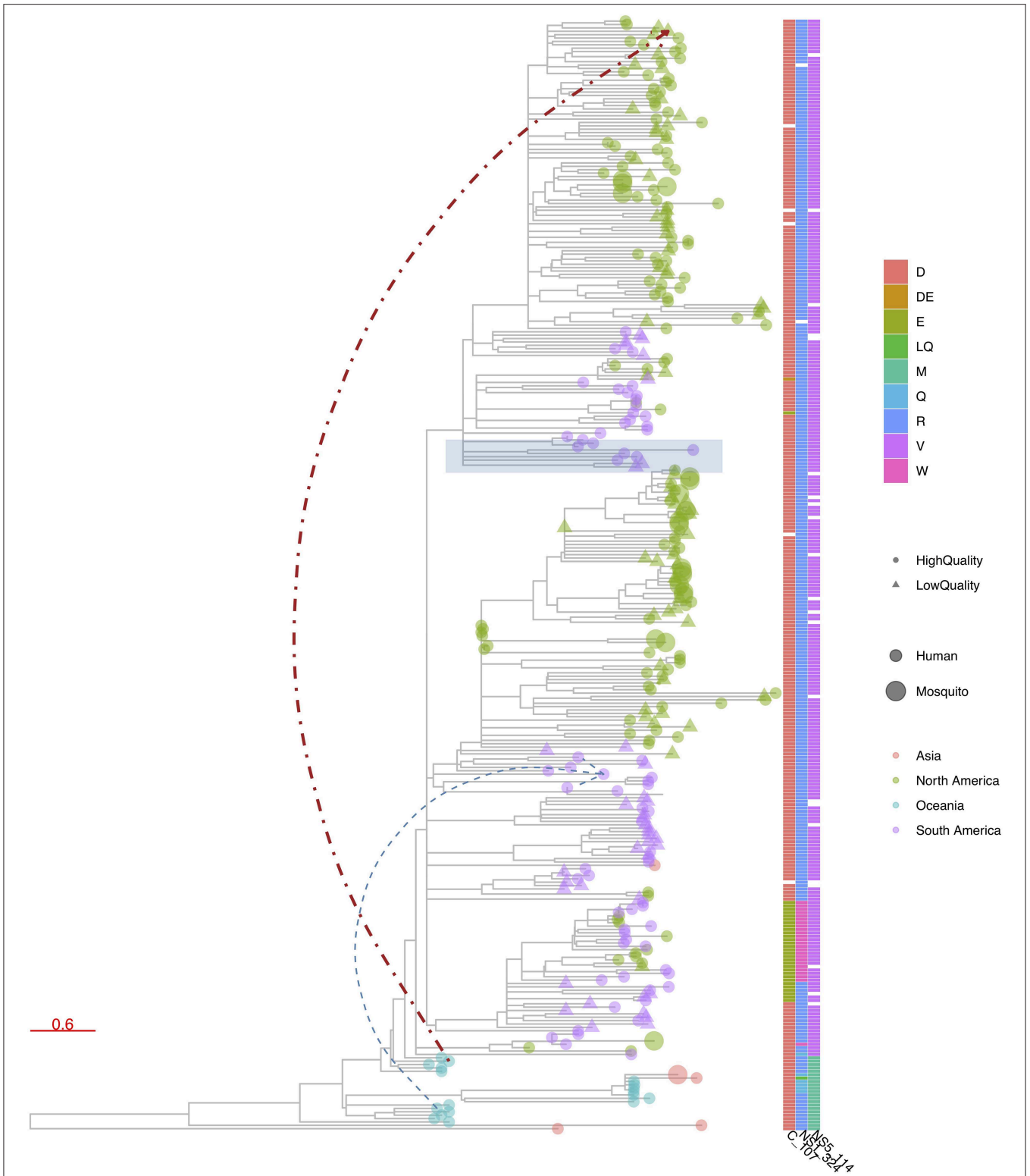


FIGURE 6 | R package ggtree (48) visualization of a phylogenetic tree constructed from publicly available Zika virus (ZIKV) genomes. ggtree allows similar advanced customized visualization of phylogenetic trees as e.g., iTOL, but by means of the traditional R scripting language. In this figure, tree leaves are colored according to continent of sampling, with a size corresponding to the host status and shape indicating the completeness of the CDS, using a cutoff of 99% of nucleotide positions being informative. A heatmap was added to denote the presence of amino acid mutations at three chosen genome positions. Finally, a particular clade was highlighted in blue based on a given internal node and two additional links between chosen taxa were added.

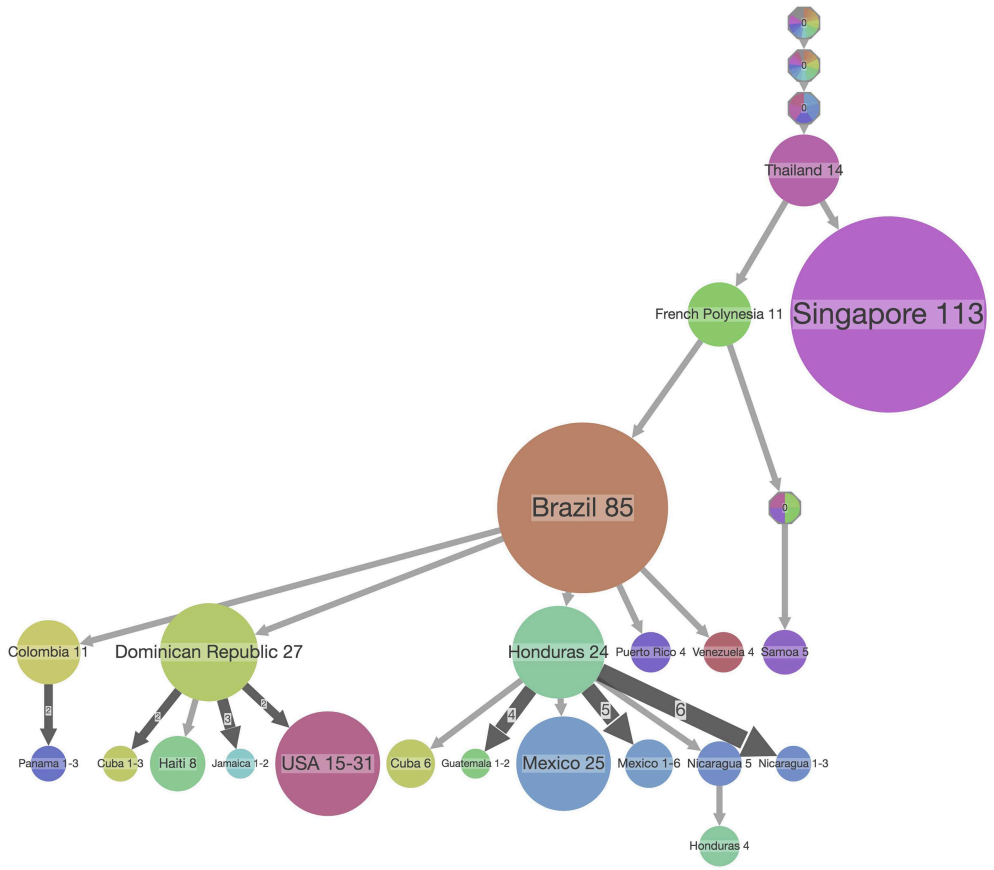


FIGURE 7 | PASTML summary visualization of the ancestral reconstruction of state locations of the ZIKV dataset used in **Figure 6**. The top-down visualization corresponds to an iterative clustering starting from the root of the tree at the top, with the size of the dot corresponding to the number of taxa in a clade which share the same ancestral state which is indicated on top of the dot. In this type of visualization, a compressed representation of the ancestral scenarios is visualized that highlights the main facts and hides minor details by performing both a vertical and horizontal merge [but see (33)]. The branch width corresponds to the number of times its subtree is found in the initial tree, and the circle size at a tip is proportional to the size of the compressed (or merged) cluster.

with real-time tracking of viral spread and the identification of sources, transmission patterns and contributing factors being key priorities (91). As a result, software packages that aim to address these questions are typically developed with an explicit emphasis on speed through the use of heuristics, and stress the importance of connectivity and interactivity to quickly respond to the availability of new data in order to develop novel insights into an ongoing epidemic. One-stop and fully-integrated analysis platforms such as MicroReact (46) and Nextstrain (19) adhere to these needs by providing the necessary visualizations of virus epidemiology and evolution across time and space, and by implementing support for collaborative analyses and sharing of genomic data and analysis outputs. A strategy of interest in these settings is the ability for phylogenetic placement of novel sequence data (92, 93), for example when updated outbreak information suggests specific cases should be investigated but the reconstruction of a new phylogeny is not desirable, as this may prove too time consuming. To avoid such *de novo* re-analyses of data sets, software tools such as iTOL (18) and PhyloGeoTool (17) offer functionalities to visualize placements of sequence data onto an existing phylogeny. A key future challenge of these approaches is to assess and visualize the associated phylogenetic placement uncertainty, or if this information would be unavailable to at least indicate the various stages in which novel sequences were added onto the (backbone) phylogeny. While methodological developments are rigorous in their accuracy assessment—for example through simulation studies—and may even provide visual options for representing the placement uncertainty [see e.g., (92)], visualization packages themselves do not offer an automated way of assessing or conveying this information and as such may project overconfidence of the power of the phylogenetic placement method used. Additionally, other flexible visualization options based on real-time outbreak monitoring can be of great interest such as highlighting locations from which cases have been reported but for which genomic data are still lacking, to clarify the potential impact of these missing data on the currently available inference results.

Investigations of more established epidemics usually involve much larger sample sizes, are more retrospective-oriented in

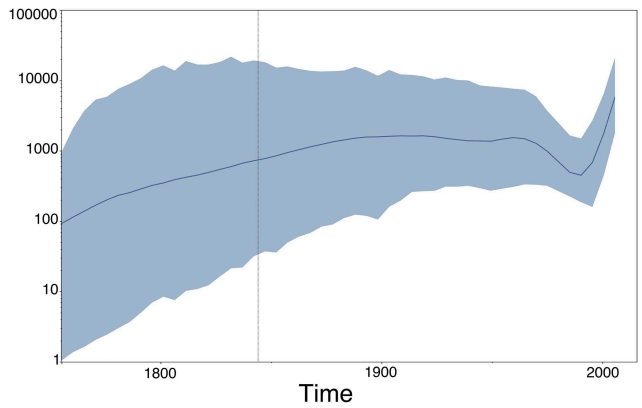


FIGURE 8 | Population (size) dynamics over time visualization of our RABV analysis (previous section) using Tracer (49). This type of output does not directly depend upon the estimated MCC tree, but rather on the estimated (log) population sizes of the Skygrid model (50), which are provided in a separate output file by BEAST (51).

design and incorporate more heterogeneous information, and therefore benefit from more extended visualization frameworks. For most of these globally prevalent pathogens, clinical and phenotypic information is often available and questions relate to the population- or patient-level dynamics of viral adaptation and the identification of transmission clusters. For example, the selection of the virus strain composition of the seasonal influenza vaccine is informed by analyses and visualizations of circulating strains and their antigenic properties using the nextflu framework (47, 91). Other diverse examples include investigations of the impact of country-specific public health interventions on transmission dynamics (94, 95), the identification of distinctive socio-demographic, clinical and epidemiological features associated with regional and global epidemics (96–99) and large-scale modeling of epidemiological links among geographical locations (100–102). In these settings, relevant software packages should consider the scalability of large phylogenies and allow user-friendly exploration of heterogeneous and customized annotations. Overall, it is anticipated that future work on visualization tools, accompanying analysis and visualization software developments as described above, will result in a merging of these two epidemic perspectives, with the development of context-independent visualization software tools that can handle both scenarios equally well.

7. CONCLUSIONS

Viral pathogens, in particular RNA viruses, have been responsible for epidemics and recurrent outbreaks associated with high morbidity and mortality in the human population, for a duration that can span from hundreds of years [e.g., HCV (103) and DENV (104)] to decades [e.g., HIV-1 (3)]. RNA viruses are known for their potential to quickly adapt to host and treatment selective pressure, but their rapid accumulation of genomic changes also provides opportunities to study their population and transmission dynamics in high resolution. Consequently, the fields of phylogenomics and phylodynamics play a pivotal role in studies on epidemiology and transmission of viral infectious diseases, and have advanced our understanding of the dynamical processes that govern virus dispersal and evolution at both population and host levels. Compared to the tremendous achievements in the performance of evolutionary and statistical inference models and hypothesis testing frameworks, software packages and resources aimed at visualizing the output of these studies have experienced difficulties to handle the increasing complexity and sizes of the analyses, for example to display levels of uncertainty and to integrate associated demographic and clinical information. Accurate and meaningful visual representation and communication are however essential tools for the interpretation and translation of outcomes into actionable insights for the design of optimal prevention, control and treatment interventions. With a plethora of applications for phylodynamics having been introduced in the last decades, in particular tailored toward reconstructions of spatiotemporal histories—which start to become useful in public health surveillance—visualization has substantially grown as an elementary discipline for investigations of infectious disease epidemiology and evolution. An extensive array of software and tools for the manipulation, editing and annotation of output visualizations in the field of pathogen phylodynamics is available to date, characterized by varying technical specifications and functionalities that respond to heterogeneous needs from the research and public health communities.

The increasing recognition for visualization tools in support of viral outbreak surveillance and control has stimulated the advent of more complex and fully integrated frameworks and platforms, all the while focusing on user experience and ease of customisation. We anticipate that future visualization developments will take further leaps in this ongoing trend by tackling remaining challenges to display increasing amounts of dense information in a human-readable manner and introducing concepts from new disciplines such as visual analytics. In particular, high expectations are stemming from the ensemble of visualization methods that allow users to work at, and move between, focused and contextual views of a data set (105). Large scientific data sets with a temporal aspect have been the subject of multi-level focus+context approaches for their interactive visualization (106), which minimize the seam between data views by displaying the focus on a specific situation or part of the data within its context. These approaches are part of an extensive series of interface mechanisms used to separate and blend views of the data, such as overview+detail, which uses a spatial separation between focused and contextual views, and zooming, which uses a temporal separation between these views (105). Phylogenetic trees can be interactively visualized as three-dimensional stacked representations (107). The field of phylogenomics and phylodynamics visualizations will increasingly implement and adapt technologies from other disciplines, as already illustrated by tools and studies using the network-oriented Cytoscape package (33, 34, 78), or through the use of virtual reality technologies including customizable mapping frameworks and high-performance geospatial analytical toolboxes. As such, concomitant to the ongoing developments

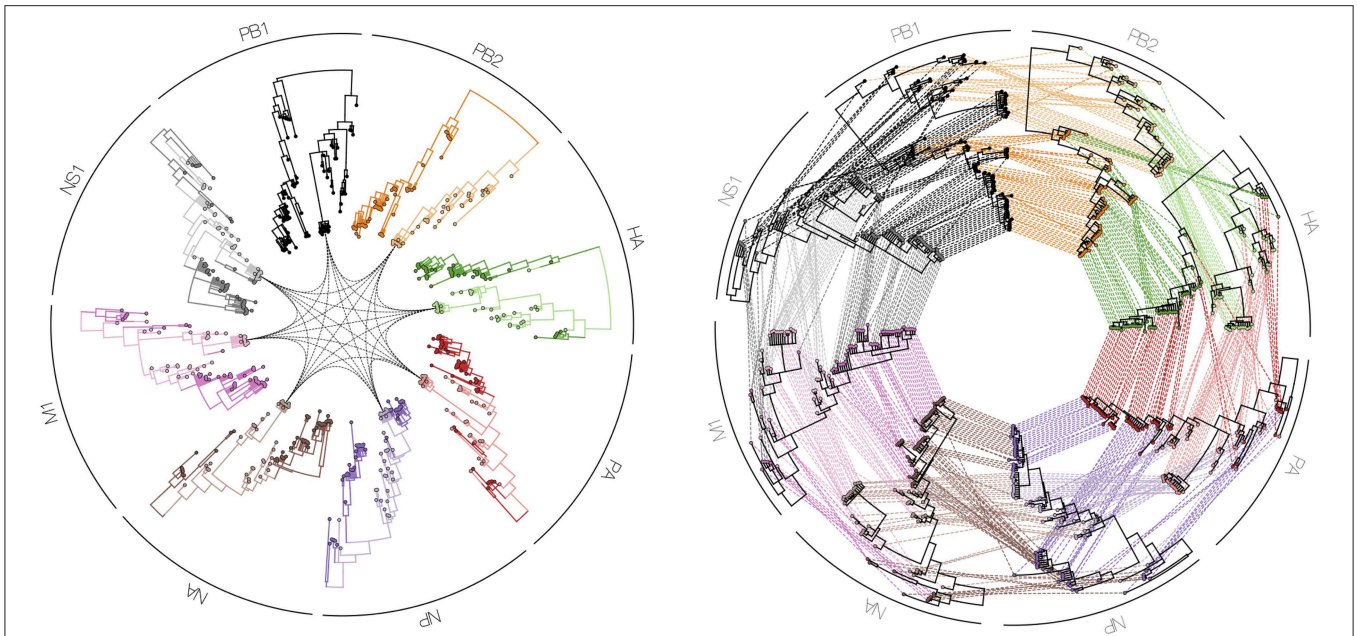


FIGURE 9 | Tanglegrams are typically shown in a side-by-side manner, in order to easily and visually identify differences in clustering between two or more phylogenetic trees, for example when inferred from different influenza proteins (PB1, PB2, PA, HA, NP, NA, M1, and NS1). Such a series of trees can also be visualized in a circle facing inwards with a particular isolate highlighted in all plotted phylogenies **(left figure)**, or with all isolates interconnected between all proteins **(right figure)**.

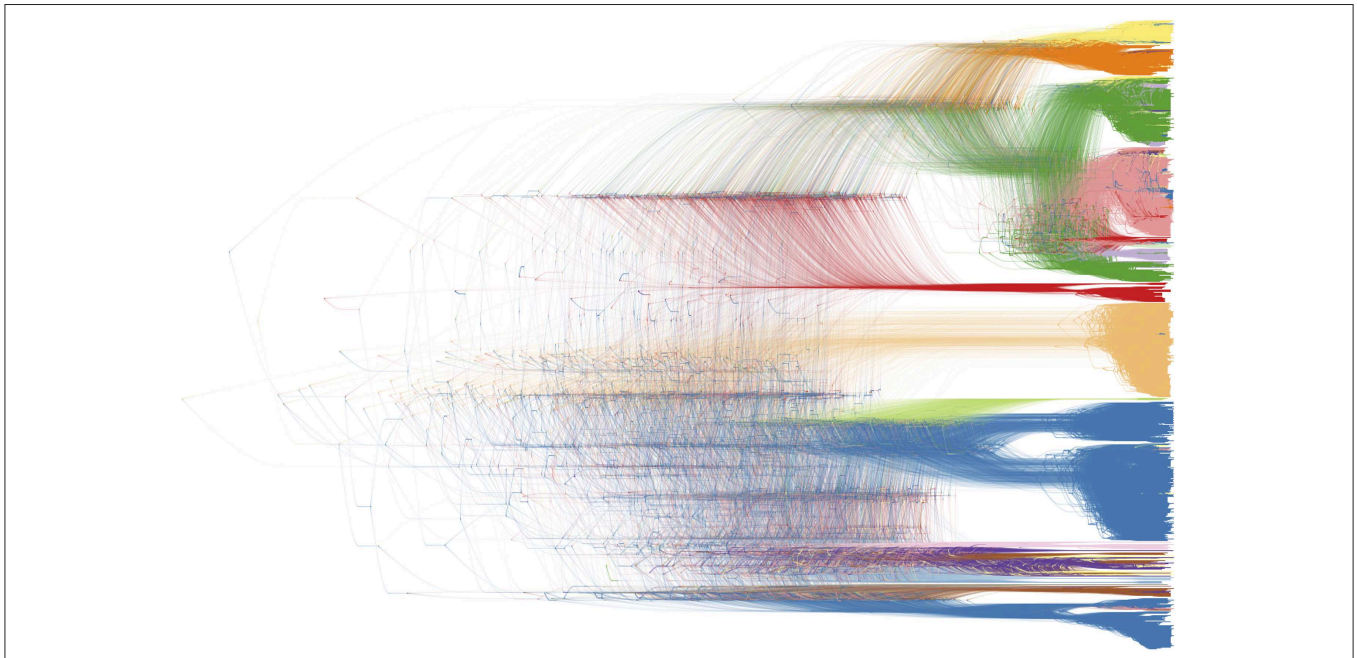


FIGURE 10 | Bayesian phylogenetic inference software packages generate a large number of posterior trees, potentially annotated with inferred ancestral traits. This collection of trees is often summarized using a consensus tree, allowing to plot a single tree with posterior support values on the internal nodes. DensiTree enables drawing all posterior trees in the collection; areas where a lot of the trees agree in topology and branch lengths show up as highly colored areas, while areas with little agreement show up as webs (39). We refer to **Figure 2** for the color legend of the host species, as the legend drawn by DensiTree was not very readable and could not be edited (in terms of its textual information).

in sample collection and sequencing, the design of more complex analytical inference models and powerful hardware infrastructure will be complemented by a new era in visualization applications that will collaboratively foster visualizations that track virus epidemics and outbreaks in real-time and with high resolution.

SEARCH STRATEGY

An initial but already comprehensive list of publications was compiled from backward and forward citation searches of the various visualization software packages the authors have (co-)developed, as well as those packages that the authors have used throughout their academic career. Complementing this already extensive list, we searched PubMed and Google Scholar, which keeps track of arXiv and bioRxiv submissions and hence decreased the risk of missing potential publications. Additional supplementary searches have been performed by backward and forward citation chasing of all of the included references throughout the writing process of writing the manuscript for the initial submission on April 7th 2019. No date restrictions were applied, but only visualization packages and publications written in English were considered.

AUTHOR CONTRIBUTIONS

KT wrote the manuscript. PL helped with the interpretation and writing. A-MV gave the idea, helped with the interpretation and writing. GB wrote the manuscript and prepared the visualizations.

FUNDING

GB acknowledges support from the Interne Fondsen KU Leuven/Internal Funds KU Leuven under grant agreement C14/18/094. The research leading to these results has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programmes (grant agreement no. 725422-ReservoirDOCS). This work was partially supported by the European Union's Horizon 2020 Research and Innovation Programme under ZIKAlliance Grant Agreement no. 734548 and under VIROGENESIS Grant Agreement no. 634650. The Artic Network receives funding from the Wellcome Trust through project 206298/Z/17/Z. PL acknowledges support by the Research Foundation – Flanders (Fonds voor Wetenschappelijk Onderzoek – Vlaanderen, G066215N, G0D5117N, and G0B9317N).

ACKNOWLEDGMENTS

We are grateful to Gytis Dudas for providing a figure from his Baltic visualization package (https://github.com/evogytis/baltic). We thank Simon Dellicour for fruitful discussions.

REFERENCES

1. Rife BD, Mavian C, Chen X, Ciccozzi M, Salemi M, Min J, et al. Phylodynamic applications in 21st century global infectious disease research. *Glob Health Res Policy*. (2017) 2:13. doi: 10.1186/s41256-017-0034-y
2. Dellicour S, Baele G, Dudas G, Faria NR, Pybus OG, Suchard MA, et al. Phylodynamic assessment of intervention strategies for the West African Ebola virus outbreak. *Nat Comms*. (2018) 9:2222. doi: 10.1038/s41467-018-03763-2
3. Faria NR, Rambaut A, Suchard MA, Baele G, Bedford T, Ward MJ, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science*. (2014) 346:56–61. doi: 10.1126/science.1256739
4. Faria NR, Kraemer MUG, Hill SC, Goes de Jesus J, Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science*. (2018) 361:894–9. doi: 10.1126/science.aat7115
5. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature*. (2017) 544:309–15. doi: 10.1038/nature22040
6. Russell CA, Kasson PM, Donis RO, Riley S, Dunbar J, Rambaut A, et al. Science Forum: improving pandemic influenza risk assessment. *eLife*. (2014) 3:e03883. doi: 10.7554/eLife.03883
7. German D, Grabowski MK, Beyrer C. Enhanced use of phylogenetic data to inform public health approaches to HIV among men who have sex with men. *Sex Health*. (2016) 14:89–96. doi: 10.1071/SH16056
8. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*. (2004) 303:327–32. doi: 10.1126/science.1090727
9. Baele G, Suchard MA, Rambaut A, Lemey P. Emerging concepts of data integration in pathogen phylodynamics. *Syst Biol*. (2016) 66:e47–e65. doi: 10.1093/sysbio/syw054
10. Baele G, Dellicour S, Suchard MA, Lemey P, Vrancken B. Recent advances in computational phylodynamics. *Curr Opin Virol*. (2018) 31:24–32. doi: 10.1016/j.coviro.2018.08.009
11. Quick J, Loman NJ, Durrafour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature*. (2016) 530:228–32. doi: 10.1038/nature16996
12. Baele G, Lemey P, Rambaut A, Suchard MA. Adaptive MCMC in Bayesian phylogenetics: an application to analyzing partitioned data in BEAST. *Bioinformatics*. (2017) 33:1798–805. doi: 10.1093/bioinformatics/btx088
13. Ayres DL, Cummings MP, Baele G, Darling AE, Lewis PO, Swofford DL, et al. BEAGLE 3: improved performance, scaling and usability for a high-performance computing library for statistical phylogenetics. *Syst Biol*. (2019). doi: 10.1093/sysbio/syz020
14. Darwin C. *On the Origin of Species by Means of Natural Selection*. London: Murray (1859).
15. Minin VM, Suchard MA. Fast, accurate and simulation-free stochastic mapping. *Philos Trans R Soc Lond B Biol Sci*. (2008) 363:3985–95. doi: 10.1098/rstb.2008.0176
16. Borner K, Bueckle A, Ginda M. Data visualization literacy: definitions, conceptual frameworks, exercises, and assessments. *Proc Natl Acad Sci USA*. (2019) 116:1857–64. doi: 10.1073/pnas.1807180116
17. Libin P, Vanden Eynden E, Incardona F, Nowe A, Bezenchek A, Sonnerborg A, et al. PhyloGeoTool: interactively exploring large phylogenies in an epidemiological context. *Bioinformatics*. (2017) 33:3993–5. doi: 10.1093/bioinformatics/btx535
18. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics*. (2007) 23:127–8. doi: 10.1093/bioinformatics/btl529
19. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*. (2018) 34:4121–3. doi: 10.1093/bioinformatics/bty407
20. Wilke CO. *Fundamentals of Data Visualization*. O'Reilly Media, Inc. (2019).
21. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science*. (2019) 363:74–7. doi: 10.1126/science.aau9343
22. Rogers JS, Swofford DL. A fast method for approximating maximum likelihoods of phylogenetic trees from nucleotide

sequences. *Syst Biol.* (1998) 47:77–89. doi: 10.1080/10635159 8261049
23. Swofford D. *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sunderland, MA: Sinauer Associates.
24. Felsenstein J. PHYLIP - phylogeny inference package (Version 3.2). *Cladistics.* (1993) 5:164–6.
25. Philippe H. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res.* (1993) 21:5264–72. doi: 10.1093/nar/21.22.5264
26. Page RDM. Tree view: an application to display phylogenetic trees on personal computers. *Bioinformatics.* (1996) 12:357–8. doi: 10.1093/bioinformatics/12. 4.357
27. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* (2012) 61:539–42. doi: 10.1093/sysbio/sys029
28. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol Biol Evol.* (2016) 33:1870–4. doi: 10.1093/molbev/msw054
29. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. Dendroscope: an interactive viewer for large phylogenetic trees. *BMC Bioinformatics.* (2007) 8:460. doi: 10.1186/1471-2105-8-460
30. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Syst Biol.* (2012) 61:1061–7. doi: 10.1093/sysbio/sys062
31. Robinson O, Dylus D, Dessimoz C. Phylo.io: interactive viewing and comparison of large phylogenetic trees on the web. *Mol Biol Evol.* (2016) 33:2163–6. doi: 10.1093/molbev/msw080
32. Vaughan TG. IcyTree: rapid browser-based visualization for phylogenetic trees and networks. *Bioinformatics.* (2017) 33:2392–4. doi: 10.1093/bioinformatics/btx155
33. Ishikawa S, Zhukova A, Iwasaki W, Gascuel O. A fast likelihood method to reconstruct and visualize ancestral scenarios. *bioRxiv.* (2018). Available online at: https://www.biorxiv.org/content/early/2018/07/29/379529
34. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* (2003) 13:2498–04. doi: 10.1101/gr.1239303
35. Chevenet F, Castel G, Jousselin E, Gascuel O. PastView: a user-friendly interface to explore evolutionary scenarios. *bioRxiv.* (2019). Available online at: https://www.biorxiv.org/content/early/2019/05/27/651661
36. Zhou Z, Alikhan NF, Sergeant MJ, Luhmann N, Vaz C, Francisco AP, et al. GrapeTree: visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res.* (2018) 28:1395–404. doi: 10.1101/gr.232397.117
37. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution.* (1985) 39:783–91. doi: 10.1111/j.1558-5646.1985. tb00420.x
38. Rambaut A. FigTree, version 1.4.3. (2009). Available online at: http://tree.bio.ed.ac.uk/software/figtree
39. Bouckaert RR. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics.* (2010) 26:1372–3. doi: 10.1093/bioinformatics/btq110
40. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. *Mol Biol Evol.* (2016) 33:2167–9. doi: 10.1093/molbev/msw082
41. Kreft L, Botzki A, Coppens F, Vandepoele K, Van Bel M. PhyD3: a phylogenetic tree viewer with extended phyloXML support for functional genomics data visualization. *Bioinformatics.* (2017) 33:2946–7. doi: 10.1093/bioinformatics/btx324
42. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol.* (2016) 33:1635–8. doi: 10.1093/molbev/msw046
43. Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. *Elife.* (2018) 1:7. doi: 10.7554/eLife.31257
44. Dellicour S, Rose R, Faria NR, Lemey P, Pybus OG. SERAPHIM: studying environmental rasters and phylogenetically informed movements. *Bioinformatics.* (2016) 32:3204–6. doi: 10.1093/bioinformatics/btw384
45. Dellicour S, Rose R, Pybus OG. Explaining the geographic spread of emerging epidemics: a framework for comparing viral phylogenies and environmental landscape data. *BMC Bioinformatics.* (2016) 17:82. doi: 10.1186/s12859-016-0924-x
46. Argimon S, Abudahab K, Goater RJ, Fedosejev A, Bhai J, Glasner C, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom.* (2016) 2:e000093. doi: 10.1099/mgen.0.000093
47. Neher RA, Bedford T. nextflu: real-time tracking of seasonal influenza virus evolution in humans. *Bioinformatics.* (2015) 31:3546–8. doi: 10.1093/bioinformatics/btv381
48. Yu G, Lam TT, Zhu H, Guan Y. Two methods for mapping and visualizing associated data on phylogeny using Ggtree. *Mol Biol Evol.* (2018) 35:3041–3. doi: 10.1093/molbev/msy194
49. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol.* (2018) 67:901–4. doi: 10.1093/sysbio/syy032
50. Gill MS, Lemey P, Faria NR, Rambaut A, Shapiro B, Suchard MA. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Mol Biol Evol.* (2013) 30:713–24. doi: 10.1093/molbev/mss265
51. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* (2018) 4:vey016. doi: 10.1093/ve/vey016
52. Streicker D, Turmelle A, Vonhof M, Kuzmin I, McCracken GF, Rupprecht C. Host phylogeny constrains cross-species emergence and establishment of rabies virus in bats. *Science.* (2010) 329:676–9. doi: 10.1126/science. 1188836
53. Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. *PLoS Comput Biol.* (2009) 5:e1000520. doi: 10.1371/journal.pcbi.1000520
54. Cardona G, Rosselló F, Valiente G. Extended Newick: it is time for a standard representation of phylogenetic networks. *BMC Bioinformatics.* (2008) 9:532. doi: 10.1186/1471-2105-9-532
55. Han MV, Zmasek CM. phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics.* (2009) 10:356. doi: 10.1186/1471-2105-10-356
56. Vos RA, Balhoff JP, Caravas JA, Holder MT, Lapp H, Maddison WP, et al. NeXML: rich, extensible, and verifiable representation of comparative data and metadata. *Syst Biol.* (2012) 61:675–89. doi: 10.1093/sysbio/sys025
57. Lemey P, Rambaut A, Welch JJ, Suchard MA. Phylogeography takes a relaxed random walk in continuous space and time. *Mol Biol Evol.* (2010) 27:1877–85. doi: 10.1093/molbev/msq067
58. Parks DH, Porter M, Churcher S, Wang S, Blouin C, Whalley J, et al. GenGIS: a geospatial information system for genomic data. *Genome Res.* (2009) 19:1896–904. doi: 10.1101/gr.095612.109
59. Maddison GR, Maddison WP. Cartographer, a Mesquite package for plotting geographic data. (2017). Available online at: http://mesquiteproject.org/packages/cartographer
60. Bielejec F, Rambaut A, Suchard MA, Lemey P. SPREAD: spatial phylogenetic reconstruction of evolutionary dynamics. *Bioinformatics.* (2011) 27:2910–2. doi: 10.1093/bioinformatics/btr481
61. Shank SD, Weaver S, Kosakovsky Pond SL. phylotree.js - a JavaScript library for application development and interactive data visualization in phylogenetics. *BMC Bioinformatics.* (2018) 19:276. doi: 10.1186/s12859-018-2283-2
62. Zuccon A, Zuccon D. MrEnt: an editor for publication-quality phylogenetic tree illustrations. *Mol Ecol Resour.* (2014) 14:1090–4. doi: 10.1111/1755-0998.12253
63. Wilkinson L. *The Grammar of Graphics (Statistics and Computing)*. Berlin; Heidelberg: Springer-Verlag (2005).
64. Franz M, Lopes CT, Huck G, Dong Y, Sumer O, Bader GD. Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics.* (2015) 32:309–11. doi: 10.1093/bioinformatics/btv557
65. Kingman JF. Origins of the coalescent. 1974-1982. *Genetics.* (2000) 156:1461–3.

66. Pybus OG, Rambaut A. GENIE: estimating demographic history from molecular phylogenies. *Bioinformatics.* (2002) 18:1404–5. doi: 10.1093/bioinformatics/18.10.1404
67. Minin VM, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol.* (2008) 25:1459–71. doi: 10.1093/molbev/msn090
68. Stadler T, Kuhnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc Natl Acad Sci USA.* (2013) 110:228–33. doi: 10.1073/pnas.1207965110
69. Nee S, Holmes EC, Rambaut A, Harvey PH. Inferring population history from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci.* (1995) 349:25–31. doi: 10.1098/rstb.1995.0087
70. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. *Proc Natl Acad Sci USA.* (2016) 113:E1701–1709. doi: 10.1073/pnas.1525578113
71. Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. *Science.* (2004) 305:371–6. doi: 10.1126/science.1097211
72. Katzelnick LC, Fonville JM, Gromowski GD, Bustos Arriaga J, Green A, James SL, et al. Dengue viruses cluster antigenically but not as discrete serotypes. *Science.* (2015) 349:1338–43. doi: 10.1126/science.aac5017
73. Lemey P, Rambaut A, Bedford T, Faria N, Bielejec F, Baele G, et al. Unifying viral genetics and human transportation data to predict the global transmission dynamics of human influenza H3N2. *PLoS Pathog.* (2014) 10:e1003932. doi: 10.1371/journal.ppat.1003932
74. Minin VN, Suchard MA. Counting labeled transitions in continuous-time Markov models of evolution. *J Math Biol.* (2008) 56:391–412. doi: 10.1007/s00285-007-0120-8
75. Bedford T, Cobey S, Beerli P, Pascual M. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathog.* (2010) 6:e1000918. doi: 10.1371/journal.ppat.1000918
76. Su YC, Bahl J, Joseph U, Butt KM, Peck HA, Koay ES, et al. Phylodynamics of H1N1/2009 influenza reveals the transition from host adaptation to immune-driven selection. *Nat Commun.* (2015) 6:7952. doi: 10.1038/ncomms8952
77. Jombart T, Cori A, Didelot X, Cauchemez S, Fraser C, Ferguson N. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput Biol.* (2014) 10:e1003457. doi: 10.1371/journal.pcbi.1003457
78. Hall MD, Woolhouse ME, Rambaut A. Using genomics data to reconstruct transmission trees during disease outbreaks. *Rev Off Int Epizoot.* (2016) 35:287–96. doi: 10.20506/rst.35.1.2433
79. Ratmann O, Grabowski MK, Hall M, Golubchik T, Wymant C, Abeler-Dorner L, et al. Inferring HIV-1 transmission networks and sources of epidemic spread in Africa with deep-sequence phylogenetic analysis. *Nat Commun.* (2019) 10:1411. doi: 10.1038/s41467-019-09139-4
80. Hall M, Woolhouse M, Rambaut A. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS Comput Biol.* (2015) 11:e1004613. doi: 10.1371/journal.pcbi.1004613
81. Ypma RJ, van Ballegooijen WM, Wallinga J. Relating phylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics.* (2013) 195:1055–62. doi: 10.1534/genetics.113.154856
82. Mollentze N, Nel LH, Townsend S, le Roux K, Hampson K, Haydon DT, et al. A Bayesian approach for inferring the dynamics of partially observed endemic infectious diseases from space-time-genetic data. *Proc Biol Sci.* (2014) 281:20133251. doi: 10.1098/rspb.2013.3251
83. Klinkenberg D, Backer JA, Didelot X, Colijn C, Wallinga J. Simultaneous inference of phylogenetic and transmission trees in infectious disease outbreaks. *PLoS Comput Biol.* (2017) 13:e1005495. doi: 10.1371/journal.pcbi.1005495
84. Dudas G, Bedford T, Lycett S, Rambaut A. Reassortment between influenza B lineages and the emergence of a coadapted PB1-PB2-HA gene complex. *Mol Biol Evol.* (2015) 32:162–72. doi: 10.1093/molbev/msu287
85. Trovao NS, Baele G, Vrancken B, Bielejec F, Suchard MA, Fargette D, et al. Host ecology determines the dispersal patterns of a plant virus. *Virus Evol.* (2015) 1:vev016. doi: 10.1093/ve/vev016
86. Dudas G, Rambaut A. MERS-CoV recombination: implications about the reservoir and potential for adaptation. *Virus Evol.* (2016) 2:vev023. doi: 10.1093/ve/vev023
87. Scornavacca C, Zickmann F, Huson DH. Tanglegrams for rooted phylogenetic trees and networks. *Bioinformatics.* (2011) 27:i248–256. doi: 10.1093/bioinformatics/btr210
88. Maddison GR, Maddison WP. Mesquite: A Modular System for Evolutionary Analysis. Version 3.51. (2018). Available online at: http://www.mesquiteproject.org
89. Kendall M, Colijn C. Mapping phylogenetic trees to reveal distinct patterns of evolution. *Mol Biol Evol.* (2016) 33:2735–43. doi: 10.1093/molbev/msw124
90. Jombart T, Kendall M, Almagro-Garcia J, Colijn C. treespace: statistical exploration of landscapes of phylogenetic trees. *Mol Ecol Resour.* (2017) 17:1385–92. doi: 10.1111/1755-0998.12676
91. Neher RA, Bedford T. Real-time analysis and visualization of pathogen sequence data. *J Clin Microbiol.* (2018) 56:e00480–18. doi: 10.1128/JCM.00480-18
92. Matsen FA, Kodner RB, Armbrust EV. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics.* (2010) 11:538. doi: 10.1186/1471-2105-11-538
93. Matsen FA, Hoffman NG, Gallagher A, Stamatakis A. A format for phylogenetic placements. *PLoS ONE.* (2012) 7:e31009. doi: 10.1371/journal.pone.0031009
94. Vasylyeva TI, du Plessis L, Pineda-Pena AC, Kuhnert D, Lemey P, Vandamme AM, et al. Tracing the impact of public health interventions on HIV-1 transmission in portugal using molecular epidemiology. *J Infect Dis.* (2019) 220:233–43. doi: 10.1093/infdis/jiz085
95. Wilkinson E, Junqueira DM, Lessells R, Engelbrecht S, van Zyl G, de Oliveira T, et al. The effect of interventions on the transmission and spread of HIV in South Africa: a phylodynamic analysis. *Sci Rep.* (2019) 9:2640. doi: 10.1038/s41598-018-37749-3
96. Jones BR, Howe AYM, Harrigan PR, Joy JB. The global origins of resistance-associated variants in the non-structural proteins 5A and 5B of the hepatitis C virus. *Virus Evol.* (2018) 4:vex041. doi: 10.1093/ve/vex041
97. Magiorkinis G, Magiorkinis E, Paraskevis D, Ho SY, Shapiro B, Pybus OG, et al. The global spread of hepatitis C virus 1a and 1b: a phylodynamic and phylogeographic analysis. *PLoS Med.* (2009) 6:e1000198. doi: 10.1371/journal.pmed.1000198
98. Pineda-Pena AC, Theys K, Stylianou DC, Demetriades I, Abecasis AB, Kostrikis LG, et al. HIV-1 infection in cyprus, the Eastern mediterranean European frontier: a densely sampled transmission dynamics analysis from 1986 to 2012. *Sci Rep.* (2018) 8:1702. doi: 10.1038/s41598-017-19080-5
99. Theys K, Van Laethem K, Gomes P, Baele G, Pineda-Pena AC, Vandamme AM, et al. Sub-epidemics explain localized high prevalence of reduced susceptibility to rilpivirine in treatment-naive HIV-1-infected patients: subtype and geographic compartmentalization of baseline resistance mutations. *AIDS Res Hum Retroviruses.* (2016) 32:427–33. doi: 10.1089/aid.2015.0095
100. Nelson MI, Simonsen L, Viboud C, Miller MA, Holmes EC. Phylogenetic analysis reveals the global migration of seasonal influenza A viruses. *PLoS Pathog.* (2007) 3:1220–28. doi: 10.1371/journal.ppat.0030131
101. Fourment M, Darling AE, Holmes EC. The impact of migratory flyways on the spread of avian influenza virus in North America. *BMC Evol Biol.* (2017) 17:118. doi: 10.1186/s12862-017-0965-4
102. Kostaki EG, Flampouris A, Karamitros T, Chueca N, Alvarez M, Casas P, et al. Spatiotemporal characteristics of the largest HIV-1 CRF02AG outbreak in Spain: evidence for onward transmissions. *Front Microbiol.* (2019) 10:370. doi: 10.3389/fmicb.2019.00370
103. Forni D, Cagliani R, Pontremoli C, Pozzoli U, Vertemara J, De Gioia L, et al. Evolutionary analysis provides insight into the origin and adaptation of HCV. *Front Microbiol.* (2018) 9:854. doi: 10.3389/fmicb.2018.00854
104. Holmes EC, Twiddy SS. The origin, emergence and evolutionary genetics of dengue virus. *Infect Genet Evol.* (2003) 3:19–28. doi: 10.1016/S1567-1348(03)00004-2

105. Cockburn A, Karlson A, Bederson BB. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput Surv.* (2009) 41:2:1–2:31. doi: 10.1145/1456650.1456652
106. Muigg P, Kehrer J, Oeltze S, Piringer H, Doleisch H, Preim B, et al. A four-level focus+context approach to interactive visual analysis of temporal features in large scientific data. *Comput Graph Forum.* (2008) 27:775–82. doi: 10.1111/j.1467-8659.2008.01207.x
107. Page RD. Space, time, form: viewing the Tree of Life. *Trends Ecol Evol (Amst).* (2012) 27:113–20. doi: 10.1016/j.tree.2011.12.002

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

REVIEW
published: 21 August 2019
doi: 10.3389/fpubh.2019.00235

Whole Genome Sequencing for Surveillance of Diphtheria in Low Incidence Settings

Helena M. B. Seth-Smith[1,2,3] *and Adrian Egli*[1,2]*

[1] *Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland,* [2] *Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland,* [3] *SIB Swiss Institute of Bioinformatics, Basel, Switzerland*

Corynebacterium diphtheriae (*C. diphtheriae*) is a relatively rare pathogen in most Western countries. While toxin producing strains can cause pharyngeal diphtheria with potentially fatal outcomes, the more common presentation is wound infections. The diphtheria toxin is encoded on a prophage and can also be carried by *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis*. Currently, across Europe, infections are mainly diagnosed in travelers and refugees from regions where diphtheria is more endemic, patients from urban areas with poor hygiene, and intravenous drug users. About half of the cases are non-toxin producing isolates. Rapid identification of the bacterial pathogen and toxin production is a critical element of patient and outbreak management. Beside the immediate clinical management of the patient, public health agencies should be informed of toxigenic *C. diphtheriae* diagnoses as soon as possible. The collection of case-related epidemiological data from the patient is often challenging due to language barriers and social circumstances. However, information on patient contacts, vaccine status and travel/refugee route, where appropriate, is critical, and should be documented. In addition, isolates should be characterized using high resolution typing, in order to identify transmissions and outbreaks. In recent years, whole genome sequencing (WGS) has become the gold standard of high-resolution typing methods, allowing detailed investigations of pathogen transmissions. De-centralized sequencing strategies with redundancy in sequencing capacities, followed by data exchange may be a valuable future option, especially since WGS becomes more available and portable. In this context, the sharing of sequence data, using public available platforms, is essential. A close interaction between microbiology laboratories, treating physicians, refugee centers, social workers, and public health officials is a key element in successful management of suspected outbreaks. Analyzing bacterial isolates at reference centers may further help to provide more specialized microbiological techniques and to standardize information, but this is also more time consuming during an outbreak. Centralized communication strategies between public health agencies and laboratories helps considerably in establishing and coordinating effective surveillance and infection control. We review the current literature on high-resolution typing of *C. diphtheriae* and share our own experience with the coordination of a Swiss-German outbreak.

Keywords: *Corynebacterium diphtheriae*, diphtheria, surveillance, public health, whole genome sequencing, molecular epidemiology, toxin

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control (ECDC), Sweden

Reviewed by:
Andreas Sing, Bavarian State Office for Health and Food Safety, Germany
Andreas Burkovski, University of Erlangen Nuremberg, Germany

***Correspondence:**
Adrian Egli
adrian.egli@usb.ch

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *14 May 2019*
Accepted: *06 August 2019*
Published: *21 August 2019*

Citation:
Seth-Smith HMB and Egli A (2019) Whole Genome Sequencing for Surveillance of Diphtheria in Low Incidence Settings. Front. Public Health 7:235. doi: 10.3389/fpubh.2019.00235

INTRODUCTION

In recent years, rare but hypervirulent pathogens have been increasingly reported in specific geographic regions (1, 2), often associated with refugees and asylum seekers (3), but also in other high-risk populations including hospitalized patients (4, 5), the elderly, and newborns (6, 7). Reports of infections in refugees over the past decade have included *Borrelia recurrentis* (8), methicillin resistant *Staphylococcus aureus* (*S. aureus)* (MRSA) (9, 10), and toxigenic *Corynebacterium diphtheriae* (11). In 2016, the European Center for Disease Control (ECDC) warned about increased rates of cutaneous *C. diphtheriae* infections in Europe due to the refugee crisis (12). This pathogen came back into the focus of attention as it is (i) associated with severe infections in humans, including respiratory diphtheria (13–15); (ii) highly transmittable, indicated by the basis reproduction number with mean 7.2 (16); and (iii) known to cause larger outbreaks (17–19). For nearly two decades, in most high-income countries, cases have been reported rarely, occasionally in travel returners (20–24), drug users and homeless people (25–29). In the last few years, in contrast, cutaneous, and respiratory infections have predominantly been reported in refugees (16, 30–38).

Providing state-of-the-art diagnostics for rare and unexpected pathogens can be a challenge for the clinician (39) and the routine microbiology laboratory (40–42). Often specific diagnostic tests are only available in reference laboratories, thus further delaying efficient therapy, surveillance reporting, and outbreak management. Once the pathogen is cultured and identified, molecular typing technologies, such as whole genome sequencing (WGS), allow a detailed comparison on the genomic level with high resolution (43–45). In the case of *C. diphtheriae*, high-resolution typing is helpful to (i) provide the epidemiological broader context (35) and (ii) include or exclude transmission events between patients (30, 31).

WGS specifically, gives the highest resolution typing, and can help to identify potential sources and transmission routes as part of modern surveillance technologies. Recent comparisons using WGS data analyzed by core genome MLST (cgMLST) or single nucleotide polymorphisms (SNP)-based methods have shown significant improvements over older technologies (46, 47). The advantages of using WGS for high-resolution typing has been seen in several pathogens, being particularly helpful in settings with (i) highly similar isolates over a long time period e.g., *Legionella pneumophila* within a city (48) or *C. difficile* (49, 50), (ii) a low endemic epidemiological background, but multiple clusters of patients from high endemic region with potential transmission events e.g., *C. diphtheriae* (31) or *M. tuberculosis* (51), and (iii) high endemic burden, where transmission events cannot easily be separated based on classical epidemiological information alone.

Alongside the availability of rapid diagnostic tests and high-resolution typing, surveillance programs are an important cornerstone of public health, as the associated framework allows data collection, communication, and coordination of public health interventions. Of note, to date no global or European surveillance network exists which integrates both classical and molecular epidemiological data into a single real-time updated platform. Future surveillance programs may not only incorporate baseline features of an isolate such as sequence type and presence or absence of the *tox* gene, but also more detailed genomic analysis and a virulence factor profile. The aim of this would be to better assess the potential of a strain to cause outbreaks with more severe clinical phenotypes. In this review article, we will focus on *C. diphtheriae* as a re-emerging but rare pathogen, and will discuss the various aspects of classical and molecular epidemiology utilizing new sequencing technologies for surveillance.

MICROBIOLOGY AND PATHOGENICITY OF *C. diphtheriae*

Corynebacterium diphtheriae was first isolated in 1884 by Loeffler (52). The classical presentation is pharyngeal diphtheria, a toxin-mediated infectious disease of the upper respiratory tract. The hallmark feature is an inflamed pseudo-membrane on the pharynx, potentially causing asphyxia (13). Beside respiratory infections, *C. diphtheriae* may cause skin infections and other invasive diseases such as endocarditis, osteomyelitis, and septic arthritis (53–58). At the moment, non-toxigenic cutaneous diphtheria is the most prevalent clinical presentation (24, 39, 57, 59, 60). Wound infections often occur with other skin pathogens, such as *Streptococcus pyogenes* or *S. aureus* (28, 31). Cutaneous diphtheria may be a source of toxigenic pathogens and may be transferred to other body sites then potentially causing respiratory diphtheria. Therefore, even wound infections with non-toxigenic strains might ideally be considered to be reported to surveillance programs in order to identify carriers, clusters of potential transmissions, and high-risk groups.

Microbiology

The species *C. diphtheriae* is divided into four biochemical biovars—belfanti, gravis, intermedius, and mitis (15, 61). Although the biochemical distinctions are not reliable, for historical reasons reference laboratories still use them. Recently, two distinct subspecies have been proposed based on genomic features: *C. diphtheriae* subsp. *diphtheriae* and *C. diphtheriae* subsp. *lausannense*. Of interest, members of the newly described subspecies *lausannense* show a larger genome size and are enriched in genes related to transport and metabolism of lipids and inorganic ion (62). On the other hand, the new subspecies lacks all genes involved in the synthesis of pili, molybdenum cofactor, and nitrate reductase. Closely related to *C. diphtheriae* are two zoonotic pathogens, *C. ulcerans* and *C. pseudotuberculosis* (63), both of which can acquire the toxin gene via a phage (64). Increasing numbers of toxigenic *C. ulcerans* infections have been reported (65, 66) e.g., in the UK (67), but these pathogens remain rare in the clinic. Host jumps from domesticated and wild animals to humans have been postulated (63, 68, 69). If either *C. ulcerans* or *C. pseudotuberculosis* is diagnosed, the isolate should be tested for the presence of the toxin and reported in surveillance programs.

Virulence Factors

The ß-corynephage encodes the diphtheria toxin, and can be transmitted between isolates. The ß-corynephage may pose a survival benefit for the bacterium by increasing the effectiveness of transmission by helping to cause local tissue damage (14, 70). The DtxR regulator is present elsewhere in the genome, and controls the transcription of the toxin gene (*tox*). This regulator is a key determinant for iron homeostasis (71). Iron is crucial for a number of cellular functions and the expression of a toxin in situations with low iron concentrations might help pathogens to compete with the host for iron or release iron via lysis of host cells. Of particular importance are pili encoded by *spa* operons (*spaABC, spaDEF*, and *spaHIG*), which contribute to the interaction with the host. Gain or loss of the function of these genes correlate to the number and expression of pili on the cell surface—especially the major pilin genes *spaA*, *spaD*, and *spaH*. The *spaA*- *spaD*- and *spaH*-type pili interact with the pharyngeal, laryngeal, and lung epithelial cell types, respectively (72). Pilus expression may strongly influence the virulence of a strain (73–78), especially in combination with the presence of the *tox* gene.

Diagnostic Aspects

Specific culture media such as tellurite agar improves the culture of *C. diphtheriae* (61, 79)—although the agar adds some selection, most diagnostic laboratories do not carry the agar as part of routine stock. The three species of interest, *C. diphtheriae*, *C. ulcerans,* and *C. pseudotuberculosis,* can be reliable identified with matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) (80–82). More specialized laboratories have the capacity to detect the presence of the diphtheria toxin either by PCR or measurement of toxin production by a modified Elek test (31, 83, 84). Other virulence factors such as pili are generally not determined in routine diagnostics. A survey of the diphtheria surveillance network (DIPNET) indicated that many centers were not able to isolate the target organisms, and most found difficulties differentiating them from specimens that contained *Corynebacterium striatum*, a commensal contaminant (85). More recently, an ECDC technical report on the diagnostic gaps has been published (86). Regular workshops and external quality assessments are important aspects in maintaining diagnostic quality for rare pathogens in the context of a surveillance program.

THE RETURN OF AN OLD FOE

Importance of Vaccination

In 2016 the EDSN reported 47 laboratory confirmed cases of *C. diphtheriae* and *C. ulcerans* in European countries—corresponding to an overall notification rate below 0.01 cases per 100,000 people (66). In contrast, worldwide, 7097 diphtheria cases were reported in 2016, mainly in low-income countries (www.who.it). In the 1900s−1950s, infections with *C. diphtheriae* were among the most severe infections during childhood, especially in pre-school children where case fatality rates of 2–25% were reported (87, 88). Prior to the availability of toxoid-vaccines, nearly 70% of the cases were in children younger than 15 years of age (89). With the introduction of vaccines in the 1940s and 1950s, a significant decrease in incidence was observed (87, 88, 90–93), although no controlled clinical trial to evaluate the efficacy of the toxoid-vaccines in preventing diphtheria has ever been performed.

The current WHO recommendation states that a series of three toxoid-vaccine doses should be provided, starting at six weeks of age, with additional booster doses based on local epidemiology (94). Vaccine effectiveness is high after three or more doses, ranging from 96 to 98% (95, 96). Although not assessed in routine, an antitoxin level of 0.01 IU/mL provides the lowest level of protection, 0.1 IU/mL is considered a protective level, and levels of >1.0 IU/mL result with long term protection (61). Interestingly, two cases of fatal diphtheria in patients with antitoxin levels above 30 IU/mL have been reported, suggesting that no absolute protection exists (97). Although immunization programs of infants started in the late 1970s, the vaccine coverage rates of infants in developing countries increased only slowly from 46% in 1985 to 79% in 1992 (98). If vaccines rates in the general population are too low, herd immunity fails to protect the non-vaccinated population, resulting in outbreaks with the potential for high mortality in younger and older age groups. An assessment of the immunity against a series of pathogens in adult asylum seekers in the Netherlands showed median 82% seroprotective anti-toxin titers against diphtheria (99). Although diphtheria vaccine rates in infants range from 89 to 98% in most European countries, a recent meta-analysis showed that vaccine rates against diphtheria and tetanus toxoids, and acellular pertussis (dTap) in healthcare workers was only 45.1% in the US and 63.9% in France (100). In Luxembourg only 2.5% of individuals under the age of 20 were seronegative, while 42% of individuals over the age of 40 years were seronegative (101). Similar low seroprotection rates have been documented in China, where only 34.1% of subjects older than 40 years were seroprotected (102). The reason for low seroprotection in some population groups in countries, where the vaccine is available, may result in a decrease in circulating toxigenic *C. diphtheriae* isolates (89), resulting in (i) an increase in non-toxigenic cases (103), and (ii) lower natural boost effects of antibody titers against the toxin (104). Especially in the adult population, gaps in herd immunity have been described due to waning of protective antibodies either from lower natural exposure or booster-vaccination. It has been found that the diphtheria vaccination only prevents symptomatic infection, and does not inhibit carriage or transmission of the pathogen. Miller and colleagues have shown that a high percentage of *C. diphtheriae* carriers were fully vaccinated, suggesting that antibodies against the toxin does not inhibit nasopharyngeal colonization (93). Based on this data, we may conclude that adults and the elderly are at higher risk of *C. diphtheriae* infection. Regular assessment of seroprotection rates in a given population should be a part of surveillance programs.

Changing Epidemiology

In the 1960–1970s, any outbreaks described in high income countries were smaller (92, 105–108) in comparison to the larger outbreaks which occurred in the late 1990s and early 2000s,

particularly in countries of the former Soviet Union (17–19, 109–113). A very large outbreak affected states of the former Soviet Union with more than 150,000 infected people and between 3,000 and 5,000 deaths (18). In this outbreak, a high proportion of adults were affected, potentially due to disruption of health services resulting in poor vaccine coverage (114, 115) and reduced "natural" exposure over the preceding decades, resulting in antibody titers below protective levels (116–118). In recent years, multiple outbreaks, or potential transmission clusters have been reported in: Bangladesh (119, 120), Brazil (121), Colombia (122), Germany (30, 35), India (123–125), Indonesia (126), Laos (127), Norway (128), Nigeria (129), Poland (130), Spain (38), South Africa (36, 131), Syria (132), Switzerland (31), Thailand (114), the United Kingdom (37), Venezuela (133, 134), and Yemen (135). The global list of affected countries indicates that (i) the disease is remains poorly controlled, (ii) the main burden lies in low-income countries, and (iii) local and global surveillance should be intensified in order to better control the disease.

EPIDEMIOLOGY: FROM CLASSICAL TO MOLECULAR

Some of the key factors driving the spread of hypervirulent pathogens include poor vaccine rates, waning antibody titers, reduced access to healthcare, failing, or collapsing healthcare systems, poor hygiene, transfer of patients between healthcare institutions, changes in travel behaviors, increased traveling to high endemic regions, and migration from high endemic regions due to violent conflicts or for economic reasons (136–138). The development of effective preventative strategies to reduce the impact of hypervirulent bacteria should, as for multidrug resistant (MDR) pathogens, have a top global priority among public health experts, clinical microbiologists, and infectious diseases physicians. The basis for preventative strategies relies on two key elements: classical and molecular epidemiological data.

Classical epidemiological methods are used to investigate an unexpected frequency of specific pathogens clustering within a certain time and/or geographical range. Determining a case definition is an important first step. Cases have to be confirmed, background rates established, and patient data collected via, for example, structured questionnaire, and accessing detailed medical history. Thus, a hypothesis for the disease transmission can be formulated and potential sources named (139, 140). Although classical epidemiological methodologies provide tremendously important information, data collection is often challenging due to delayed or incomplete reporting of cases, lack of centralized communication strategies, especially at the beginning of an outbreak, vague medical history, language barriers, and cultural differences. Especially in the case of refugees, where classical epidemiological data are often not reliable, available or re-constructible, in many cases classical methods cannot provide the required data.

Molecular epidemiological methods are based on detailed comparison of pathogens, using some or all of the genomic information. The relatedness of pathogens can be visualized in trees, thereby helping to cluster isolates and provide information on potential molecular epidemiological links. Several genotyping approaches have been used for *C. diphtheriae* including ribotyping, amplified fragment length polymorphisms, PFGE, random amplified polymorphic DNA (RAPD), clustered regularly interspaced short palindromic repeat (CRISPR)-based spoligotyping and MLST (141–149). Some typing methods show better resolution than others: ribotyping outperforms PFGE and AFLP in terms of discriminatory power (143). Ribotyping was for many years considered the gold-standard before the introduction of a robust MLST approach. Many ribotypes were allocated a geographical name based on the location of the initial isolate, however some followed an arbitrary nomenclature (144). CRISPR-based spoligotyping can offer additional resolution within ribotypes, and be used successfully to further characterize outbreak-associated strains (147, 148): the epidemic strains from the former Soviet Union belonged to two ribotypes (St. Petersburg and Rossija) that could be subdivided into 45 additional spoligotypes (146, 147). Data from various outbreaks shows the relative high molecular diversity of isolates indicating that new strains are emerging regularly within this species (150).

A robust MLST scheme was developed in 2010, including the genes *atpA*, *dnaE*, *dnaK*, *fusA*, *leuA*, *odhA*, and *rpoB* (www.pubmlst.org/cdiphtheriae). The advantages of an MLST scheme include transferability and comparability. The sequence types were shown to be consistent with the previously determined ribotypes and offered higher resolution in most cases (141). MLST diversity has grown continuously, with 608 types currently categorized (March 2019). Of note, the MLST scheme lacks the biochemical correlation of the biovar system and STs have not been able to be associated with a more severe clinical phenotype (141, 151, 152).

Comparison of the performance of various typing techniques is important, as low resolution typing methods may overcall transmission events masking the real transmission steps and potentially delaying the identification of the source. Stucki et al. showed this for *M. tuberculosis* transmissions events in Switzerland, where a VNTR low-resolution typing gave evidence of a significantly higher rate of transmissions events in comparison to WGS based typing on the same set of isolates (153). Similarly, *C. diphtheriae* SNP-based WGS comparisons improved the typing resolution in comparison to cgMLST (35).

WHOLE GENOME SEQUENCING OF *C. diphtheriae*

The first complete genome sequence of *C. diphtheriae* (strain NCTC13129) was analyzed in 2003, a UK clinical isolate containing a series of pathogenicity factors including iron-uptake systems, adhesins and fimbrial proteins (154). The genome of *C. diphtheriae* is 2.45 Mbp with a G+C content of 53.5% (154). Through WGS analysis we can determine the presence of virulence factors such as the toxin gene (and ß-corynephage) and pili, and genes encoding antimicrobial resistance determinants (62, 155, 156). During outbreak and public health investigations, WGS SNP-based typing clearly shows important benefits due to its high resolution (31). Although MLST may be more cost

effective, MLST data can also be extracted from WGS data, providing the ST as well as high resolution phylogeny and additional important genetic information. WGS can identify additional toxins and adherence factors, which may allow the generation of a specific risk profile for the pathogen.

Comparative studies have shown that the species has a set of ∼1,630 core genes which almost every representative of this species possesses [60% of the genome], and a relatively large, open pan-genome (155, 156). The difference in genome content across the species is largely due to the presence of genomic islands, prophages, transposons, restriction-modification systems, and CRISPR elements. Horizontal transfer substantially helps to shape the bacterial genome (62, 155). Some of the identified genomic islands carry genes for siderophore synthesis and transportation and degradation of polysaccharides, and heavy metal resistance. Interestingly, prophages are genetically more similar within specific clusters of bacterial isolates than between clusters, suggesting that prophages do not randomly mix between isolates, but rather cluster within specific clades (31, 157, 158).

While MLST analysis first suggested, that there is significant recombination within *C. diphtheriae* (141), this has been confirmed through analysis of whole genome sequences (159). Recombination plays an important role in bacterial evolution and has been linked to increased virulence in some pathogens (160–162). Especially in the upper respiratory tract, where *C. diphtheriae* can form a colonizing state, horizontal gene transfer can commonly happen (163). WGS allowed to study genetic ancestry of multiple bacterial species—including *C. diphtheriae*. This challenged sometimes our current understanding and groups based on biochemistry or serotypes may change. As an example, it has also been shown that biovars of *C. diphtheriae* do not correlate to genetic ancestry (152, 159). In recent years, several cohorts of *C. diphtheriae* isolates have been analyzed using WGS (30, 31, 35, 36, 62, 152, 155, 156, 164–167). Comparison of WGS data across a species generally uses one of two approaches: cgMLST, or SNP-based variant calling across the whole genome based on a reference, which provides more information and higher resolution. Dangel et al. have generated a cgMLST scheme including 1553 target loci and an extended cgMLST scheme including 2154 target loci, providing higher resolution (35).

cgMLST and SNP-based analyses of all publicly available whole genome sequences (**Figures 1**, **2** and **Supplementary Table 1**) shows vast diversity, and geographic mixing: isolates identified in Malaysia, India, Australia, and Switzerland are found throughout the trees. Relatively few cgMLST clusters are defined at the five allele cut off, yet some clades/clusters clearly show geographic association, such as those from South Africa, Belarus and Germany (35), suggestive of local outbreaks. The largest clade of highly related isolates, at the top of **Figure 2**, includes those from Germany, Poland, the UK and the former Soviet Union, suggesting that these may have had a common source, but spread prior to diagnosis (This clustering is not represented in the minimum spanning tree of **Figure 1**). However, the dates of the isolates in this clade range from 1996 to 2017, also suggesting some stability of the isolates over time. This is also evidenced as closely related isolates throughout the tree may have been isolated many decades apart.

Clustering and Likelihood of Transmission

There is an ongoing debate about defining diversity thresholds to separate clusters of pathogens. Determining a threshold of diversity to reliably describe a transmission cluster is a question commonly asked, yet difficult to answer, particularly in recombinogenic bacteria. Dangel et al. defined a cluster in their cgMLST scheme as five or fewer allele differences, with higher resolution of subclusters analyzed through an extended cgMLST scheme (35).

In order to determine a reliable cut-off, it is beneficial to combine the genomic analysis with more classical epidemiological data, which significantly contributes to understanding the transmission risks. However, in the literature and epidemiological data associated with WGS, few such cases have been described: in one case of direct transmission between siblings, the isolates show no allele differences in the defined core genome or accessory genome (30); and one case of direct transmission from mother to twin newborns showed a single SNP between the isolates on a whole genome level (unpublished data) and zero allele differences in the cgMLST scheme (**Figure 1**).

During our study on isolates from refugees in Basel, we asked ourselves if the observed whole genome diversity of 50–150 SNPs within clusters could represent a recent transmission event. We considered two different mutation rates representing extremes of plausible ranges, and estimated the approximate transmission dynamic. Even using a very high mutational rate of 0.00018 substitutions/bp/year, the estimation indicated that transmission occurred more than four to 6 weeks prior to sampling. In that paper, we played with substitution rates an picked the mutation rate of *Helicobacter pylori*, in order to have a highly conservative estimated if the transmission occurred on European ground to trigger potential outbreak investigations. This helped to exclude a transmission event within Europe, as the affected refugees arrived 2 weeks prior in South Italy (31). Analyzing these clusters by cgMLST shows that the isolates diverge by 0–4 alleles (**Figure 1**), within the cluster threshold, despite possessing at least 50 SNP differences and not representing recent transmission (31). This exemplifies the increased resolution of using whole genome SNP-based methods, and the difficulty of inferring direct transmissions from cgMLST data alone. As *C. diphtheriae* can also undergo recombination, it is crucial to consider a recent recombination by studying the distribution of SNPs across the genome: if many SNPs cluster in one or more genomic loci, then a recombination event is likely to have occurred, bringing the putative transmission event more recent.

SURVEILLANCE

Although country specific surveillance systems for hypervirulent pathogens such as *C. diphtheriae* exist, the interoperability of data and the exchange across countries presents problems (170). In 2014, a WHO-recommended surveillance standard of diphtheria was published. This included a case definition, laboratory

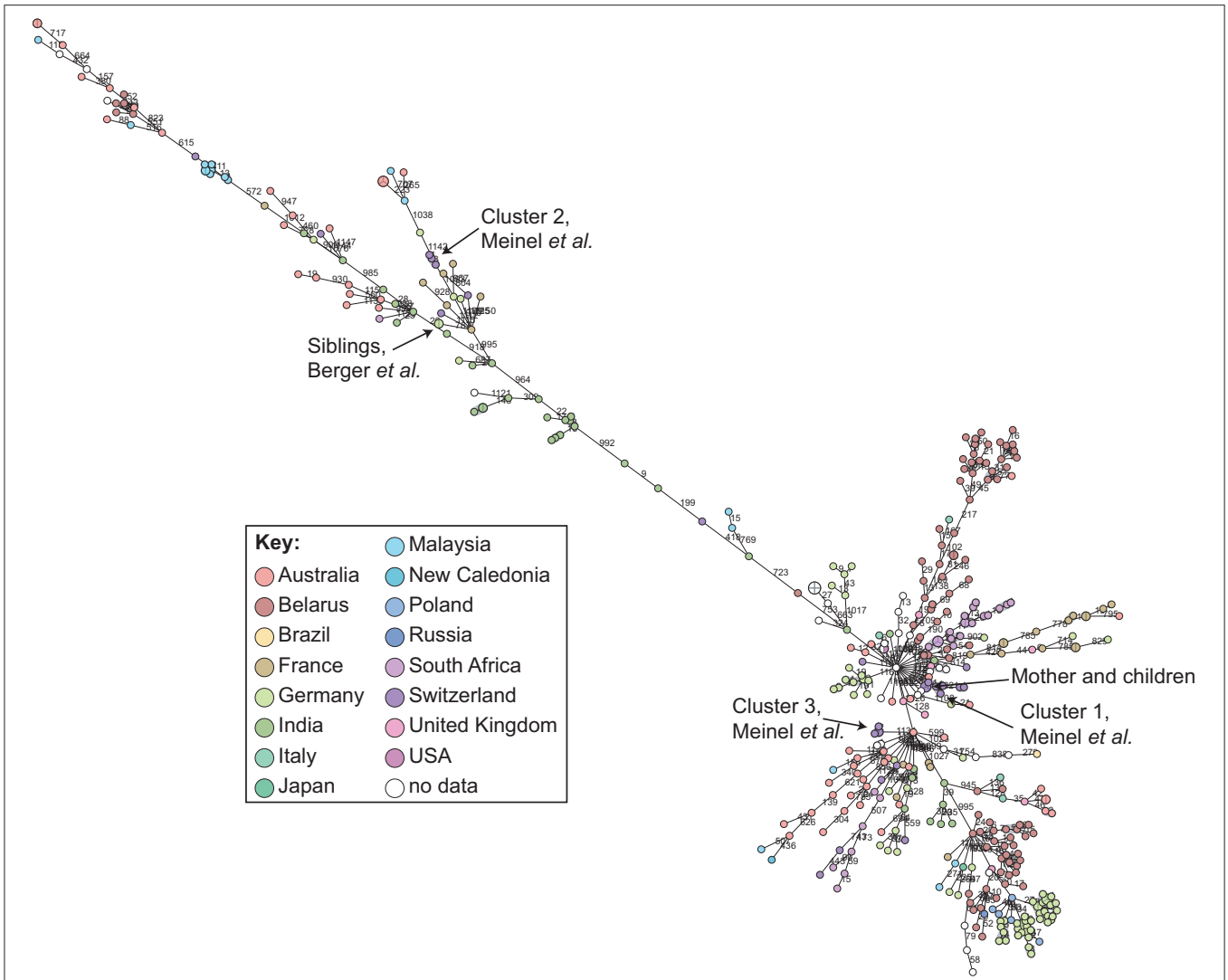


FIGURE 1 | Minimum spanning tree showing relationship of all available *Corynebacterium diphtheriae* genomes by cgMLST. All *C. diphtheriae* assemblies available from NCBI on 27.02.2019 were retrieved, and compared to isolates from our laboratory, as well as all reads available from NCBI on 08.10.2018, and those published, which were assembled using unicycler (168), and duplicate samples removed, giving $n = 419$ genomes. The cgMLST scheme of (35) was used within Ridom SeqSphere+ v4.1.6, with clusters given between nodes with five or fewer differences. Nodes are colored according to country of isolation.

criteria for diagnosis, and minimum data elements which should be collected (171). Similarly, the ECDC has established a surveillance program for diphtheria. Founded in 1993 as European Laboratory Working Group in Diphtheria in 2006 it became the European Diphtheria Surveillance Network (EDSN, www.ecdc.europa.eu) (172). The network provides valuable information and aims to standardize surveillance activities and ensure availability of more comparable data between countries. It also includes laboratory components focusing on trainings and external quality assessments (EQAs), strengthening the laboratory capacity to characterize isolates and develop novel tools for molecular typing of *C. diphtheriae*.

While the EDSN provides an important framework for surveillance of *C. diphtheriae*, in the current refugee crisis, multi-national coordination of outbreak investigation is clearly a challenge. Rapid and effective mechanisms of communication are crucial. Patients may be evaluated several times on their journey, and the same pathogen may be isolated in different countries. A recent report on the tracing of an MDR *M. tuberculosis* cluster was very well-coordinated by a joint effort from multiple centers (51). Similarly, for *C. diphtheriae*, we directed an investigation with multiple refugees presenting with wound infection across different hospitals and diagnostic laboratories in Switzerland in 2015 (31). In both situations, a multi-national taskforce organized a coordinated effort to collect isolates and information, using case report forms to collect structured epidemiological information on migration routes, vaccine status, and other affected travelers.

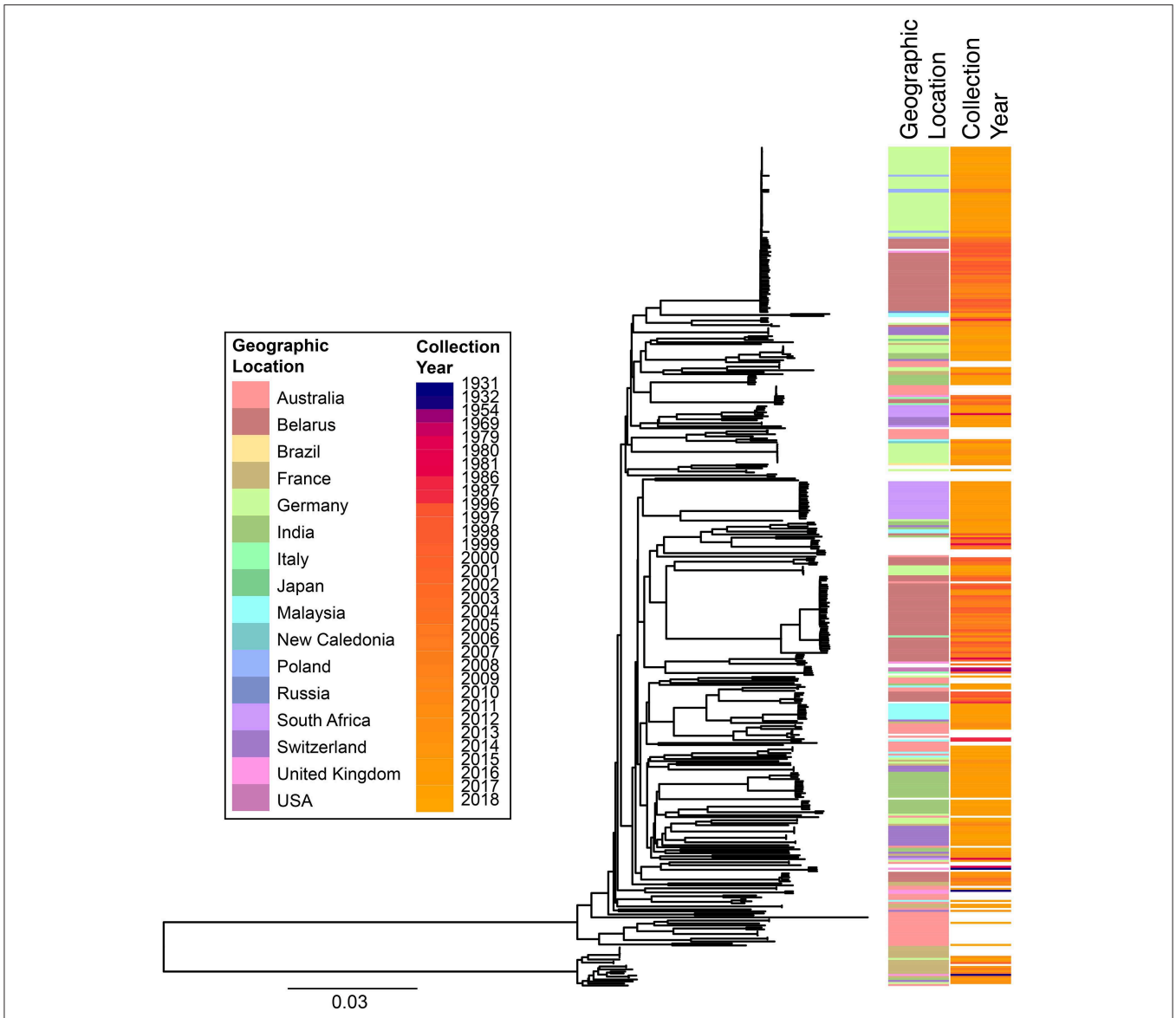


FIGURE 2 | Phylogenetic overview of all available *C. diphtheriae* genomes. All *C. diphtheriae* reads available from NCBI on 08.10.2018 were retrieved, and those published, and compared to isolates from our laboratory, as well as all assemblies available from NCBI on 27.02.2019, which were shredded to reads using wgsim in samtools (https://github.com/lh3/wgsim), and duplicate samples removed, giving $n = 419$ genomes. All reads were mapped against the reference genome CP003210 (155) within CLC Genomics Workbench 10.1.1, also used to generate a single nucleotide polymorphism (SNP) phylogeny with parameters that differed from the default as: variant calling with 10x minimum coverage, 10 minimum count and 70% minimum frequency, and SNP tree creation with 10x minimum coverage, 10% minimum coverage, 0 prune distance and including multi-nucleotide variants (MNVs). Metadata was retrieved from the NCBI database and was associated with the phylogeny using phandango (169). Colors use the same key as **Figure 1**; years are shown on a continuous scale. The bottom clade shows the clearly separate cluster proposed as *C. diphtheriae* subsp. *lausannense*.

Individual responsible experts, such as representatives of the EDSN or reference laboratories, should be assigned in each country in order to keep track of potential movements of refugees. In the *C. diphtheriae* situation, refugees were rapidly lost to follow-up, for example due to relocation in other refugee centers. Communication to refugee center responsible personnel and physicians should be established. The molecular epidemiology of diphtheria would certainly benefit from implementation of WGS. Such analysis offers improvements over the current model of global tracing of large clonal clusters toward fine-tuned strain discrimination. At the same time, a multicenter evaluation of recently developed inexpensive and discriminatory VNTR and CRISPR methods is warranted to see if and how they could complement regional surveillance (150). Beside the molecular definition of an outbreak, a centralized database allows running the standardized bioinformatic algorithms and

thereby may provide a benefit for investigations. Isolates could be registered with particular coded identifiers to avoid re-sequencing the same isolate (173).

To date, no database can integrate classical epidemiological data in the form of coded patient identification, vaccine status, potential exposures, spatiotemporal information of cases, socioeconomic and immunological data on a population level, with high-resolution molecular epidemiological data from sequenced strains. We are developing such a platform, initially for MDR pathogens (173), which could easily be expanded to hypervirulent species including *C. diphtheriae*. This Swiss Pathogen Surveillance Platform (www.spsp.ch) aims to integrate all relevant data in the near future, thereby providing various stakeholders with important information in real-time. Such a platform may provide a public health data sharing hub not only for Switzerland, but for European countries and beyond.

Warning Systems

In many countries, reporting of *C. diphtheriae* cases to public health authorities is mandatory. Information is collected and reported back to the diagnostic laboratories and infectious diseases specialist in order to heighten awareness. Various email alerting system for surveillance exists, one of the most well-known being PROMED (https://www.promedmail.org/), a subscription service which has been in place since the early 2000s (174). Those warning systems collect information from media reports, official reports, online summaries, local observers, subscribers, and others. However, those services rely on reporting toward the service and also inaccurate interpretation and privacy issues may be an issue. Nevertheless, there is still room for faster, more targeted and international ways of communication to be established. The connection of various data sources will require the usage of standardized and specific epidemiological ontologies being used across various databases such as SNOMED CT (www.snomed.org) or IRIDA (www.irida.ca). The ethical and legal implications of such big-data driven surveillance programs need to be clarified in the near future. Clearly individual patient data should be protected, but those rights should be balanced in situations where outbreaks with hypervirulent pathogens may put the general population at risk—in the case of *C. diphtheriae* the risk for the general healthy population in Western countries seems rather low and therefore surveillance efforts should rather focus on at-risk populations. Social media may be used to generate epidemiological data but could also be used as a tool to inform the general public and health care specialists. We could imagine internet-based warning systems being combined with a more detailed platform allowing clinicians to assess classical and molecular epidemiological aspects.

Machine Learning for Investigation and Surveillance of Rare Pathogens

In the near future, we can foresee interconnected databases containing epidemiological data on individual cases, incidence rates of particular infections, spatiotemporal clusters, WGS data, travel and migration information, social and print media reports, and vaccine rates in populations. These may then be used for machine learning based epidemiological surveillance, such as that recently published on prediction of dengue outbreaks (175).

Machine learning based algorithms may also be used to predict the case severity of a particular infection based on NGS and other clinical data, as similar performed by Njage et al. in the case of shigatoxigenic *E. coli* (176). Bacterial genome wide association studies (GWAS) using machine learning in *C. diphtheriae* may help to identify critical biomarkers, linking bacterial genomic features such as virulence or resistance with specific host outcomes. Such work often requires hundreds to thousands of bacterial genomes to compensate for host variability effects (177) as shown for *M. tuberculosis*, *Campylobacter* spp. and *Bordetella* spp. (178–180).

The advances in machine learning algorithms may allow the development of revolutionary surveillance programs, potentially providing valuable information to public health policy makers about potential epidemiological trends and risks for the general public. Although such databases are likely to first be established for more common epidemic scenarios such as annual influenza, MDR pathogens, and foodborne pathogens, particular risks may also be calculated for rare pathogens such as measles, ebola, or hypervirulent bacteria such as *C. diphtheriae*. As we live in an increasingly globalized world with rapid spread of pathogens, new concepts for epidemiological surveillance are needed, to enable rapid and effective interventions.

CONCLUSIONS

Corynebacterium diphtheriae is reemerging in clinics in high income countries, partly as a result of refugee movement, and requiring greater awareness of the issue. WGS offers the opportunity to describe potential transmission events and infection sources with the highest resolution. Data provided from molecular typing methods should, where possible, be analyzed in the context of classical epidemiological information, for which information has to be rapidly shared with local public health authorities. In addition, surveillance for *C. diphtheriae* and other re-emerging hypervirulent pathogens would benefit from rapid data collection and sharing platforms sharing information on classical and molecular epidemiology.

AUTHOR CONTRIBUTIONS

HS-S performed data analysis and wrote the manuscript. AE wrote the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2019.00235/full#supplementary-material

REFERENCES

1. Liu C, Guo J. Hypervirulent *Klebsiella pneumoniae* (hypermucoviscous and aerobactin positive) infection over 6 years in the elderly in China: antimicrobial resistance patterns, molecular epidemiology and risk factor. *Ann Clin Microbiol Antimicrob*. (2019) 18:4. doi: 10.1186/s12941-018-0302-9
2. Acevedo R, Bai X, Borrow R, Caugant DA, Carlos J, Ceyhan M, et al. The global meningococcal initiative meeting on prevention of meningococcal disease worldwide: epidemiology, surveillance, hypervirulent strains, antibiotic resistance and high-risk populations. *Exp Rev Vaccines*. (2019) 18:15–30. doi: 10.1080/14760584.2019.1557520
3. Isenring E, Fehr J, Gultekin N, Schlagenhauf P. Infectious disease profiles of Syrian and Eritrean migrants presenting in Europe: a systematic review. *Travel Med Infect Dis*. (2018) 25:65–76. doi: 10.1016/j.tmaid.2018.04.014
4. Couturier J, Davies K, Gateau C, Barbut F. Ribotypes and new virulent strains across Europe. *Adv Exp Med Biol*. (2018) 1050:45–58. doi: 10.1007/978-3-319-72799-8_4
5. Yakob L, Riley TV, Paterson DL, Marquess J, Magalhaes RJ, Furuya-Kanamori L, et al. Mechanisms of hypervirulent *Clostridium difficile* ribotype 027 displacement of endemic strains: an epidemiological model. *Sci Rep*. (2015) 5:12666. doi: 10.1038/srep12666
6. Joubrel C, Tazi A, Six A, Dmytruk N, Touak G, Bidet P, et al. Group B streptococcus neonatal invasive infections, France 2007–2012. *Clin Microbiol Infect*. (2015) 21:910–6. doi: 10.1016/j.cmi.2015.05.039
7. Giufre M, Accogli M, Ricchizzi E, Barbanti F, Farina C, Fazii P, et al. Multidrug-resistant infections in long-term care facilities: extended-spectrum β-lactamase-producing *Enterobacteriaceae* and hypervirulent antibiotic resistant *Clostridium difficile*. *Diagn Microbiol Infect Dis*. (2018) 91:275–81. doi: 10.1016/j.diagmicrobio.2018.02.018
8. Goldenberger D, Claas GJ, Bloch-Infanger C, Breidthardt T, Suter B, Martinez M, et al. Louse-borne relapsing fever (*Borrelia recurrentis*) in an Eritrean refugee arriving in Switzerland, August 2015. *Euro surveillance*. (2015) 20:2–5. doi: 10.2807/1560-7917.ES2015.20.32.21204
9. Aro T, Kantele A. High rates of meticillin-resistant *Staphylococcus aureus* among asylum seekers and refugees admitted to Helsinki University Hospital, 2010 to 2017. *Euro surveillance*. (2018) 23:1–14. doi: 10.2807/1560-7917.ES.2018.23.45.1700797
10. Piso RJ, Kach R, Pop R, Zillig D, Schibli U, Bassetti S, et al. Correction: a Cross-sectional study of colonization rates with methicillin-resistant *Staphylococcus aureus* (MRSA) and Extended-Spectrum Beta-Lactamase (ESBL) and carbapenemase-producing enterobacteriaceae in four swiss refugee centres. *PLoS ONE*. (2017) 12:e0174911. doi: 10.1371/journal.pone.0174911
11. Bloch-Infanger C, Battig V, Kremo J, Widmer AF, Egli A, Bingisser R, et al. Increasing prevalence of infectious diseases in asylum seekers at a tertiary care hospital in Switzerland. *PLoS ONE*. (2017) 12:e0179537. doi: 10.1371/journal.pone.0179537
12. ECDC. *Cutaneous Diphtheria Among Recently Arrived Refugees and Asylum Seekers in the Eu* (2016).
13. Hadfield TL, McEvoy P, Polotsky Y, Tzinserling VA, Yakovlev AA. The pathology of diphtheria. *J Infect Dis*. (2000) 181(Suppl 1):S116–20. doi: 10.1086/315551
14. Collier RJ. Diphtheria toxin: mode of action and structure. *Bacteriol Rev*. (1975) 39:54–85.
15. Funke G, von Graevenitz A, Clarridge JE III, Bernard KA. Clinical microbiology of coryneform bacteria. *Clin Microbiol Rev*. (1997) 10:125–59. doi: 10.1128/CMR.10.1.125
16. Matsuyama R, Akhmetzhanov AR, Endo A, Lee H, Yamaguchi T, Tsuzuki S, et al. Uncertainty and sensitivity analysis of the basic reproduction number of diphtheria: a case study of a Rohingya refugee camp in Bangladesh, November-December 2017. *PeerJ*. (2018) 6:e4583. doi: 10.7717/peerj.4583
17. Rakhmanova AG, Lumio J, Groundstroem K, Valova E, Nosikova E, Tanasijchuk T, et al. Diphtheria outbreak in St. Petersburg: clinical characteristics of 1860 adult patients. *Scand J Infect Dis*. (1996) 28:37–40. doi: 10.3109/00365549609027147
18. Markina SS, Maksimova NM, Vitek CR, Bogatyreva EY, Monisov AA. Diphtheria in the russian federation in the 1990s. *J Infect Dis*. (2000) 181(Suppl 1):S27–34. doi: 10.1086/315535
19. Popovic T, Kombarova SY, Reeves MW, Nakao H, Mazurova IK, Wharton M, et al. Molecular epidemiology of diphtheria in Russia, 1985–1994. *J Infect Dis*. (1996) 174:1064–72. doi: 10.1093/infdis/174.5.1064
20. FitzGerald RP, Rosser AJ, Perera DN. Non-toxigenic penicillin-resistant cutaneous *C. diphtheriae* infection: a case report and review of the literature. *J Infect Public Health*. (2015) 8:98–100. doi: 10.1016/j.jiph.2014.05.006
21. Lindhusen-Lindhe E, Dotevall L, Berglund M. Imported laryngeal and cutaneous diphtheria in tourists returning from western Africa to Sweden, March 2012. *Euro surveillance*. (2012) 17:20189.
22. May ML, McDougall RJ, Robson JM. *Corynebacterium diphtheriae* and the returned tropical traveler. *J Travel Med*. (2014) 21:39–44. doi: 10.1111/jtm.12074
23. Jakovljev A, Steinbakk M, Mengshoel AT, Sagvik E, Brugger-Synnes P, Sakshaug T, et al. Imported toxigenic cutaneous diphtheria in a young male returning from Mozambique to Norway, March 2014. *Euro surveillance*. (2014) 19:1–4. doi: 10.2807/1560-7917.ES2014.19.24.20835
24. Nelson TG, Mitchell CD, Sega-Hall GM, Porter RJ. Cutaneous ulcers in a returning traveller: a rare case of imported diphtheria in the UK. *Clin Exp Dermatol*. (2016) 41:57–9. doi: 10.1111/ced.12763
25. Gruner E, Opravil M, Altwegg M, von Graevenitz A. Nontoxigenic *Corynebacterium diphtheriae* isolated from intravenous drug users. *Clin Infect Dis*. (1994) 18:94–6. doi: 10.1093/clinids/18.1.94
26. Monsuez JJ, Mathieu D, Arnoult F, Passeron J. Cutaneous diphtheria in a homeless man. *Lancet*. (1995) 346:649–50. doi: 10.1016/S0140-6736(95)91490-0
27. Ryan TA. Infectious disease in the homeless. *Md Med*. (2008) 9:26–7.
28. Lowe CF, Bernard KA, Romney MG. Cutaneous diphtheria in the urban poor population of Vancouver, British Columbia, Canada: a 10-year review. *J Clin Microbiol*. (2011) 49:2664–6. doi: 10.1128/JCM.00362-11
29. Gubler J, Huber-Schneider C, Gruner E, Altwegg M. An outbreak of nontoxigenic *Corynebacterium diphtheriae* infection: single bacterial clone causing invasive infection among Swiss drug users. *Clin Infect Dis*. (1998) 27:1295–8. doi: 10.1086/514997
30. Berger A, Dangel A, Schober T, Schmidbauer B, Konrad R, Marosevic D, et al. Whole genome sequencing suggests transmission of *Corynebacterium diphtheriae*-caused cutaneous diphtheria in two siblings, Germany, 2018. *Euro surveillance*. (2019) 24:1–4. doi: 10.2807/1560-7917.ES.2019.24.2.1800683
31. Meinel DM, Kuehl R, Zbinden R, Boskova V, Garzoni C, Fadini D, et al. Outbreak investigation for toxigenic *Corynebacterium diphtheriae* wound infections in refugees from Northeast Africa and Syria in Switzerland and Germany by whole genome sequencing. *Clin Microbiol Infect*. (2016) 22:1003 e1–8. doi: 10.1016/j.cmi.2016.08.010
32. Reynolds GE, Saunders H, Matson A, O'Kane F, Roberts SA, Singh SK, et al. Public health action following an outbreak of toxigenic cutaneous diphtheria in an Auckland refugee resettlement centre. *Commun Dis Intell Q Rep*. (2016) 40:E475–81.
33. Sane J, Sorvari T, Widerstrom M, Kauma H, Kaukoniemi U, Tarkka E, et al. Respiratory diphtheria in an asylum seeker from Afghanistan arriving to Finland via Sweden, December 2015. *Euro surveillance*. (2016) 21:14–17. doi: 10.2807/1560-7917.ES.2016.21.2.30105
34. Scheifer C, Rolland-Debord C, Badell E, Reibel F, Aubry A, Perignon A, et al. Re-emergence of *Corynebacterium diphtheriae*. *Med Mal Infect*. (2018) doi: 10.1016/j.medmal.2018.12.001. [Epub ahead of print].
35. Dangel A, Berger A, Konrad R, Bischoff H, Sing A. Geographically diverse clusters of nontoxigenic *Corynebacterium diphtheriae* infection, Germany, 2016–2017. *Emerging Infect Dis*. (2018) 24:1239–45. doi: 10.3201/eid2407.172026
36. du Plessis M, Wolter N, Allam M, de Gouveia L, Moosa F, Ntshoe G, et al. Molecular characterization of *Corynebacterium diphtheriae* outbreak isolates, South Africa, March-June 2015. *Emerging Infect Dis*. (2017) 23:1308–15. doi: 10.3201/eid2308.162039
37. Edwards D, Kent D, Lester C, Brown CS, Murphy ME, Brown NM, et al. Transmission of toxigenic *Corynebacterium diphtheriae* by a fully immunised resident returning from a visit to West Africa, United Kingdom, 2017. *Euro surveillance*. (2018) 23:1700681. doi: 10.2807/1560-7917.ES.2018.23.39.1700681

38. Jane M, Vidal MJ, Camps N, Campins M, Martinez A, Balcells J, et al. A case of respiratory toxigenic diphtheria: contact tracing results and considerations following a 30-year disease-free interval, Catalonia, Spain, 2015. *Euro surveillance*. (2018) 23:1–6. doi: 10.2807/1560-7917.ES.2018.23.13.17-00183
39. Cassir N, Bagneres D, Fournier PE, Berbis P, Brouqui P, Rossi PM. Cutaneous diphtheria: easy to be overlooked. *Int J Infect Dis.* (2015) 33:104–5. doi: 10.1016/j.ijid.2015.01.008
40. Bloss S, Klemann C, Rother AK, Mehmecke S, Schumacher U, Mucke U, et al. Diagnostic needs for rare diseases and shared prediagnostic phenomena: results of a urvey. *PLoS ONE.* (2017) 12:e0172532. doi: 10.1371/journal.pone.0172532
41. Caliendo AM, Gilbert DN, Ginocchio CC, Hanson KE, May L, Quinn TC, et al. Better tests, better care: improved diagnostics for infectious diseases. *Clin Infect Dis.* (2013) 57(Suppl 3):S139–70. doi: 10.1093/cid/cit578
42. van Duin D. Diagnostic challenges and opportunities in older adults with infectious diseases. *Clin Infect Dis.* (2012) 54:973–8. doi: 10.1093/cid/cir927
43. Kozinska A, Seweryn P, Sitkiewicz I. A crash course in sequencing for a microbiologist. *J Appl Genet.* (2019) 60:103–11. doi: 10.1007/s13353-019-00482-2
44. McNerney R, Zignol M, Clark TG. Use of whole genome sequencing in surveillance of drug resistant tuberculosis. *Exp Rev Anti Infect Ther.* (2018) 16:433–42. doi: 10.1080/14787210.2018.1472577
45. Mirande C, Bizine I, Giannetti A, Picot N, van Belkum A. Epidemiological aspects of healthcare-associated infections and microbial genomics. *Eur J Clin Microbiol Infect Dis.* (2018) 37:823–31. doi: 10.1007/s10096-017-3170-x
46. Revez J, Espinosa L, Albiger B, Leitmeyer KC, Struelens MJ, Points ENMF, et al. Survey on the use of whole-genome sequencing for infectious diseases surveillance: rapid expansion of european national capacities, 2015–2016. *Front Public Health.* (2017) 5:347. doi: 10.3389/fpubh.2017.00347
47. Schurch AC, van Schaik W. Challenges and opportunities for whole-genome sequencing-based surveillance of antibiotic resistance. *Ann N Y Acad Sci.* (2017) 1388:108–20. doi: 10.1111/nyas.13310
48. Wuthrich D, Gautsch S, Spieler-Denz R, Dubuis O, Gaia V, Moran-Gilad J, et al. Air-conditioner cooling towers as complex reservoirs and continuous source of Legionella pneumophila infection evidenced by a genomic analysis study in 2017, Switzerland. *Euro Surveill.* (2019) 24:1800192. doi: 10.2807/1560-7917.ES.2019.24.4.1800192
49. Endres BT, Dotson KM, Poblete K, McPherson J, Lancaster C, Basseres E, et al. Environmental transmission of *Clostridioides difficile* ribotype 027 at a long-term care facility; an outbreak investigation guided by whole genome sequencing. *Infect Control Hosp Epidemiol.* (2018) 39:1322–9. doi: 10.1017/ice.2018.230
50. Halstead FD, Ravi A, Thomson N, Nuur M, Hughes K, Brailey M, et al. Whole genome sequencing of toxigenic *Clostridium difficile* in asymptomatic carriers: insights into possible role in transmission. *J Hosp Infect.* (2018) 2:125–134. doi: 10.1016/j.jhin.2018.10.012
51. Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, et al. A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. *Lancet Infect Dis.* (2018) 18:431–40. doi: 10.1016/S1473-3099(18)30004-5
52. Loeffler F. Untersuchungen über die Bedeuteung der Mikroorganismen für die Entstehung der Diphtherie beim Menschen, bei der Taube und beim Kalbe. *Mitt KJ in Gesundh.* (1884) 2:421–99.
53. Barakett V, Morel G, Lesage D, Petit JC. Septic arthritis due to a nontoxigenic strain of *Corynebacterium diphtheriae* subspecies mitis. *Clin Infect Dis.* (1993) 17:520–1. doi: 10.1093/clinids/17.3.520
54. Belko J, Wessel DL, Malley R. Endocarditis caused by *Corynebacterium diphtheriae*: case report and review of the literature. *Pediatr Infect Dis J.* (2000) 19:159–63. doi: 10.1097/00006454-200002000-00015
55. Farfour E, Badell E, Zasada A, Hotzel H, Tomaso H, Guillot S, et al. Characterization and comparison of invasive *Corynebacterium diphtheriae* isolates from France and Poland. *J Clin Microbiol.* (2012) 50:173–5. doi: 10.1128/JCM.05811-11
56. Poilane I, Fawaz F, Nathanson M, Cruaud P, Martin T, Collignon A, et al. *Corynebacterium diphtheriae* osteomyelitis in an immunocompetent child: a case report. *Eur J Pediatr.* (1995) 154:381–3. doi: 10.1007/BF02072108
57. Romney MG, Roscoe DL, Bernard K, Lai S, Efstratiou A, Clarke AM. Emergence of an invasive clone of nontoxigenic *Corynebacterium diphtheriae* in the urban poor population of Vancouver, Canada. *J Clin Microbiol.* (2006) 44:1625–9. doi: 10.1128/JCM.44.5.1625-1629.2006
58. Tiley SM, Kociuba KR, Heron LG, Munro R. Infective endocarditis due to nontoxigenic *Corynebacterium diphtheriae*: report of seven cases and review. *Clin Infect Dis.* (1993) 16:271–5. doi: 10.1093/clind/16.2.271
59. Gordon CL, Fagan P, Hennessy J, Baird R. Characterization of *Corynebacterium diphtheriae* isolates from infected skin lesions in the Northern Territory of Australia. *J Clin Microbiol.* (2011) 49:3960–2. doi: 10.1128/JCM.05038-11
60. Huhulescu S, Hirk S, Zeinzinger V, Hasenberger P, Skvara H, Mullegger R, et al. Letter to the editor: cutaneous diphtheria in a migrant from an endemic country in east Africa, Austria May 2014. *Euro surveillance*. (2014) 19:1–2. doi: 10.2807/1560-7917.ES2014.19.26.20845
61. Efstratiou A, Maple PAC. *Manual for the Laboratory Diagnosis of Diphtheria.* WHO: The Expanded Programme on Immunization in the European Region of WHO (1994).
62. Tagini F, Pillonel T, Croxatto A, Bertelli C, Koutsokera A, Lovis A, et al. Distinct genomic features characterize two clades of *Corynebacterium diphtheriae*: proposal of *Corynebacterium diphtheriae* Subsp. diphtheriae Subsp. nov. and *Corynebacterium diphtheriae* Subsp. lausannense Subsp. nov. *Front Microbiol.* (2018) 9:1743. doi: 10.3389/fmicb.2018.01743
63. Meinel DM, Konrad R, Berger A, Konig C, Schmidt-Wieland T, Hogardt M, et al. Zoonotic transmission of toxigenic *Corynebacterium ulcerans* strain, Germany, 2012. *Emerg Infect Dis.* (2015) 21:356–8. doi: 10.3201/eid2102.141160
64. Bardsdale WL, Pappenheimer AM Jr. Phage-host relationships in nontoxigenic and toxigenic diphtheria bacilli. *J Bacteriol.* (1954) 67:220–32.
65. Sangal V, Nieminen L, Weinhardt B, Raeside J, Tucker NP, Florea CD, et al. Diphtheria-like disease caused by toxigenic *Corynebacterium ulcerans* strain. *Emerg Infect Dis.* (2014) 20:1257–8. doi: 10.3201/eid2007.140216
66. Control ECfDPa. Diphtheria. *ECDC Annual Epidemiological Report for 2016* (2018).
67. Wagner KS, White JM, Crowcroft NS, De Martin S, Mann G, Efstratiou A. Diphtheria in the United Kingdom, 1986–2008: the increasing role of *Corynebacterium ulcerans*. *Epidemiol Infect.* (2010) 138:1519–30. doi: 10.1017/S0950268810001895
68. Berger A, Boschert V, Konrad R, Schmidt-Wieland T, Hormansdorfer S, Eddicks M, et al. Two cases of cutaneous diphtheria associated with occupational pig contact in Germany. *Zoonoses Public Health.* (2013) 60:539–42. doi: 10.1111/zph.12031
69. Sing A, Konrad R, Meinel DM, Mauder N, Schwabe I, Sting R. *Corynebacterium diphtheriae* in a free-roaming red fox: case report and historical review on diphtheria in animals. *Infection*. (2016) 44:441–5. doi: 10.1007/s15010-015-0846-y
70. Pappenheimer AM Jr, Gill DM. Diphtheria. *Science*. (1973) 182:353–8. doi: 10.1126/science.182.4110.353
71. Brune I, Werner H, Huser AT, Kalinowski J, Puhler A, Tauch A. The DtxR protein acting as dual transcriptional regulator directs a global regulatory network involved in iron metabolism of *Corynebacterium glutamicum*. *BMC Genom*. (2006) 7:21. doi: 10.1186/1471-2164-7-21
72. Mandlik A, Swierczynski A, Das A, Ton-That H. *Corynebacterium diphtheriae* employs specific minor pilins to target human pharyngeal epithelial cells. *Mol Microbiol.* (2007) 64:111–24. doi: 10.1111/j.1365-2958.2007.05630.x
73. Bertuccini L, Baldassarri L, von Hunolstein C. Internalization of non-toxigenic *Corynebacterium diphtheriae* by cultured human respiratory epithelial cells. *Microb Pathog.* (2004) 37:111–8. doi: 10.1016/j.micpath.2004.06.002
74. Colombo AV, Hirata R Jr, de Souza CM, Monteiro-Leal LH, Previato JO, Formiga LC, et al. *Corynebacterium diphtheriae* surface proteins as adhesins to human erythrocytes. *FEMS Microbiol Lett*. (2001) 197:235–9. doi: 10.1016/S0378-1097(01)00113-6
75. Hirata R, Napoleao F, Monteiro-Leal LH, Andrade AF, Nagao PE, Formiga LC, et al. Intracellular viability of toxigenic *Corynebacterium*

diphtheriae strains in HEp-2 cells. *FEMS Microbiol Lett.* (2002) 215:115–9. doi: 10.1016/S0378-1097(02)00930-8
76. Peixoto RS, Antunes CA, Louredo LS, Viana VG, Santos CSD, Fuentes Ribeiro da Silva J, et al. Functional characterization of the collagen-binding protein DIP2093 and its influence on host-pathogen interaction and arthritogenic potential of *Corynebacterium diphtheriae*. *Microbiology*. (2017) 163:692–701. doi: 10.1099/mic.0.000467
77. Peixoto RS, Pereira GA, Sanches Dos Santos L, Rocha-de-Souza CM, Gomes DLR, Silva Dos Santos C, et al. Invasion of endothelial cells and arthritogenic potential of endocarditis-associated *Corynebacterium diphtheriae*. *Microbiology*. (2014) 160(Pt 3):537–46. doi: 10.1099/mic.0.069948-0
78. Mandlik A, Swierczynski A, Das A, Ton-That H. Pili in Gram-positive bacteria: assembly, involvement in colonization and biofilm development. *Trends Microbiol.* (2008) 16:33–40. doi: 10.1016/j.tim.2007.10.010
79. Hoyle L, Leeds MB. A tellurite blood-agar medium for the rapid diagnosis of diphtheria. *Lancet.* (1941) 237:175–6. doi: 10.1016/S0140-6736(00)77533-7
80. Alibi S, Ferjani A, Gaillot O, Marzouk M, Courcol R, Boukadida J. Identification of clinically relevant *Corynebacterium* strains by Api Coryne, MALDI-TOF-mass spectrometry and molecular approaches. *Pathol Biol.* (2015) 63:153–7. doi: 10.1016/j.patbio.2015.07.007
81. Barberis C, Almuzara M, Join-Lambert O, Ramirez MS, Famiglietti A, Vay C. Comparison of the Bruker MALDI-TOF mass spectrometry system and conventional phenotypic methods for identification of Gram-positive rods. *PLoS ONE.* (2014) 9:e106303. doi: 10.1371/journal.pone.0106303
82. Konrad R, Berger A, Huber I, Boschert V, Hormansdorfer S, Busch U, et al. Matrix-assisted laser desorption/ionisation time-of-flight (MALDI-TOF) mass spectrometry as a tool for rapid diagnosis of potentially toxigenic *Corynebacterium* species in the laboratory management of diphtheria-associated bacteria. *Euro surveillance.* (2010) 15:1–5. doi: 10.2807/ese.15.43.19699-en
83. Engler KH, Glushkevich T, Mazurova IK, George RC, Efstratiou A. A modified Elek test for detection of toxigenic corynebacteria in the diagnostic laboratory. *J Clin Microbiol.* (1997) 35:495–8.
84. Schuhegger R, Lindermayer M, Kugler R, Heesemann J, Busch U, Sing A. Detection of toxigenic *Corynebacterium diphtheriae* and *Corynebacterium ulcerans* strains by a novel real-time PCR. *J Clin Microbiol.* (2008) 46:2822–3. doi: 10.1128/JCM.01010-08
85. Neal SE, Efstratiou A, International Diphtheria Reference Laboratories. International external quality assurance for laboratory diagnosis of diphtheria. *J Clin Microbiol.* (2009) 47:4037–42. doi: 10.1128/JCM.00473-09
86. Czumbel I, Efstratiou A, Head V, Trindall A, West D, M. R. *European Centre for Disease Prevention and Control.* Gap analysis on securing diphtheria diagnostic capacity and diphtheria antitoxin availability in the EU/EEA: European Center for Disease Prevention and Control (2017). p. 33.
87. Wheeler SM, Morton AR. Epidemiological observations in the halifax epidemic. *Am J Public Health Nations Health.* (1942) 32:947–56. doi: 10.2105/AJPH.32.9.947
88. Stuart G. Diphtheria Incidence in European Countries. *Br Med J.* (1945) 2:613–5. doi: 10.1136/bmj.2.4426.613
89. Galazka AM, Robertson SE. Diphtheria: changing patterns in the developing world and the industrialized world. *Eur J Epidemiol.* (1995) 11:107–17. doi: 10.1007/BF01719955
90. Pezzotti P, Bellino S, Prestinaci F, Iacchini S, Lucaroni F, Camoni L, et al. The impact of immunization programs on 10 vaccine preventable diseases in Italy: 1900–2015. *Vaccine.* (2018) 36:1435–43. doi: 10.1016/j.vaccine.2018.01.065
91. van Wijhe M, Tulen AD, Korthals Altes H, McDonald SA, de Melker HE, Postma MJ, et al. Quantifying the impact of mass vaccination programmes on notified cases in the Netherlands. *Epidemiol Infect.* (2018) 146:716–22. doi: 10.1017/S0950268818000481
92. Marcuse EK, Grand MG. Epidemiology of diphtheria in San Antonio, Tex., 1970. *JAMA.* (1973) 224:305–10. doi: 10.1001/jama.224.3.305
93. Miller LW, Older JJ, Drake J, Zimmerman S. Diphtheria immunization. Effect upon carriers and the control of outbreaks. *Am J Dis Child.* (1972) 123:197–9. doi: 10.1001/archpedi.1972.02110090067004
94. WHO. *Diphtheria vaccine - WHO position paper* (2006).
95. Bisgard KM, Rhodes P, Hardy IR, Litkina IL, Filatov NN, Monisov AA, et al. Diphtheria toxoid vaccine effectiveness: a case-control study in Russia. *J Infect Dis.* (2000) 181(Suppl 1):S184–7. doi: 10.1086/315562
96. Chen RT, Hardy IR, Rhodes PH, Tyshchenko DK, Moiseeva AV, Marievsky VF. Ukraine, 1992: first assessment of diphtheria vaccine effectiveness during the recent resurgence of diphtheria in the Former Soviet Union. *J Infect Dis.* (2000) 181(Suppl 1):S178–83. doi: 10.1086/315561
97. Ipsen J. Circulating antitoxin at the onset of diphtheria in 425 patients. *J Immunol.* (1946) 54:325–47.
98. Galazka AM, Robertson SE, Oblapenko GP. Resurgence of diphtheria. *Eur J Epidemiol.* (1995) 11:95–105. doi: 10.1007/BF01719954
99. Freidl GS, Tostmann A, Curvers M, Ruijs WLM, Smits G, Schepp R, et al. Immunity against measles, mumps, rubella, varicella, diphtheria, tetanus, polio, hepatitis A and hepatitis B among adult asylum seekers in the Netherlands, 2016. *Vaccine.* (2018) 36:1664–72. doi: 10.1016/j.vaccine.2018.01.079
100. Randi BA, Sejas ONE, Miyaji KT, Infante V, Lara AN, Ibrahim KY, et al. A systematic review of adult tetanus-diphtheria-acellular (Tdap) coverage among healthcare workers. *Vaccine.* (2019) 37:1030–7. doi: 10.1016/j.vaccine.2018.12.046
101. Mossong J, Putz L, Shkedy Z, Schneider F. Seroepidemiology of diphtheria and pertussis in Luxembourg in 2000. *Epidemiol Infect.* (2006) 134:573–8. doi: 10.1017/S0950268805005662
102. Li X, Chen M, Zhang T, Li J, Zeng Y, Lu L. Seroepidemiology of diphtheria and pertussis in Beijing, China: a cross-sectional study. *Hum Vaccin Immunother.* (2015) 11:2434–9. doi: 10.1080/21645515.2015.1062954
103. Zasada AA, Baczewska-Rej M, Wardak S. An increase in non-toxigenic *Corynebacterium diphtheriae* infections in Poland–molecular epidemiology and antimicrobial susceptibility of strains isolated from past outbreaks and those currently circulating in Poland. *Int J Infect Dis.* (2010) 14:e907–12. doi: 10.1016/j.ijid.2010.05.013
104. Galazka AM, Robertson SE. Immunization against diphtheria with special emphasis on immunization of adults. *Vaccine.* (1996) 14:845–57. doi: 10.1016/0264-410X(96)00021-7
105. Anderson GS, Penfold JB. An outbreak of diphtheria in a hospital for the mentally subnormal. *J Clin Pathol.* (1973) 26:606–15. doi: 10.1136/jcp.26.8.606
106. Belsey MA, Sinclair M, Roder MR, LeBlanc DR. *Corynebacterium diphtheriae* skin infections in Alabama and Louisiana. A factor in the epidemiology of diphtheria. *N Engl J Med.* (1969) 280:135–41. doi: 10.1056/NEJM196901162800304
107. Sinclair MC, Overton R, Donald WJ. Alabama diphtheria outbreak, 1967. *HSMHA Health Rep.* (1971) 86:1107–11. doi: 10.2307/4594395
108. Zalma VM, Older JJ, Brooks GF. The Austin, Texas, diphtheria outbreak. Clinical and epidemiological aspects. *JAMA.* (1970) 211:2125–9. doi: 10.1001/jama.211.13.2125
109. Filonov VP, Zakharenko DF, Vitek CR, Romanovsky AA, Zhukovski VG. Epidemic diphtheria in Belarus, 1992–1997. *J Infect Dis.* (2000) 181(Suppl 1):S41–6. doi: 10.1086/315537
110. Griskevica A, Ching P, Russo G, Kreysler J. Diphtheria in Latvia, 1986–1996. *J Infect Dis.* (2000) 181(Suppl 1):S60–4. doi: 10.1086/315540
111. Jogiste A, Ching P, Trei T, Kreysler J. Diphtheria in Estonia, 1991–1996. *J Infect Dis.* (2000) 181(Suppl 1):S65–8. doi: 10.1086/315541
112. Nekrassova LS, Chudnaya LM, Marievski VF, Oksiuk VG, Gladkaya E, Bortnitska, II, et al. Epidemic diphtheria in Ukraine, 1991–1997. *J Infect Dis.* (2000) 181(Suppl 1):S35–40. doi: 10.1086/315536
113. Usonis V, Bakasenas V, Morkunas B, Valentelis R, Ching P, Kreysler J. Diphtheria in Lithuania, 1986–1996. *J Infect Dis.* (2000) 181(Suppl 1):S55–9. doi: 10.1086/315539
114. Wanlapakorn N, Yoocharoen P, Tharmaphornpilas P, Theamboonlers A, Poovorawan Y. Diphtheria outbreak in Thailand, (2012) seroprevalence of diphtheria antibodies among Thai adults and its implications for immunization programs. *South Asian J Trop Med Public Health.* (2014) 45:1132–41.
115. Pool V, Tomovici A, Johnson DR, Greenberg DP, Decker MD. Humoral immunity 10years after booster immunization with an adolescent and adult formulation combined tetanus, diphtheria, and 5-component acellular pertussis vaccine in the USA. *Vaccine.* (2018) 36:2282–7. doi: 10.1016/j.vaccine.2018.03.029

116. Golaz A, Hardy IR, Glushkevich TG, Areytchiuk EK, Deforest A, Strebel P, et al. Evaluation of a single dose of diphtheria-tetanus toxoids among adults in Odessa, Ukraine, 1995: immunogenicity and adverse reactions. *J Infect Dis.* (2000) 181(Suppl 1):S203–7. doi: 10.1086/315558
117. Ronne T, Valentelis R, Tarum S, Griskevica A, Wachmann CH, Aggerbeck H, et al. Immune response to diphtheria booster vaccine in the Baltic states. *J Infect Dis.* (2000) 181(Suppl 1):S213–9. doi: 10.1086/315560
118. Sutter RW, Hardy IR, Kozlova IA, Tchoudnaia LM, Gluskevich TG, Marievsky V, et al. Immunogenicity of tetanus-diphtheria toxoids (Td) among Ukrainian adults: implications for diphtheria control in the Newly Independent States of the Former Soviet Union. *J Infect Dis.* (2000) 181(Suppl 1):S197–202. doi: 10.1086/315557
119. Finger F, Funk S, White K, Siddiqui MR, Edmunds WJ, Kucharski AJ. Real-time analysis of the diphtheria outbreak in forcibly displaced Myanmar nationals in Bangladesh. *BMC Med.* (2019) 17:58. doi: 10.1186/s12916-019-1288-7
120. Rahman MR, Islam K. Massive diphtheria outbreak among Rohingya refugees: lessons learnt. *J Travel Med.* (2019) 26:tay141. doi: 10.1093/jtm/tay122
121. Santos LS, Sant'anna LO, Ramos JN, Ladeira EM, Stavracakis-Peixoto R, Borges LL, et al. Diphtheria outbreak in Maranhao, Brazil: microbiological, clinical and epidemiological aspects. *Epidemiol Infect.* (2015) 143:791–8. doi: 10.1017/S0950268814001241
122. Landazabal Garcia N, Burgos Rodriguez MM, Pastor D. Diphtheria outbreak in Cali, Colombia, August-October 2000. *Epidemiol Bull.* (2001) 22:13–5.
123. Saikia L, Nath R, Saikia NJ, Choudhury G, Sarkar M. A diphtheria outbreak in Assam, India. Southeast Asian *J Trop Med Public Health.* (2010) 41:647–52.
124. Das PP, Patgiri SJ, Saikia L, Paul D. Recent outbreaks of diphtheria in Dibrugarh District, Assam, India. *J Clin Diagn Res.* (2016) 10:DR01–3. doi: 10.7860/JCDR/2016/20212.8144
125. Parande MV, Roy S, Mantur BG, Parande AM, Shinde RS. Resurgence of diphtheria in rural areas of North Karnataka, India. *Indian J Med Microbiol.* (2017) 35:247–51. doi: 10.4103/ijmm.IJMM_17_48
126. Hughes GJ, Mikhail AF, Husada D, Irawan E, Kafatos G, Bracebridge S, et al. Seroprevalence and determinants of immunity to diphtheria for children living in two districts of contrasting incidence during an outbreak in east Java, Indonesia. *Pediatr Infect Dis J.* (2015) 34:1152–6. doi: 10.1097/INF.0000000000000846
127. Nanthavong N, Black AP, Nouanthong P, Souvannaso C, Vilivong K, Muller CP, et al. Diphtheria in Lao PDR: insufficient coverage or ineffective vaccine? *PLoS ONE.* (2015) 10:e0121749. doi: 10.1371/journal.pone.0121749
128. Rasmussen I, Wallace S, Mengshoel AT, Hoiby EA, Brandtzaeg P. Diphtheria outbreak in Norway: lessons learned. *Scand J Infect Dis.* (2011) 43:986–9. doi: 10.3109/00365548.2011.600326
129. Besa NC, Coldiron ME, Bakri A, Raji A, Nsuami MJ, Rousseau C, et al. Diphtheria outbreak with high mortality in northeastern Nigeria. *Epidemiol Infect.* (2014) 142:797–802. doi: 10.1017/S0950268813001696
130. Czajka U, Wiatrzyk A, Mosiej E, Forminska K, Zasada AA. Changes in MLST profiles and biotypes of *Corynebacterium diphtheriae* isolates from the diphtheria outbreak period to the period of invasive infections caused by nontoxigenic strains in Poland (1950–2016). *BMC Infect Dis.* (2018) 18:121. doi: 10.1186/s12879-018-3020-1
131. Mahomed S, Archary M, Mutevedzi P, Mahabeer Y, Govender P, Ntshoe G, et al. An isolated outbreak of diphtheria in South Africa, 2015. *Epidemiol Infect.* (2017) 145:2100–8. doi: 10.1017/S0950268817000851
132. Raad I, Chaftari AM, Dib RW, Graviss EA, Hachem R. Emerging outbreaks associated with conflict and failing healthcare systems in the Middle East. *Infect Control Hosp Epidemiol.* (2018) 39:1230–6. doi: 10.1017/ice.2018.177
133. Lodeiro-Colatosti A, Reischl U, Holzmann T, Hernandez-Pereira CE, Risquez A, Paniz-Mondolfi AE. Diphtheria outbreak in amerindian communities, Wonken, Venezuela, 2016–2017. *Emerg Infect Dis.* (2018) 24:1340–4. doi: 10.3201/eid2407.171712
134. Paniz-Mondolfi AE, Tami A, Grillet ME, Marquez M, Hernandez-Villena J, Escalona-Rodriguez MA, et al. Resurgence of vaccine-preventable diseases in venezuela as a regional public health threat in the americas. *Emerg Infect Dis.* (2019) 25:625–32. doi: 10.3201/eid2504.181305
135. Dureab F, Muller O, Jahn A. Resurgence of diphtheria in Yemen due to population movement. *J Travel Med.* (2018) 25:1–2. doi: 10.1093/jtm/tay094
136. Doi Y, Iovleva A, Bonomo RA. The ecology of extended-spectrum β-lactamases (ESBLs) in the developed world. *J Travel Med.* (2017) 24(Suppl 1):S44–51. doi: 10.1093/jtm/taw102
137. Navon-Venezia S, Kondratyeva K, Carattoli A. *Klebsiella pneumoniae*: a major worldwide source and shuttle for antibiotic resistance. *FEMS Microbiol Rev.* (2017) 41:252–75. doi: 10.1093/femsre/fux013
138. Vilches TN, Bonesso MF, Guerra HM, Fortaleza C, Park AW, Ferreira CP. The role of intra and inter-hospital patient transfer in the dissemination of heathcare-associated multidrug-resistant pathogens. *Epidemics.* (2018) 3:362–72. doi: 10.1016/j.epidem.2018.11.001
139. Goodman RA, Buehler JW, Koplan JP. The epidemiologic field investigation: science and judgment in public health practice. *Am J Epidemiol.* (1990) 132:9–16. doi: 10.1093/oxfordjournals.aje.a115647
140. Reingold AL. Outbreak investigations–a perspective. *Emerg Infect Dis.* (1998) 4:21–7. doi: 10.3201/eid0401.980104
141. Bolt F, Cassiday P, Tondella ML, Dezoysa A, Efstratiou A, Sing A, et al. Multilocus sequence typing identifies evidence for recombination and two distinct lineages of *Corynebacterium diphtheriae*. *J Clin Microbiol.* (2010) 48:4177–85. doi: 10.1128/JCM.00274-10
142. Damian M, Grimont F, Narvskaya O, Straut M, Surdeanu M, Cojocaru R, et al. Study of *Corynebacterium diphtheriae* strains isolated in Romania, northwestern Russia and the Republic of Moldova. *Res Microbiol.* (2002) 153:99–106. doi: 10.1016/S0923-2508(01)01294-3
143. De Zoysa A, Hawkey P, Charlett A, Efstratiou A. Comparison of four molecular typing methods for characterization of *Corynebacterium diphtheriae* and determination of transcontinental spread of *C. diphtheriae* based on BstEII rRNA gene profiles. *J Clin Microbiol.* (2008) 46:3626–35. doi: 10.1128/JCM.00300-08
144. Grimont PA, Grimont F, Efstratiou A, De Zoysa A, Mazurova I, Ruckly C, et al. International nomenclature for *Corynebacterium diphtheriae* ribotypes. *Res Microbiol.* (2004) 155:162–6. doi: 10.1016/j.resmic.2003.12.005
145. Kolodkina V, Titov L, Sharapa T, Grimont F, Grimont PA, Efstratiou A. Molecular epidemiology of *C. diphtheriae* strains during different phases of the diphtheria epidemic in Belarus. *BMC Infect Dis.* (2006) 6:129. doi: 10.1186/1471-2334-6-129
146. Mokrousov I, Limeschenko E, Vyazovaya A, Narvskaya O. *Corynebacterium diphtheriae* spoligotyping based on combined use of two CRISPR loci. *Biotechnol J.* (2007) 2:901–6. doi: 10.1002/biot.200700035
147. Mokrousov I, Narvskaya O, Limeschenko E, Vyazovaya A. Efficient discrimination within a *Corynebacterium diphtheriae* epidemic clonal group by a novel macroarray-based method. *J Clin Microbiol.* (2005) 43:1662–8. doi: 10.1128/JCM.43.4.1662-1668.2005
148. Mokrousov I, Vyazovaya A, Kolodkina V, Limeschenko E, Titov L, Narvskaya O. Novel macroarray-based method of *Corynebacterium diphtheriae* genotyping: evaluation in a field study in Belarus. *Eur J Clin Microbiol Infect Dis.* (2009) 28:701–3. doi: 10.1007/s10096-008-0674-4
149. Titov L, Kolodkina V, Dronina A, Grimont F, Grimont PA, Lejay-Collin M, et al. Genotypic and phenotypic characteristics of *Corynebacterium diphtheriae* strains isolated from patients in belarus during an epidemic period. *J Clin Microbiol.* (2003) 41:1285–8. doi: 10.1128/JCM.41.3.1285-1288.2003
150. Mokrousov I. Resolution threshold of current molecular epidemiology of diphtheria. *Emerg Infect Dis.* (2014) 20:1937–8. doi: 10.3201/eid2011.140094
151. Fourel G, Phalipon A, Kaczorek M. Evidence for direct regulation of diphtheria toxin gene transcription by an Fe2+-dependent DNA-binding repressor, DtoxR, in *Corynebacterium diphtheriae*. *Infect Immun.* (1989) 57:3221–5.
152. Sangal V, Burkovski A, Hunt AC, Edwards B, Blom J, Hoskisson PA. A lack of genetic basis for biovar differentiation in clinically important *Corynebacterium diphtheriae* from whole genome sequencing. *Infect Genet Evol.* (2013) 21:54–7. doi: 10.1016/j.meegid.2013.10.019
153. Stucki D, Ballif M, Egger M, Furrer H, Altpeter E, Battegay M, et al. Standard genotyping overestimates transmission of *Mycobacterium tuberculosis* among immigrants in a low-incidence country. *J Clin Microbiol.* (2016) 54:1862–70. doi: 10.1128/JCM.00126-16

154. Cerdeno-Tarraga AM, Efstratiou A, Dover LG, Holden MT, Pallen M, Bentley SD, et al. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res.* (2003) 31:6516–23. doi: 10.1093/nar/gkg874
155. Trost E, Blom J, Soares Sde C, Huang IH, Al-Dilaimi A, Schroder J, et al. Pangenomic study of *Corynebacterium diphtheriae* that provides insights into the genomic diversity of pathogenic isolates from cases of classical diphtheria, endocarditis, and pneumonia. *J Bacteriol.* (2012) 194:3199–215. doi: 10.1128/JB.00183-12
156. Sangal V, Blom J, Sutcliffe IC, von Hunolstein C, Burkovski A, Hoskisson PA. Adherence and invasive properties of *Corynebacterium diphtheriae* strains correlates with the predicted membrane-associated and secreted proteome. *BMC Genom.* (2015) 16:765. doi: 10.1186/s12864-015-1980-8
157. Arnold JW, Koudelka GB. The Trojan Horse of the microbiological arms race: phage-encoded toxins as a defence against eukaryotic predators. *Environ Microbiol.* (2014) 16:454–66. doi: 10.1111/1462-2920.12232
158. Mokrousov I. *Corynebacterium diphtheriae*: genome diversity, population structure and genotyping perspectives. *Infect Genet Evol.* (2009) 9:1–15. doi: 10.1016/j.meegid.2008.09.011
159. Grosse-Kock S, Kolodkina V, Schwalbe EC, Blom J, Burkovski A, Hoskisson PA, et al. Genomic analysis of endemic clones of toxigenic and non-toxigenic *Corynebacterium diphtheriae* in Belarus during and after the major epidemic in 1990s. *BMC Genom.* (2017) 18:873. doi: 10.1186/s12864-017-4276-3
160. Chen J, Ram G, Penades JR, Brown S, Novick RP. Pathogenicity island-directed transfer of unlinked chromosomal virulence genes. *Mol Cell.* (2015) 57:138–49. doi: 10.1016/j.molcel.2014.11.011
161. Joseph B, Schwarz RF, Linke B, Blom J, Becker A, Claus H, et al. Virulence evolution of the human pathogen *Neisseria meningitidis* by recombination in the core and accessory genome. *PLoS ONE.* (2011) 6:e18441. doi: 10.1371/journal.pone.0018441
162. Sabharwal V, Stevenson A, Figueira M, Orthopoulos G, Trzcinski K, Pelton SI. Capsular switching as a strategy to increase pneumococcal virulence in experimental otitis media model. *Microbes Infect.* (2014) 16:292–9. doi: 10.1016/j.micinf.2013.12.002
163. Marks LR, Reddinger RM, Hakansson AP. High levels of genetic recombination during nasopharyngeal carriage and biofilm formation in *Streptococcus pneumoniae*. *MBio.* (2012) 3:e00200–12. doi: 10.1128/mBio.00200-12
164. Sangal V, Tucker NP, Burkovski A, Hoskisson PA. Draft genome sequence of *Corynebacterium diphtheriae* biovar intermedius NCTC 5011. *J Bacteriol.* (2012) 194:4738. doi: 10.1128/JB.00939-12
165. Sangal V, Tucker NP, Burkovski A, Hoskisson PA. The draft genome sequence of *Corynebacterium diphtheriae* bv. mitis NCTC 3529 reveals significant diversity between the primary disease-causing biovars. *J Bacteriol.* (2012) 194:3269. doi: 10.1128/JB.00503-12
166. Hong KW, Asmah Hani AW, Nurul Aina Murni CA, Pusparani RR, Chong CK, Verasahib K, et al. Comparative genomic and phylogenetic analysis of a toxigenic clinical isolate of *Corynebacterium diphtheriae* strain B-D-16–78 from Malaysia. *Infect Genet Evol.* (2017) 54:263–70. doi: 10.1016/j.meegid.2017.07.015
167. Timms VJ, Nguyen T, Crighton T, Yuen M, Sintchenko V. Genome-wide comparison of *Corynebacterium diphtheriae* isolates from Australia identifies differences in the Pan-genomes between respiratory and cutaneous strains. *BMC Genom.* (2018) 19:869. doi: 10.1186/s12864-018-5147-2
168. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computat Biol.* (2017) 13:e1005595. doi: 10.1371/journal.pcbi.1005595
169. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics*. (2017) doi: 10.1093/bioinformatics/btx610. [Epub ahead of print].
170. Lawpoolsri S, Kaewkungwal J, Khamsiriwatchara A, Sovann L, Sreng B, Phommasack B, et al. Data quality and timeliness of outbreak reporting system among countries in Greater Mekong subregion: challenges for international data sharing. *PLoS Negl Trop Dis.* (2018) 12:e0006425. doi: 10.1371/journal.pntd.0006425
171. WHO. *WHO-Recommended Surveillance Standard Of Diphtheria* (2014).
172. Neal S, Efstratiou A. DIPNET - establishment of a dedicated surveillance network for diphtheria in Europe. *Euro surveillance*. (2007) 12:E9–10. doi: 10.2807/esm.12.12.00754-en
173. Egli A, Blanc DS, Greub G, Keller PM, Lazarevic V, Lebrand A, et al. Improving the quality and workflow of bacterial genome sequencing and analysis: paving the way for a Switzerland-wide molecular epidemiological surveillance platform. *Swiss Med Wkly.* (2018) 148:w14693. doi: 10.4414/smw.2018.14693
174. Choi J, Cho Y, Shim E, Woo H. Web-based infectious disease surveillance systems and public health perspectives: a systematic review. *BMC Public Health.* (2016) 16:1238. doi: 10.1186/s12889-016-3893-0
175. Jain R, Sontisirikit S, Iamsirithaworn S, Prendinger H. Prediction of dengue outbreaks based on disease surveillance, meteorological and socio-economic data. *BMC Infect Dis*. (2019) 19:272. doi: 10.1186/s12879-019-3874-x
176. Njage PMK, Leekitcharoenphon P, Hald T. Improving hazard characterization in microbial risk assessment using next generation sequencing data and machine learning: predicting clinical outcomes in shigatoxigenic *Escherichia coli*. *Int J Food Microbiol.* (2019) 292:72–82. doi: 10.1016/j.ijfoodmicro.2018.11.016
177. Power RA, Parkhill J, de Oliveira T. Microbial genome-wide association studies: lessons from human GWAS. *Nat Rev Genet*. (2017) 18:41–50. doi: 10.1038/nrg.2016.132
178. Coll F, Phelan J, Hill-Cawthorne GA, Nair MB, Mallard K, Ali S, et al. Genome-wide analysis of multi- and extensively drug-resistant *Mycobacterium tuberculosis*. *Nat Genet.* (2018) 50:307–16.
179. Park J, Zhang Y, Buboltz AM, Zhang X, Schuster SC, Ahuja U, et al. Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genom.* (2012) 13:545. doi: 10.1186/1471-2164-13-545
180. Yahara K, Meric G, Taylor AJ, de Vries SP, Murray S, Pascoe B, et al. Genome-wide association of functional traits linked with *Campylobacter jejuni* survival from farm to fork. *Environ Microbiol.* (2017) 19:361–80. doi: 10.1111/1462-2920.13628

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

ORIGINAL RESEARCH
published: 27 August 2019
doi: 10.3389/fpubh.2019.00241

Evaluation of Rapid Library Preparation Protocols for Whole Genome Sequencing Based Outbreak Investigation

Helena M. B. Seth-Smith[1,2,3]*, *Ferdinando Bonfiglio*[2,4], *Aline Cuénod*[1,2], *Josiane Reist*[2], *Adrian Egli*[1,2] *and Daniel Wüthrich*[1,2,3]

[1] Division of Clinical Bacteriology and Mycology, University Hospital Basel, Basel, Switzerland, [2] Applied Microbiology Research, Department of Biomedicine, University of Basel, Basel, Switzerland, [3] DBM Bioinformatics Core Facility, SIB Swiss Institute of Bioinformatics, Basel, Switzerland, [4] Personalized Health Basel, University of Basel, Basel, Switzerland

Whole genome sequencing (WGS) has become the new gold standard for bacterial outbreak investigation, due to the high resolution available for typing. While sequencing is currently predominantly performed on Illumina devices, the preceding library preparation can be performed using various protocols. Enzymatic fragmentation library preparation protocols are fast, have minimal hands-on time, and work with small quantities of DNA. The aim of our study was to compare three library preparation protocols for molecular typing: Nextera XT (Illumina); Nextera Flex (Illumina); and QIAseq FX (Qiagen). We selected 12 ATCC strains from human Gram-positive and Gram-negative pathogens with %G+C-content ranging from 27% (*Fusobacterium nucleatum*) to 73% (*Micrococcus luteus*), each having a high quality complete genome assembly available, to allow in-depth analysis of the resulting Illumina sequence data quality. Additionally, we selected isolates from previously analyzed cases of vancomycin-resistant *Enterococcus faecium* (VRE) ($n = 7$) and a local outbreak of *Klebsiella aerogenes* ($n = 5$). The number of protocol steps and time required were compared, in order to test the suitability for routine laboratory work. Data analyses were performed with standard tools commonly used in outbreak situations: Ridom SeqSphere+ for cgMLST; CLC genomics workbench for SNP analysis; and open source programs. Nextera Flex and QIAseq FX were found to be less sensitive than Nextera XT to variable %G+C-content, resulting in an almost uniform distribution of read-depth. Therefore, low coverage regions are reduced to a minimum resulting in a more complete representation of the genome. Thus, with these two protocols, more alleles were detected in the cgMLST analysis, producing a higher resolution of closely related isolates. Furthermore, they result in a more complete representation of accessory genes. In particular, the high data quality and relative simplicity of the workflow of Nextera Flex stood out in this comparison. This thorough comparison within an ISO/IEC 17025 accredited environment will be of interest to those aiming to optimize their clinical microbiological genome sequencing.

Keywords: NGS, next generation sequencing, library, Illumina, whole genome sequencing, comparison, bacteria, prokaryotes

OPEN ACCESS

Edited by:
Vitali Sintchenko, University of Sydney, Australia

Reviewed by:
Rebecca Rockett, University of Sydney, Australia
Avram Levy, University of Western Australia, Australia

***Correspondence:**
Helena M. B. Seth-Smith
helena.seth-smith@usb.ch

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *09 April 2019*
Accepted: *12 August 2019*
Published: *27 August 2019*

Citation:
Seth-Smith HMB, Bonfiglio F, Cuénod A, Reist J, Egli A and Wüthrich D (2019) Evaluation of Rapid Library Preparation Protocols for Whole Genome Sequencing Based Outbreak Investigation. Front. Public Health 7:241. doi: 10.3389/fpubh.2019.00241

INTRODUCTION

Whole genome sequences currently provide the highest resolution for typing bacterial pathogens. The implementation of next generation sequencing (NGS) in routine clinical microbiology laboratories provides the foundation to analyze bacteria with high resolution, reproducibility and accuracy. Decreasing costs and increasing ease of implementation through increasingly flexible platform options, means that more laboratories will seek this technology over time.

Whole genome sequencing (WGS) has shown its value in molecular epidemiology, from seminal papers on MRSA and *Mycobacterium tuberculosis* helping to trace and resolve epidemics (1, 2), to implementation in routine laboratories (3–5), and local molecular epidemiological studies (6, 7). Methods of analysis range from determination of multi-locus sequence type (MLST; low resolution) through core genome MLST (cgMLST; high resolution) to whole genome phylogenies based on single nucleotide polymorphisms (SNPs; highest resolution). Using WGS in outbreak detection ideally takes account of all mutations and genomic variability in order to fully resolve outbreak scenarios and transmission chains (5, 8–11). Factors encoded within the genomes, such as antimicrobial resistance (AMR) and virulence factors, can also be determined from good quality assemblies (3–5, 12). Quality assurance, backward compatibility, communication between experts in different fields, and reporting to clinicians are issues currently being addressed (13–17).

Behind all these analyzes lies the all-important data. Several technologies have been used over the past decade for WGS: Ion Torrent PGM, Roche 454, PacBio and most recently Oxford Nanopore Technologies. But it is predominantly data from Illumina machines, from the MiniSeq, MiSeq, NextSeq, or HiSeq platforms, that is used for molecular epidemiology or bacterial genomics, as evidenced by the vast amounts of Illumina data deposited in databases (>90% at the Short Read Archive). Prior to the sequencing step, DNA libraries need to be made, protocols for which can vary greatly. Given the relatively high cost of library preparation compared to sequencing, and the time required to perform it, library preparation is a critical and rate-limiting step. Although many aspects of WGS can be optimized for routine diagnostic microbiology (17), to date few studies have addressed the data quality produced by different library methods.

Mechanical shearing of DNA often offers the most even and controllable DNA fragmentation (18), but requires high amounts of input DNA and hands-on time. Automation of mechanical shearing is problematic, limiting throughput. The most popular and implementable library protocols use proprietary transposases to cleave the DNA and ligate the adapters in one step, a method which is rapid but dependent on the DNA/enzyme concentration ratio, and is subject to sequence bias. The impact of this bias on the %G+C rich *Mycobacterium tuberculosis* genome has been explored, and the TruSeq (Illumina) method, involving mechanical shearing of DNA, was found to be superior to the enzymatic Nextera XT (Illumina) (19). On the AT-rich *Plasmodium falciparum* genome, Nextera was again found to give highly biased results (20). This phenomenon has also been observed in human leukocyte antigen (HLA) genotyping (21).

With QIAseq FX, Qiagen have recently released a library preparation protocol that is based on fully enzymatic fragmentation (nuclease). The advantage of this approach is that the efficiency of the fragmentation is not as strongly affected by %G+C-content as the transposase from the Nextera XT approach. As QIAseq FX uses only an enzyme and not a whole complex, the adaptor ligation must then be applied in a separate step (QIAseq FX DNA Library Handbook). Another recent launch, Nextera Flex (Illumina) is also a transposome based library preparation kit, promising consistent yield and fragment size, and less sequence bias (22). The development over Nextera XT involves bead-conjugated transposomes, meaning that the tagmentation sites are positionally better defined by the DNA binding to the beads.

The costs of the different compared kits are quite similar, and up-to-date prices are listed on the manufacturer's websites. Currently, the difference across all is <20%. Some laboratories implement protocols using lower reagent volumes to reduce the per sample costs, however this study used the manufacturer's standard protocols.

Our aim was to compare the data quality from three commercial library preparation kits, for use in clinical routine microbiology WGS. The optimal protocol is rapid, performs consistently across all genome types without optimization, and produces high quality data for both rapid and reliable outbreak analysis and AMR gene detection.

MATERIALS AND METHODS

Strain Selection

In order to evaluate the usability of the different library preparation kits, we made a selection of 12 ATCC strains representing Gram-positive and negative pathogenic bacterial species, with a high range of %G+C-content (**Table 1**). A complete high-quality reference genome exists for each strain. Additionally, we included seven local patient isolates of *Enterococcus faecium* and five isolates from a *Klebsiella aerogenes* outbreak from 2018.

DNA Extraction and Sequencing

All work was performed in an ISO/IEC 17025 accredited environment, although only the Nextera XT protocol is currently accredited. DNA from all isolates was extracted by Qiagen EZ1 (Qiagen, Hilden, Germany) using the DNeasy blood and tissue kit (Qiagen), from a single colony. Prior to this, some isolate were subject to pretreatment: *Mycobacterium tuberculosis* was inactivated at 95°C for 1 h and disrupted in a TissueLyser (Qiagen) for 2 min at highest frequency; *Streptococcus pyogenes* was pre-treated with the TissueLyser (Qiagen) for 2 min at frequency 30; *Staphylococcus* were pre-treated with lysozyme und lysostaphin for 30 min at 37°C; all other bacteria were pre-treated using Proteinase K for 10 min at 56°C. Extracts were quantified by Qubit (Invitrogen), separated into three aliquots, and frozen at 20°C.

TABLE 1 | List of sequenced isolates, characteristics, reference genomes, and sample accessions.

Unique name	Species	DNA extraction conczentration (ng/μl)	Reference used	%G+C- content reference	Reference sequence accession	Number of reads produced			Sample accession
						XT	Flex	Qia	
ATCC25586	*Fusobacterium nucleatum*	36.4	ATCC25586	27.15	NC_003454.1	1,16,59,182	54,71,621	38,92,304	ERS3207828 (SAMEA5402510)
ATCC700819	*Campylobacter jejuni*	34.2	ATCC700819	30.55	NC_002163.1	51,04,723	80,67,749	5,37,051	ERS3207833 (SAMEA5402515)
ATCC25923	*Staphylococcus aureus*	88.4	ATCC25923	32.86	NZ_CP009361.1, NZ_CP009362.1	90,93,138	57,42,025	71,39,563	ERS3207824 (SAMEA5402506)
ATCC29212	*Enterococcus faecalis*	39.8	ATCC29212	37.35	NZ_CP008816.1, NZ_CP008815.1, NZ_CP008814.1	71,99,132	68,06,105	69,81,047	ERS3207826 (SAMEA5402508)
ATCC19615	*Streptococcus pyogenes*	20.8	ATCC19615	38.48	NZ_CP008926.1	78,95,584	60,46,735	94,81,835	ERS3207823 (SAMEA5402505)
ATCC25845	*Prevotella melaninogenica*	92.0	ATCC25845	40.98	NC_014370.1, NC_014371.1	69,93,760	21,62,813	52,57,867	ERS3207831 (SAMEA5402513)
ATCC25922	*Escherichia coli*	27.2	ATCC25922	50.37	CP009072.1	64,19,681	53,21,879	61,12,711	ERS3207827 (SAMEA5402509)
ATCC700603	*Klebsiella quasipneumoniae*	42.8	ATCC700603	57.73	NZ_CP014696.2, NZ_CP014697.2, NZ_CP014698.2	48,25,887	58,53,388	84,17,937	ERS3207829 (SAMEA5402511)
ATCC25177 (H37Ra)	*Mycobacterium tuberculosis*	1.2	ATCC25177	65.61	NC_009525.1	47,94,204	96,95,720	2,54,69,645	ERS3207832 (SAMEA5402514)
ATCC27853	*Pseudomonas aeruginosa*	42.4	ATCC27853	66.08	CP015117.1	45,32,729	48,88,025	68,12,269	ERS3207825 (SAMEA5402507)
ATCCBAA-67	*Burkholderia stabilis*	72.0	ATCCBAA-67	66.42	NZ_CP016442.1, NZ_CP016443.1, NZ_CP016444.1	87,99,551	55,77,758	63,87,296	ERS3207822 (SAMEA5402504)
ATCC4698	*Micrococcus luteus*	45.6	ATCC4698	73.00	CP001628.1	50,81,588	93,96,130	85,84,319	ERS3207830 (SAMEA5402512)
NMB004374	*Enterococcus faecium*	55.8	Aus0004	37.80	NC_017022.1	53,87,832	52,12,078	77,60,492	ERS3207811 (SAMEA5402493)
NMB004375	*Enterococcus faecium*	55.8	Aus0004	37.80	NC_017022.1	52,85,502	44,62,856	55,05,430	ERS3207812 (SAMEA5402494)
NMB004376	*Enterococcus faecium*	55.4	Aus0004	37.80	NC_017022.1	49,36,762	28,48,407	88,145	ERS3207813 (SAMEA5402495)
NMB003061	*Enterococcus faecium*	56.2	Aus0004	37.80	NC_017022.1	41,98,651	52,13,009	84,72,370	ERS3207814 (SAMEA5402496)
NMB003076	*Enterococcus faecium*	47.2	Aus0004	37.80	NC_017022.1	61,97,648	64,57,841	75,28,868	ERS3207815 (SAMEA5402497)

(Continued)

TABLE 1 | Continued

Unique name	Species	DNA extraction conczentration (ng/µl)	Reference used	%G+C- content reference	Reference sequence accession	Number of reads produced			Sample accession
						XT	Flex	Qia	
NMB003240	*Enterococcus faecium*	57.6	Aus0004	37.80	NC_017022.1	71,97,873	75,30,750	81,14,687	ERS3207816 (SAMEA5402498)
NMB003062 (VRECH001)	*Enterococcus faecium*	40.6	Aus0004	37.80	NC_017022.1	51,80,248	77,99,226	76,15,044	ERS2595418 (SAMEA4775467)
NMB004427	*Klebsiella aerogenes*	38.6	KCTC2190	55.00	NC_015663.1	62,65,626	34,05,502	68,37,549	ERS3207817 (SAMEA5402499)
NMB004428	*Klebsiella aerogenes*	25	KCTC2190	55.00	NC_015663.1	9,64,566	78,95,179	1,09,31,859	ERS3207818 (SAMEA5402500)
NMB004429	*Klebsiella aerogenes*	29	KCTC2190	55.00	NC_015663.1	4,27,144	67,01,537	60,70,575	ERS3207819 (SAMEA5402501)
NMB004430	*Klebsiella aerogenes*	24.8	KCTC2190	55.00	NC_015663.1	33,39,975	38,92,829	69,09,301	ERS3207820 (SAMEA5402502)
NMB004431	*Klebsiella aerogenes*	28.2	KCTC2190	55.00	NC_015663.1	46,70,831	86,19,114	56,84,285	ERS3207821 (SAMEA5402503)

Libraries were created from the aliquots using Nextera XT ("XT"; Illumina), Nextera DNA Flex ("flex"; Illumina) or QIAseq FX ("Qia"; Qiagen). The recommended amounts and concentrations of DNA for each protocol were used where possible (1 ng for XT, 100 ng for Flex, 200 ng for Qia). To simulate a more realistic situation for *M. tuberculosis*, for which DNA extraction is not trivial, we used less DNA for the three kits (1 ng for XT, 10 ng for flex, 10 ng for Qia).

Each pool of libraries was loaded and sequenced separately on a NextSeq 500 device (cluster densities: XT 202, flex 189, Qia 244 K/mm^2) and were sequenced using 2 × 151 bp paired end reads, within the Division of Clinical Microbiology, University Hospital Basel. The data was demultiplexed using bcl2fastq (version v2.17.1.14; Illumina).

Genomic Data Quality Analysis

Reads were trimmed using trimmomatic (version 0.38) (23) using default parameters (ILLUMINACLIP:2:30:10 SLIDINGWINDOW:4:15 MINLEN:125), and randomly subsampled using seqtk (version 1.3-r106, -s100; https://github.com/lh3/seqtk) to provide mean 10, 20, 50, 100, and 200-fold coverage of the genomes.

Assemblies were produced by unicycler (v0.3.0b) (24), with assembly parameters derived using QUAST (version 5.0.2) (25). The annotation was performed using Prokka (version 1.13) (26). AMR genes were predicted by using ABRicate (version 0.8.10; https://github.com/tseemann/abricate) with the NCBI database (accession: PRJNA313047).

Reads from ATCC strains were mapped using BWA (version 0.7.17) (27) against the complete references with all replicons concatenated (**Table 1**). The read depth at the different positions was determined using pilon (version 1.23). The insert size was calculated from the sam files using an in-house python script (https://github.com/danielwuethrich87/collection/blob/master/scripts/parse_sam_for_insertsize.py). The base-composition at the difference positions within the reads was calculated using FastQC (version 0.11.5; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) on the mapped 10-fold subsampled reads from ATCC25586.

K-mer signatures of sub-sampled reads and corresponding reference genomes were computed with Sourmash v2.0.0 (28) using the suggested MinHash resolution (1000:1 compression ratio) and a k-mer size of 31. The k-mer signature of the subsampled assemblies was assessed with the Jaccard distance metric, which is calculated by asking how many k-mers are shared between two samples vs. how many k-mers in total are in the combined samples [(Sample1 ∩ Sample2)/(Sample1 ∪ Sample2)]. A Jaccard distance of 1 means the samples are identical; a Jaccard distance of 0 means the samples are completely different. Overlaps with reference genome were also calculated in terms of containment [(Sample1 ∩ Sample2)/Sample1].

The orthologous groups were determined using the stand-alone Roary pipeline v3.12.0 (29), which takes annotated assemblies in GFF3 format produced by Prokka as above. Principal component analysis (PCA) was performed on the output table of gene presence/absence and the coordinates of the

first two principal components (weighted by the proportion of variance explained) were used to calculate the distance of each sample from the reference as a metric to determine the similarity in terms of gene content.

Outbreak Analysis

For outbreak isolate genomes, data was analyzed in Ridom SeqSphere+ v4.1.6 for *Enterococcus faecium* cgMLST (30), and *Klebsiella aerogenes* cgMLST using an *ad-hoc* scheme comprising 3282 target loci based on the KCTC2190 genome (NC_015663.1) and 41 additional genomes from NCBI. Additionally, MentaLiST (version 1.0.0) (31) was used to identify the cgMLST alleles from the *Enterococcus faecium* isolates.

CLC Genomics Workbench 10.1 was used to generate Single Nucleotide Polymorphism (SNP) phylogenies. Mapping was performed using default parameters, variant calling used the parameters: 10x min coverage, 10 min count and 70% min frequency. SNP trees used a neighbor joining method: minimum coverage 10, minimum coverage 10%, minimum z-score 1.96, multi-nucleotide variants included. The mapping reference for the *Klebsiella aerogenes* outbreak was that of KCTC2190, accession number CP002824.

The *Enterococcus faecium* data was also analyzed using snippy (version 4.3.6, –minfrac 0.8; https://github.com/tseemann/snippy) for SNP calling comparing to the Aus0004 as reference (accession number NC_017022.1) For the phylogenetic analysis, only the core genome SNPs were used. The phylogenetic tree was calculated using the neighbor joining tree algorithm of the scikit-bio (version 0.2.0) package (http://scikit-bio.org/).

RESULTS

Library Preparation and Ease of Use in Routine Laboratories

We selected three different rapid library preparation kits, all of which are based on enzymatic fragmentation: Nextera XT ("XT"), Nextera DNA Flex ("flex"), and QIAseq FX DNA ("Qia") as they each provide a complete solution kit. The required DNA input amount of the three kits is very variable: XT needs exactly 1 ng of input DNA; Qia and flex support a wide range of DNA inputs that can affect the library preparation. The insert size of the Qia kit can be controlled by adjusting the fragmentation time and DNA input amount (1–1,000 ng). Flex accepts a wide range of input DNA (1–500 ng) resulting in the same insert size (300–350 bp). However, in Qia and flex, different DNA input amounts require the number of cycles in the PCR amplification step to be adjusted. Qia also supports a PCR free protocol if more than 100 ng are applied. We decided to use 1 ng of input DNA for XT, 100 ng for flex and 200 ng for Qia. This amount of DNA is reliably produced by our routine DNA extraction techniques, and simplifies the Qia protocol through elimination of PCR. For *M. tuberculosis* we used only 10 ng for Qia and flex. For the Qia protocol we were aiming for a fragment peak size of 550 bp by using 6 min fragmentation time for 100 ng and 10 min for the 10 ng input. The other kits do not allow specific adjustments for fragment length in the standard protocol. Each of the final libraries using XT and Qia were quantified and the 24 samples were equimolarly pooled. As sample normalization is already included in flex, we pooled the samples by taking the same volume from each library.

The application of the three kits revealed their strengths and weaknesses in the laboratory. For routine work, time is of course a major factor. The provider of all three kits state that the library preparation takes 2.5 h. However, we were only able to reach this time with XT, and only if time taken for DNA quantification before and after is not included. We also have to mention that the XT protocol has been established in our laboratory for 3 years and therefore the technicians are highly experienced. The Qia and flex protocols both took ∼4 h. The Qia kit requires long fragmentation time (∼60 min) and ligation time (45 min). It also has to be considered that if the input DNA amount of Qia is below 100 ng, PCR and clean-up must be included, which adds a further 90 min. For flex, the resuspension of the beads with the transposomes requires optimization, as they stick to the walls of PCR plates. Saving hands-on time, especially with larger sample numbers, flex includes bead-based concentration normalization. Also of importance in routine laboratory work, the Illumina kits provide plenty of consumable, which allows for potential inaccurate pipetting and still allows the indicated number of samples to be processed. In contrast, the Qia fragmentation mixture volume delivered in the kit was too limited and resulted in the sequencing failure of one sample (NMB004375).

Taken together, XT has the most convenient protocol to use in the laboratory. However, flex provides some features that allow a very streamlined process. Even though the flex protocol takes longer than XT, the wide range of DNA input amount and the normalized output can lead to a significant time gain. The Qia protocol take also longer than the XT protocol and needs more adjustments according to the DNA input amounts. On the other hand, it offers the easy adjustment of insert sizes.

Genome Coverage Evaluation of ATCC Strains

As a first estimate for the quality of the sequencing we mapped the reads of the ATCC strains against their published reference genomes. For this purpose, we aligned the 100-fold subsampled reads from each sample to the reference and visualized the read depth distribution (**Figure 1A**). The read depth is most variable using XT. This is especially obvious in genomes with low %G+C-content, resulting in many genomic regions with low coverage. Qia and flex, on the other hand, show a more even distribution of read depth in all samples and therefore provide a more complete representation of the genomes. The unevenness of coverage is less pronounced in genomes with G+C-content of 40% or more: these show similar pictures with XT, flex and Qia.

Based on the alignments of the reads to the reference genomes we calculated the insert sizes of the different library preparation kits (**Figure 1B**). With XT we see a clear trend that the genomes with higher %G+C-content have larger insert sizes, showing again that this method is highly sensitive to %G+C-content. The insert sizes of the flex and Qia are stable across the different genomes, with the exception of the low input *Mycobacterium tuberculosis*, and seem unaffected by %G+C-content. Using flex, the insert sizes are well above 300 bp, which allows an optimal use

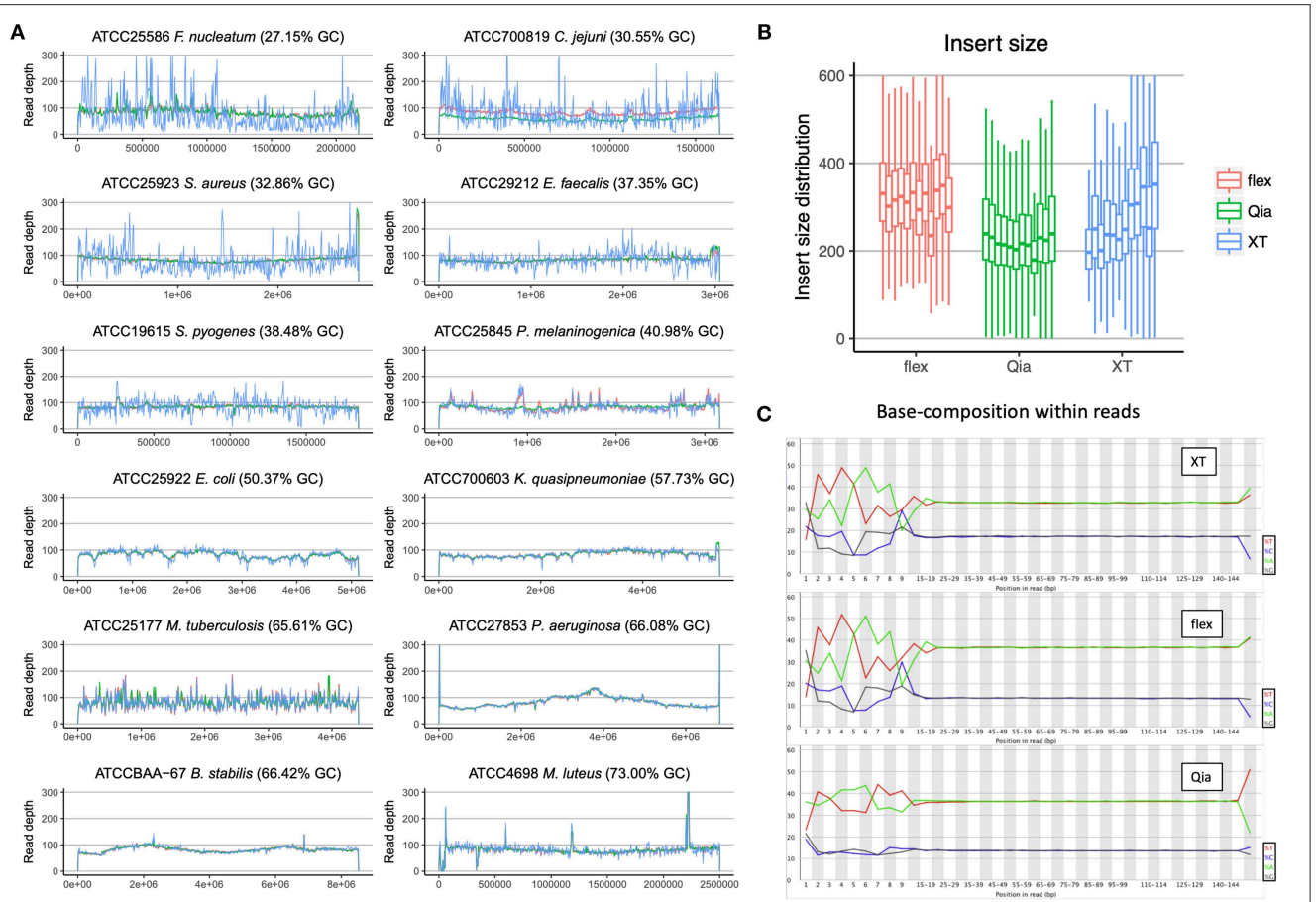


FIGURE 1 | Quality assessment of WGS data. **(A)** The reads of the three library kits subsampled to 100-fold were mapped against the 12 reference genomes and the read depth called was measured. The colors indicate the different library preparation kits. The x-axis reflects the position along the genomes and the y-axis the read depth. **(B)** The insert size of the different libraries was calculated using the alignment of the paired-end reads to the reference. The boxplots represent the calculations from the different species, with the lowest %G+C-content on the left, and the highest on the right. In the boxplots the lower and upper hinges correspond to the first and third quartiles. The whiskers are located at 1.5x of the interquartile range. **(C)** The base composition of all the nucleotide sites in the reads was determined. The bases on the left side show the composition around the fragmentation site.

of 151 × 151 paired-end reads. With Qia we have an insert size slightly above 200 bp, despite having aimed for 400 bp (550 bp fragment size). This value should be able to be adjusted through in-depth protocol optimization.

Looking at %G+C-content variation within the reads (**Figure 1C**), overall the reads produced by Qia and flex are closer to the actual genomic %G+C-content than those from XT. Focusing on the beginning of the reads, which represent the fragmentation sites, flex and XT give a strong variation of the %G+C-content, which is characteristic for the transposome used by Nextera. Surprisingly this fragmentation preference does not affect the read depth distribution of flex. Using Qia we see that the beginning of the reads are very similar to the %G+C-content of the genomes.

Evaluation of Assembly Quality

In order to study the genome representation in the different library preparation kits, we analyzed the subsampled reads of the ATCC strains at mean 10, 20, 50, 100, and 200-fold coverage. K-mer containment was used to compare the k-mers in the reads of the difference subsamples against the k-mers in the reference assemblies (**Figure 2A**). With this analysis we found that, using Qia and flex with an average read depth of 10-fold, more than 99% of all k-mers were found in most of the genomes. At 50-fold with these two kits, k-mers were already completely saturated, indicating that all the genome is represented. XT shows a different picture: while increasing read depth increases the percentage of k-mers found, the k-mer pool of the reference is not completely represented using XT even with 100 and 200-fold coverage, indicating that there will always be regions absent, leading to incomplete representation of the genome.

De novo assembly of the subsampled reads was performed, and the k-mers of the assemblies compared to those from the references using the Jaccard index (**Figure 2B**). This analysis shows a similar picture. Qia and flex show a good representation of the genome with a 50-fold coverage upwards, whereas using XT with 100-fold and 200-fold coverage, reads do not completely represent the genome.

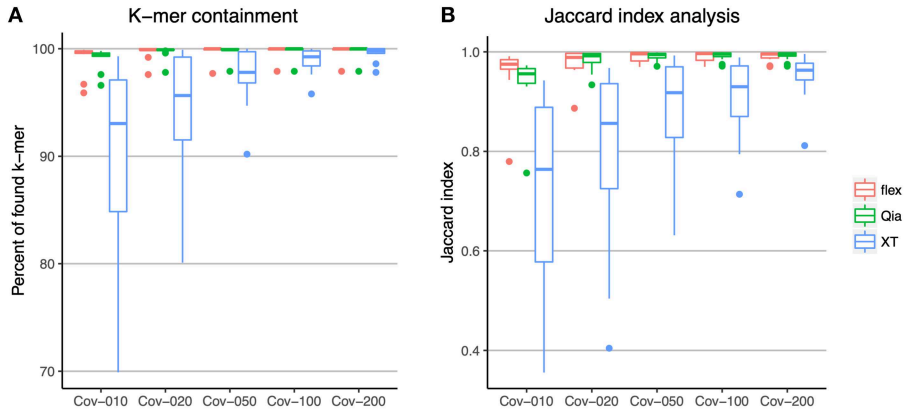


FIGURE 2 | Comparison of the sequencing content using k-mers. **(A)** All k-mers identified within the reads were compared to those k-mer from the reference genomes. The x-axis shows the different subsampling of the reads and the y-axis shows the percent of k-mers that were found in the reads. **(B)** The assemblies of the sequenced strains were compared against the reference assemblies using the Jaccard index of the k-mers. The x-axis shows the different subsampling of the reads used for each assembly. The y-axis shows the Jaccard index. The colors indicate the different library preparation kits. In the boxplots the lower and upper hinges correspond to the first and third quartiles. The whiskers are located at 1.5x of the interquartile range.

Assembly quality measures (NG50, number of contigs, genome representation, mismatches) were calculated using Quast (**Figure S1**). With increasing coverage, contig length (NG50) increases, as does genome fraction compared against the reference genomes, the number of contigs in the assembly decreases, and so do the number of mismatches called between the assemblies and the references. This analysis again shows that we can obtain an almost complete representation of the genome with 50-fold coverage using Qia and flex; XT on the other hand needs 100-fold or more coverage. In order to compare the gene content of assemblies from the different library preparation kits and subsamples, we performed a PCA on the presence and absence of orthologous groups (**Figure S2**). In general, we found that low coverage assemblies (10- and 20-fold) are more likely to result in different gene content (less genes) to the references, which cluster with the high coverage assemblies. However, we also found that in strains with low %G+C-content (≤50.37%; ATCC25586, ATCC700819, ATCC25923, ATCC29212, ATCC19615, ATCC25845, and ATCC25922) the genes found in the XT assemblies, even at high coverage, are separated from the references, Qia and flex assemblies.

AMR genes were analyzed in the published complete ATCC reference genomes and the assemblies from our experiment, using ABRicate (**Table 2**). We found that we can find every resistance gene from flex and Qia reads if the coverage is 50-fold or over. With XT many genes are not found with a coverage of 50-fold and some genes are even absent from the assemblies produced from a coverage of 100 or 200-fold.

Estimation of Coverage Required for cgMLST Analysis

In 2018, we sequenced, as routine, several cases of vancomycin-resistant *Enterococcus faecium* (VRE) and a small outbreak of *Klebsiella aerogenes* (*K. aerogenes*) that was not associated with our hospital. For this study, we selected five *K. aerogenes* isolates and seven VRE isolates to evaluate the performance of the three kits on samples from the routine clinical microbiology laboratory.

After subsampling, we typed the seven VRE strains using the cgMLST scheme of *Enterococcus faecium*, in the commercial software Ridom SeqSphere+, and using the open source software MentaLiST. In order to determine the resolution, we compared the number of core genes found in each sample and subsample (**Figures 3A,B**). Using reads from Qia and flex libraries, most of the core genes are found with a 50-fold coverage and over. With XT, over 25% of the core genes are not identified using a coverage of 50-fold, and 10–20% are still missing at 100-fold coverage.

For the five *K. aerogenes* we created an *ad-hoc* cgMLST scheme using Ridom SeqSphere+. In comparison to the VRE, 50-fold coverage was sufficient for all three kits to assign alleles to over 85% of core genes (**Figure 3C**). As *K. aerogenes* has a higher %G+C-content than *E. faecium*, we have seen that this results in more equal genome coverage from all kits, especially XT, leading to better assemblies and increased core gene identification.

Analysis of Vancomycin-Resistant *Enterococcus faecium* Isolates

A previous investigation showed an outbreak of VRE from Switzerland carried the same MLST type (ST796) as an outbreak in Australia (32–34). Out of this investigation we selected four isolates from an outbreak, as well as three isolates (ST117) from an example of in-patient acquisition of a vancomycin-resistance carrying transposon (Tn1549) in the same strain background. To test the performance of the different kits we aligned the reads against a reference (Aus0004) for the construction of a SNP scheme. For the analysis we only selected SNPs from the core genome to reduce false SNPs caused by the distance to the reference. Using high coverage samples (≥50-fold Qia, 50-fold flex, 100-fold XT), we found more than 2,000 SNPs between the isolates of ST117 and ST796. In contrast, within each sequence type only 1-2 SNP differences were identified (**Figure S3A**). All strains showed the same distance to the reference at the root of

TABLE 2 | Prediction of AMR determinants in sequenced ATCC strains compared to reference genomes.

ATCC strain	Resistance mechanism*			Automatically detected in															
				XT					Flex					Qia					
	Name	% coverage	% identity	10	20	50	100	200	10	20	50	100	200	10	20	50	100	200
ATCC25177	*aac(2')-Ic*	100	100	Y	Y	Y	Y	N	Y	N	Y	Y	Y	Y	Y	Y	Y	Y
	erm(37)	100	100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y
	blaA	100	100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCC25922	*blaEC-5*	100	100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCC25923	*tet(38)*	100	100	2	2	2	2	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	fosD	100	79.05	N	N	N**	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCC27853	*fosA*	100	98.53	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	catB7	100	99.22	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	blaOXA-396	100	100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	aph(3')-IIb	100	98.39	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	blaPDC-303	100	99.92	P	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCC29212	*dfrE*	100	97.98	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	P	Y	Y	Y	Y
	tet(M)	100	100	Y	Y	Y	Y	Y	P	Y	Y	Y	Y	Y	Y	Y	Y	Y
	lsa(A)	100	99.8	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCC700603	*blaOKP-B-23*	100	99.42	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	oqxA10	100	93.79	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	oqxB11	100	95.94	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	fosA	100	95.24	P	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	blaSHV-18	100	100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
	ant(2")-Ia	100	100	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	aphA16	100	100	N	N	2	2	Y	N	Y	Y	Y	Y	2	Y	Y	Y	Y
	aadA10	100	87.59	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	blaOXA-2	100	100	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	qacEdelta1	100	100	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
	sul1	100	100	N	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCC700819	*blaOXA-605*	99.75	99.63	P	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
ATCCBAA-67	*penA*	95.74	84.89	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y

**All under 70% coverage and/or 70% identity were screened out. Y, identified; N, not identified (red); P, partial (yellow); 2, split over 2 contigs (yellow). **This sequence also assembled a contig of 896 bp which is predicted to carry a dfrC resistance determinant: % coverage 91; % identity 76.*

the tree (**Figure S3B**). However, if we also include samples with lower sequencing depth (≤20-fold Qia, ≤20-fold flex, ≤50-fold XT), we find a higher diversity in the pairwise comparison of the strains (**Figure S3C**): up to 49 SNPs among ST117 isolates, and in up to 51 SNPs among the ST796 isolates. This is an indication that we are discovering falsely called SNPs. The neighbor joining phylogeny also shows that that subsamples with lower sequencing depth have a smaller distance to the root, as not all SNPs are called (**Figure S3D**). Therefore, we conclude that we can improve SNP typing: lowering the number of falsely called SNPs and increasing the number of "real" SNPs, by using higher sequencing coverage, and Qia and flex kits.

We identified AMR genes in the VRE isolates (**Table S1**). Using Qia and flex, all AMR genes are found if at least a coverage of 50-fold is used. For XT, a coverage of at least 100-fold is needed to ensure the detection of all genes.

Analysis of *Klebsiella aerogenes* Outbreak Analysis

Five *K. aerogenes* isolates from a small outbreak were investigated using the commercial software Ridom SeqSphere+ for cgMLST, and CLC genomics for SNP analysis. The cgMLST analysis gave the same results for the all library kits at 200-fold (**Figure 4A**), showing small numbers of allelic discriminations between the isolates. Using lower coverage subsampled datasets, the number of identified allelic differences becomes smaller (**Figure S4**). Using flex, all the strains could be differentiated even with 10-fold coverage, which was not the case for Qia and XT, where isolates began to collapse into clusters. In the SNP analysis we did not find any differences between the kits with 200-fold coverage (**Figure 4B**), with each identifying 15 SNPs separating the five isolates. However, if we perform the analysis with the lower subsampled reads, again the resolution declines (**Figure 4C**). We could still capture the whole diversity using flex with 100-fold and 50-fold coverage, and Qia with 100-fold. Using XT we identified 13 and 14 SNPs (1 SNP was falsely called) in the 100-fold and 50-fold dataset, respectively.

DISCUSSION

Sequence Quality

This in-depth comparison of three commercial library preparations kits shows the superiority of the Qia and flex kits over XT concerning the quality of the data produced.

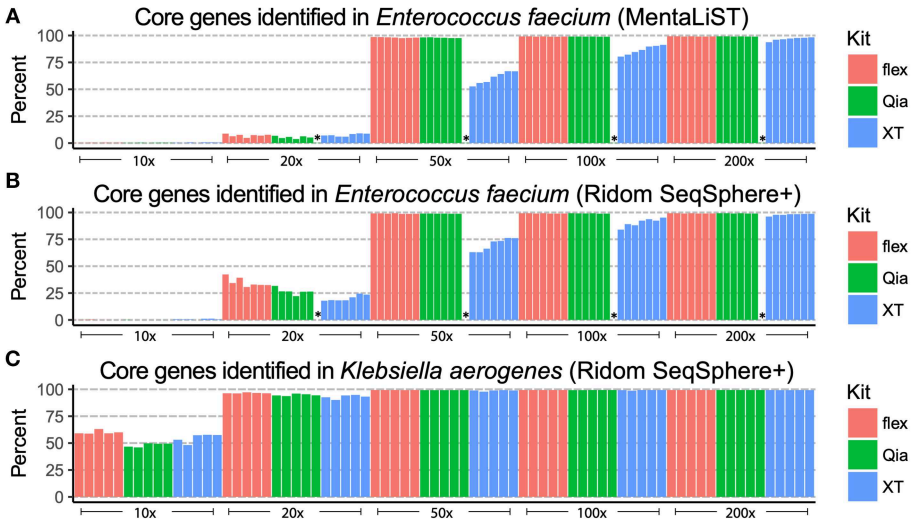


FIGURE 3 | cgMLST alleles identified from the patient isolates. The different subsamples (x-axis) were used to determine of the alleles of the core genome. The different strains are depicted as bars. The y-axis shows the percentage of core genes that can be used for allelic typing. The colors indicate the different library preparation kits. The *E. faecium* isolates were analyzed using Mentalist **(A)** and Ridom SeqSphere+ **(B)**. The *K. aerogenes* isolates were analyzed only using Ridom SeqSphere+ **(C)**. The failed Qia library is labeled with "*".

Through our strategic study design, including a range of human pathogens, we have shown that these two methods produce high quality NGS data that represent the whole genome, at a mean coverage of at least 50-fold. The fragmentation step of these methods is highly stable to variability in the %G+C-content in the genome, resulting in almost even distribution of read depth. In contrast, XT is highly affected by the %G+C-content variation within and between the genomes. This results in an incomplete representation of the genome, especially if lower read depths (<100-fold) are used. Therefore, we suggest that, while Qia and flex libraries can be relied on at mean coverages 50-fold and above, a higher sequencing depth for libraries prepared with XT is required (over 100-fold), which will affect the number of samples that can be pooled on a sequencing run (**Table S2**). This is crucial for the highest resolution of typing, and for comprehensive surveillance of genomic elements such as AMR and virulence genes. We note that we found limitations of the XT data in terms of genome representation in some cases even at a mean read depth of 200-fold.

Our study protocol used a single DNA extraction protocol, and as such we cannot exclude that the tested kits show a different performance with other protocols. Additionally, we conducted this study without technical replicates, therefore variability between batches could not be assessed.

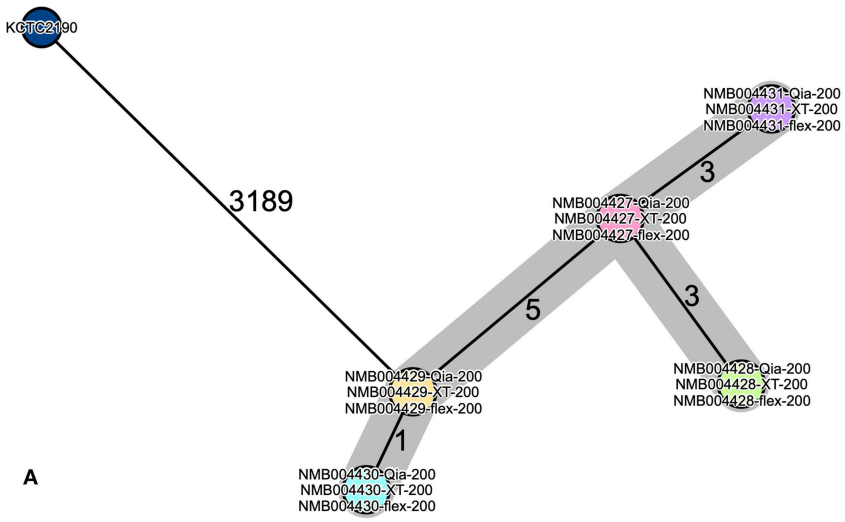
Outbreak Investigation

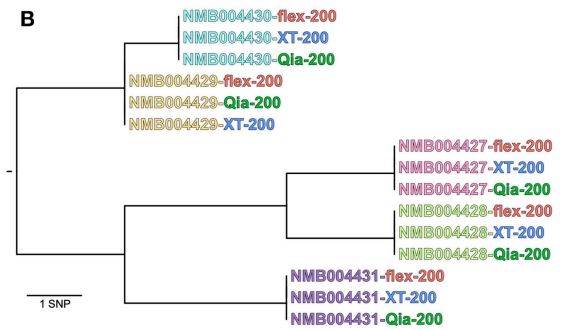
The investigations of the *E. faecium* and *K. aerogenes* patient isolates show the strength of WGS for bacterial typing. Even though the three kits are based on significantly different enzymatic and chemical reactions, the typing results are identical between the methods at high coverage. If the data quality is low, resolution is lost, both in cgMLST and SNP analysis. This is a very important finding for typing laboratories, and especially large-scale projects that want to compile NGS data from nationwide labs to establish national surveillance (35). In all settings, however: local, national or global, the quality control and bioinformatic analysis remain key for epidemiological analysis, as low-quality data can affect the outcomes by lowering the resolution or allowing the false calling of SNPs.

Usability in the Laboratory

The evaluation of the usability in the laboratory showed that XT is the quickest protocol (2.5 h). The core protocols of Qia and flex take a least 1 h longer. However, in the flex protocol, library normalization is included, which reduces the time needed to pool the libraries. This protocol also offers a flexible input amount (50–500 ng) that does not require optimization, saving time in the DNA preparation. However, if <50 ng is available, the number of PCR cycles has to be increased. The Qia protocol needs an accurate measurement of the input DNA, as the resulting fragmentation depends on the DNA amount. If <100 ng input DNA is used, additional PCR and clean-up steps are required that prolong the library preparation by a further 60–90 min. Therefore, XT is superior in time efficiency, but closely followed by flex.

The fragment length is also a very important factor in the sequencing process. Libraries with insert sizes that are smaller than the read length lead to overlaps in the reads pair and therefore loss of sequence information. Fragment length also affects the cluster density calculation, the clustering efficiency and the sequencing depth. If the fragment length is stable across different sample types, the amount of DNA is sufficient data to calculate the molarity and therefore the number of clusters. Long fragments lead to inefficiency in the clustering





C

Library kit	Subsampling	#SNPs in outbreak
flex	200	15
Qia	200	15
XT	200	15
flex	100	15
Qia	100	15
XT	100	13
flex	50	15
Qia	50	14
XT	50	14*
all combined	50,100,200	13
all combined	20,50,100,200	7
all combined	10,20,50,100,200	0

*2 SNPs not called, and 1 anomalous SNP called.

FIGURE 4 | Analysis of the *K. aerogenes* outbreak isolates. **(A)** The isolates (200-fold subsamples) were analyzed using cgMLST in Ridom SeqSphere+ and are depicted in a minimum spanning tree (MST). The isolates are shown as circles. If two strains are identical they collapse into one circle. The numbers on the lines connecting the different circles show the number of different alleles between two isolates (not to scale). **(B)** The genomic distances between the isolates (200-fold subsamples) is show as a phylogenetic tree representing all SNP differences across the whole genome. **(C)** SNP numbers across the tree called using the different subsamples.

and can result in very low cluster densities; short fragments can lead to over clustering and the failure of a run. Therefore, it is important to produce a stable insert size for the libraries, independent of the input DNA. Our comparisons showed that the insert length in Qia and flex are stable across varying %G+C-content. The insert size from the XT is much more affected by the %G+C-content. In our experience with XT, we obtain much higher cluster densities (occasionally leading to over clustering) when sequencing AT-rich species such as *Campylobacter*, as opposed to *Klebsiella*. It is worth mentioning that we suggest to use 2 × 150 bp reagent kits for these libraries, as the tested libraries generally show an insert size of <350 bp. We do not recommend using the MiSeq Reagent Kit v3 (600-cycle) for these libraries, as it produces 2 × 300 bp reads, and the resulting read pairs would overlap with libraries of this length.

Current trends indicate that WGS will be used more often in routine diagnostics and therefore also the number of samples processed will increase. Thus, an implementation of the library preparations kits on automated liquid handling systems will reduce the time and cost associated with this technology. All three protocols discussed in this study can be implemented on liquid handling systems that are equipped with a thermocycler and a magnetic stand.

We have summarized the important features of the different kits that should support other labs in deciding on the most appropriate library preparation kit (**Table 3**).

FINAL CONCLUSION

The evaluation of the three kits clearly showed that the data quality from libraries made with Qia and flex are superior

TABLE 3 | Key features of the compared library preparation kits.

	Nextera XT	Nextera DNA Flex	QIAseq FX
Time required	2.5 h	4 h	4 h
DNA input amount range (ng)	1–1	1–500	1–1,000
Adjustments required for variable input	No variable input supported	PCR cycles required to be adjusted, using <50 ng	Additional PCR step is required if using <100 ng (+ 90 min)
Insert size behavior	Affected by DNA input amount and %G+C-content	Barely affected by the input DNA	Affected by DNA input amount
Available barcodes	384	384	96
Limitations	Highly affected by input DNA	Bead-linked transposomes (BLT) handling needs practice	Reagent volumes are tight
Key advantage	Simple protocol	Highly standardized output (input DNA independent)	PCR-free (>100 ng input DNA)
Special feature	Fast protocol	Produces normalized libraries (>100 ng input DNA)	Insert size can easily be adjusted to needs
Data quality	Highly variable read depth	High quality data	High quality data
Recommended read depth	G+C < 50%: 200 x G+C ≥ 50%: 100 x	50 x	50 x

to those from XT. The comparison of laboratory processes of the Qia and flex kits shows that flex is superior, as the protocol needs very few adjustments, and less hands-on time for routine questions. Therefore, flex best enables streamlining of the laboratory processes for WGS in the context of surveillance.

DATA AVAILABILITY

The datasets generated for this study can be found in the ENA repository under project PRJEB31421.

AUTHOR CONTRIBUTIONS

HS-S, FB, and DW performed the data analysis and wrote the manuscript. AC and JR performed the data analysis. AE wrote the manuscript.

FUNDING

FB was funded through a grant to AE from Personalized Health Basel.

ACKNOWLEDGMENTS

We thank Christine Kiessling, Magdalena Schneider, Elisabeth Schultheiss, Clarisse Straub, and Rosa-Maria Vesco (University Hospital Basel) for excellent technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2019.00241/full#supplementary-material

REFERENCES

1. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L, Brodkin E, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. *N Engl J Med.* (2011) 364:730–9. doi: 10.1056/NEJMoa1003176
2. Harris SR, Cartwright EJ, Torok ME, Holden MT, Brown NM, Ogilvy-Stuart AL, et al. Whole-genome sequencing for analysis of an outbreak of meticillin-resistant *Staphylococcus aureus*: a descriptive study. *Lancet Infect Dis.* (2013) 13:130–6. doi: 10.1016/s1473-3099(12)70268-2
3. Deurenberg RH, Bathoorn E, Chlebowicz MA, Couto N, Ferdous M, Garcia-Cobos S, et al. Application of next generation sequencing in clinical microbiology and infection prevention. *J Biotechnol.* (2017) 243:16–24. doi: 10.1016/j.jbiotec.2016.12.022
4. Balloux F, Bronstad Brynildsrud O, van Dorp L, Shaw LP, Chen H, Harris KA, et al. From theory to practice: translating Whole-Genome Sequencing (WGS) into the clinic. *Trends Microbiol.* (2018) 26:1035–48. doi: 10.1016/j.tim.2018.08.004
5. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. *Clin Microbiol Infect.* (2018) 24:335–41. doi: 10.1016/j.cmi.2017.10.013
6. Hinic V, Seth-Smith H, Stockle M, Goldenberger D, Egli A. First report of sexually transmitted multi-drug resistant *Shigella sonnei* infections in Switzerland, investigated by whole genome sequencing. *Swiss Med Wkly.* (2018) 148:w14645. doi: 10.4414/smw.2018.14645
7. Wuthrich D, Gautsch S, Spieler-Denz R, Dubuis O, Gaia V, Moran-Gilad J, et al. Air-conditioner cooling towers as complex reservoirs and continuous source of *Legionella pneumophila* infection evidenced by a genomic analysis study in 2017, Switzerland. *Euro Surveill.* (2019) 24:1800192. doi: 10.2807/1560-7917.ES.2019.24.4.1800192.
8. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics.* (2010) 11:595. doi: 10.1186/1471-2105-11-595
9. Maiden MC, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol.* (2013) 11:728–36. doi: 10.1038/nrmicro3093
10. Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a

European *Salmonella* serovar Enteritidis outbreak. *Int J Food Microbiol.* (2018) 274:1–11. doi: 10.1016/j.ijfoodmicro.2018.02.023
11. Schurch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene-based approaches. *Clin Microbiol Infect.* (2018) 24:350–4. doi: 10.1016/j.cmi.2017.12.016
12. Tagini F, Greub G. Bacterial genome sequencing in clinical microbiology: a pathogen-oriented review. *Eur J Clin Microbiol Infect Dis.* (2017) 36:2007–20. doi: 10.1007/s10096-017-3024-6
13. Gargis AS, Kalman L, Lubin IM. Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J Clin Microbiol.* (2016) 54:2857–65. doi: 10.1128/jcm.00949-16
14. Hardwick SA, Deveson IW, Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet.* (2017) 18:473–84. doi: 10.1038/nrg.2017.44
15. Carrico JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect.* (2018) 24:342–9. doi: 10.1016/j.cmi.2017.12.015
16. Crisan A, McKee G, Munzner T, Gardy JL. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. *PeerJ.* (2018) 6:e4218. doi: 10.7717/peerj.4218
17. Rossen JWA, Friedrich AW, Moran-Gilad J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect.* (2018) 24:355–60. doi: 10.1016/j.cmi.2017.11.001
18. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome center's improvements to the Illumina sequencing system. *Nat Methods.* (2008) 5:1005–10. doi: 10.1038/nmeth.1270
19. Tyler AD, Christianson S, Knox NC, Mabon P, Wolfe J, Van Domselaar G, et al. Comparison of sample preparation methods used for the next-generation sequencing of *Mycobacterium tuberculosis*. *PLoS ONE.* (2016) 11:e0148676. doi: 10.1371/journal.pone.0148676
20. Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics.* (2012) 13:341. doi: 10.1186/1471-2164-13-341
21. Lan JH, Yin Y, Reed EF, Moua K, Thomas K, Zhang Q. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. *Hum Immunol.* (2015) 76:166–75. doi: 10.1016/j.humimm.2014.12.016
22. Bruinsma S, Burgess J, Schlingman D, Czyz A, Morrell N, Ballenger C, et al. Bead-linked transposomes enable a normalization-free workflow for NGS library preparation. *BMC Genomics.* (2018) 19:722. doi: 10.1186/s12864-018-5096-9
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170
24. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol.* (2017) 13:e1005595. doi: 10.1371/journal.pcbi.1005595
25. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics.* (2013) 29:1072–5. doi: 10.1093/bioinformatics/btt086
26. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics.* (2014) 30:2068–9. doi: 10.1093/bioinformatics/btu153
27. Li H, Durbin R. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics.* (2009) 25:1754–60. doi: 10.1093/bioinformatics/btp324
28. Brown CT, Irber L. sourmash: a library for MinHash sketching of DNA. *J Open Source Softw.* (2016) 1:27. doi: 10.21105/joss.00027
29. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics.* (2015) 31:3691–3. doi: 10.1093/bioinformatics/btv421
30. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, et al. Core genome multilocus sequence typing scheme for high- resolution typing of *Enterococcus faecium*. *J Clin Microbiol.* (2015) 53:3788–97. doi: 10.1128/jcm.01946-15
31. Feijao P, Yao HT, Fornika D, Gardy J, Hsiao W, Chauve C, et al. MentaLiST - a fast MLST caller for large MLST schemes. *Microb Genom.* (2018) 4:e000146. doi: 10.1099/mgen.0.000146.
32. Buultjens AH, Lam MM, Ballard S, Monk IR, Mahony AA, Grabsch EA, et al. Evolutionary origins of the emergent ST796 clone of vancomycin resistant *Enterococcus faecium*. *PeerJ.* (2017) 5:e2916. doi: 10.7717/peerj.2916
33. Mahony AA, Buultjens AH, Ballard SA, Grabsch EA, Xie S, Seemann T, et al. Vancomycin-resistant *Enterococcus faecium* sequence type 796 - rapid international dissemination of a new epidemic clone. *Antimicrob Resist Infect Control.* (2018) 7:44. doi: 10.1186/s13756-018-0335-z
34. Wassilew N, Seth-Smith HM, Rolli E, Fietze Y, Casanova C, Fuhrer U, et al. Outbreak of vancomycin-resistant *Enterococcus faecium* clone ST796, Switzerland, December 2017 to April 2018. *Euro Surveill.* (2018) 23:1800351. doi: 10.2807/1560-7917.Es.2018.23.29.1800351
35. Egli A, Blanc DS, Greub G, Keller PM, Lazarevic V, Lebrand A, et al. Improving the quality and workflow of bacterial genome sequencing and analysis: paving the way for a Switzerland-wide molecular epidemiological surveillance platform. *Swiss Med Wkly.* (2018) 148:w14693. doi: 10.4414/smw.2018.14693

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

REVIEW
published: 04 September 2019
doi: 10.3389/fpubh.2019.00242

Using Genomics to Track Global Antimicrobial Resistance

Rene S. Hendriksen[1], Valeria Bortolaia[1], Heather Tate[2], Gregory H. Tyson[2], Frank M. Aarestrup[1] and Patrick F. McDermott[2]*

[1] European Union Reference Laboratory for Antimicrobial Resistance, World Health Organisation, Collaborating Center for Antimicrobial Resistance and Genomics in Food borne Pathogens, FAO Reference Laboratory for Antimicrobial Resistance, National Food Institute, Technical University of Denmark, Lyngby, Denmark, [2] Center for Veterinary Medicine, Office of Research, United States Food and Drug Administration, Laurel, MD, United States

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control (ECDC), Sweden

Reviewed by:
Sergey Eremin, World Health Organization (Switzerland), Switzerland
Ana Afonso, University of São Paulo, Brazil

***Correspondence:**
Rene S. Hendriksen
rshe@food.dtu.dk

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *18 June 2019*
Accepted: *13 August 2019*
Published: *04 September 2019*

Citation:
Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM and McDermott PF (2019) Using Genomics to Track Global Antimicrobial Resistance. Front. Public Health 7:242. doi: 10.3389/fpubh.2019.00242

The recent advancements in rapid and affordable DNA sequencing technologies have revolutionized diagnostic microbiology and microbial surveillance. The availability of bioinformatics tools and online accessible databases has been a prerequisite for this. We conducted a scientific literature review and here we present a description of examples of available tools and databases for antimicrobial resistance (AMR) detection and provide future perspectives and recommendations. At least 47 freely accessible bioinformatics resources for detection of AMR determinants in DNA or amino acid sequence data have been developed to date. These include, among others but not limited to, ARG-ANNOT, CARD, SRST2, MEGARes, Genefinder, ARIBA, KmerResistance, AMRFinder, and ResFinder. Bioinformatics resources differ for several parameters including type of accepted input data, presence/absence of software for search within a database of AMR determinants that can be specific to a tool or cloned from other resources, and for the search approach employed, which can be based on mapping or on alignment. As a consequence, each tool has strengths and limitations in sensitivity and specificity of detection of AMR determinants and in application, which for some of the tools have been highlighted in benchmarking exercises and scientific articles. The identified tools are either available at public genome data centers, from GitHub or can be run locally. NCBI and European Nucleotide Archive (ENA) provide possibilities for online submission of both sequencing and accompanying phenotypic antimicrobial susceptibility data, allowing for other researchers to further analyze data, and develop and test new tools. The advancement in whole genome sequencing and the application of online tools for real-time detection of AMR determinants are essential to identify control and prevention strategies to combat the increasing threat of AMR. Accessible tools and DNA sequence data are expanding, which will allow establishing global pathogen surveillance and AMR tracking based on genomics. There is however, a need for standardization of pipelines and databases as well as phenotypic predictions based on the data.

Keywords: global, antimicrobial resistance, surveillance, genomic, bioinformatics tools, microbiology

INTRODUCTION

The science of infectious disease, along with other medical and biological specialties, is undergoing rapid change brought on by the advent of affordable whole genomic sequencing (WGS) technologies (1–3). These technologies are rapidly gaining acceptance as routine methods, and in the process, are transforming laboratory procedures.

The amount of bacterial genomic data being generated is immense. As of this writing, for example, over 190,000 *Salmonella* genomes alone are in the public domain with hundreds being added weekly. A complete genomic DNA sequence represents the highest practicable level of structural detail on the individuating traits of an organism or population. As such, it can be used to provide more reliable microbial identification, definitive phylogenetic relationships, and a comprehensive catalog of traits relevant for epidemiological investigations. This is having a major impact on outbreak investigations and the diagnosis and treatment of infectious diseases, as well as the practice of microbiology and epidemiology (4). Furthermore, DNA sequences are a universal dataset from which, theoretically, any biological feature can be inferred. In clinical applications, this includes the ability to detect antimicrobial resistance (AMR), and to track the evolution and spread of AMR bacteria in a hospital or the community.

AMR is a global health problem that contributes to tens of thousands of deaths per year [Chaired by Jim O'Neill, (5)]. Historically, AMR has been detected as a measurement of the growth inhibitory effects of a chemotherapeutic agent on a bacterial population cultured under specific laboratory conditions. Despite some ancillary enhancements, clinical laboratories to this day rely mainly on diffusion and dilution methods to guide clinical therapy and to monitor AMR over time. Accumulating data show that AMR can be accurately predicted from the genomic sequence for many bacteria. The sequence-based approach to AMR detection requires robust bioinformatics tools to analyze and visualize the genomic structure of the microbial "resistome," defined by AMR genes and their precursors (6). This review summarizes the state of the science in using single isolate WGS to track global AMR.

THE ADVANTAGES OF WHOLE GENOME SEQUENCING

A major advancement enabling resistome surveillance is the demonstrated power to predict AMR from genomic data alone. Several studies including those focused on foodborne pathogens and *Enterobacteriaceae* have shown a high concordance (>96%) between the presence of known AMR genes or mutations and Minimum Inhibitory Concentration (MIC) of several antimicrobials at or above the epidemiological cut-off value or clinical breakpoint for resistance. High sensitivity of >87%, defined by the ability to correctly identify AMR determinants associated with an antimicrobial resistance phenotype (true positive rate) and high specificity of >98%, defined by the ability to correctly identify the absence of AMR determinants in an antimicrobial susceptible phenotype (true negative rate), have been observed depending on the bacterial species analyzed (**Table 1**) (7–18). Furthermore, a growing body of data shows that it is possible to predict AMR, and perhaps the MIC of an antimicrobial, applying machine or deep learning to genome sequence data (19–21). The comparison between phenotype and genotype as well as the application of machine or deep learning are however still in their infancy and additional data on bacterial species beyond the foodborne pathogen domain are needed.

The most obvious advantage of WGS for microbial typing and AMR surveillance is the unprecedented level of detail in one assay that can be used to describe current trends and distinguish emerging tendencies (22). AMR bacteria can be typed and traced by specific allele profiles, rather than just according to phenotypic patterns by drug class. This is exemplified by a study of emerging aminoglycoside-resistant *Campylobacter* in the USA, where WGS revealed that the rising trend was driven by nine different resistance alleles, six of which had never been detected in *Campylobacter* previously and would not have been found easily using PCR (10). Similarly, in one of the first large-scale applications of WGS to investigate a drug-resistant foodborne outbreak in the US in 2011, inconsistent resistance patterns among indistinguishable PFGE types of *Salmonella* serovar Heidelberg were revealed by sequence analysis to be a polymicrobic contamination event, involving various combinations of plasmids and strain types (23).

DNA sequence-based surveillance makes it possible also to define multidrug-resistance (MDR) with much greater precision compared to phenotypic tests (22). It has long been a common practice to define MDR as resistance to compounds from three or more drug classes (24), a definition with limited practical value. Bioinformatic analysis can reveal the co-carriage of specific genes underlying different MDR patterns, allelic trends over time, their genetic context including the potential for horizontal transfer, and their distribution by source. In addition, the presence of co-resistances not assayed on standard drug panels is revealed, such as disinfectant and heavy metal resistance. This level of "deep surveillance" can uncover other potential drivers of AMR persistence and evolution, and the opportunity for a more refined microbial risk analysis based on the association of resistance traits with specific sources.

ONLINE RESOURCES FOR *IN SILICO* ANTIMICROBIAL RESISTANCE DETECTION

The high level of agreement between phenotype and genotype coincides with the development of new and updated versions of bioinformatics tools to predict AMR, and the maturation of well-curated AMR gene databases. In principle, *in silico* AMR detection is performed by using a search algorithm to query input DNA or amino acid sequence data for the presence of a pre-determined set of AMR determinants contained in AMR reference databases (**Figure 1**). This can be performed using proprietary systems offered by commercial companies or open-access systems requiring different levels of user expertise.

TABLE 1 | Concordance between phenotypic susceptibility testing and WGS based predicted antimicrobial resistance.

	Pathogen	No. of pathogens	AST method	No. of antimicrobials	Bioinformatic tool	Sequencing data	Concordance	Sensitivity	Specificity	Comment	References
2013	*S. Typhimurium*	49	MIC	17	ResFinder	Assembled, Velvet	99.74%			Disagreement: 7 isolates including 6 *E. coli* resistent to Spec	(7)
	E. coli	48									
	E. faecalis	50		14							
	E. faecium	50									
2013	*E. coli* (ESBL)	74	DD	7	BLASTn, selected panel	Assembled, Velvet		96%	97%	VM rate: 1.2%/M rate: 2.1%	(8)
	K. pneumonia (ESBL)	69									
2014	*S. aureus*	501	DD/MIC (Vitek)	12	BLASTn, selected panel	Assembled, Velvet		97%	99%	VM rate: 0.5%/M rate: 0.7%	(9)
2016	*C. jejuni*	32	MIC	9	BLASTx	Assembled, CLC-bio	99.2%			Lower concordance to	(10)
	C. coli	82								Gen, Azi, Clin, Tel	
2016	*S. enterica*	104	MIC	14	ResFinder/ ARG-ANNOT/ CARD/BLAST	Assembled, CLC-bio	99.0%	99.2%	99.3%	Lower concordance to	(11)
		536						97.6%	98.0%	aminoglycosides/β-lactams	
2017	*E. coli*	31	MIC	4	Custom DB based on ARDB/CARD/β-lactamase allelles			87%	98%	Neg. predictive value: 97%	(12)
	K. pneumonia	24								Pos. Predictive value: 91%	
	P. aeruginosa	22									
	E. cloacae	13									
2017	*S. enterica*	50	MIC	4	ResFinder/ PointFinder	Assembled, SPAdes	98.4%			Disagreement: 2/2 C.jejuni to FQ/ERY	(13)
	E. coli	50		6							
	C. jejuni	50		4						5 *E. coli* to COL (pmrB)	
2018	*E. faecalis*	97	MIC	11	ResFinder/NCBI Pathogen DB/BLAST	Assembled, CLC-bio	96.5%				(14)
	E. faecium	100									
2018	*S. aureus*	501	DD/MIC	12	GeneFinder/ Mykrobe/ Typewriter	FASTQ/assembled, BLAST	98.3%			Disagreements: 0.7% predicted resistant	(15)
		491									
		397	MIC							0.6% predicted susceptible	
2018	*M. tuberculosis*	10,209	MGIT 960	4	Cortex	Assembled	89.5%			97.1%/99.0% predicted R/S	(16)
				4						97.5%/98.8% predicted R/S	
				4						94.6%/93.6% predicted R/S	
				4						91.3%/96.8% predicted R/S	
2019	*H. pylori*	140	MIC (E-test)	5	ARIBA	FASTQ	99%			Phenotype issues to metronidazole	(17)

1) ESBL: Extended Spectrum Beta-Lactamase, 2) MIC: Minimum Inhibitory Concentration, 3) DD: Disk diffusion, 4) VM: Very Major, 5) M: Major, 6) R/S: Resistant/Susceptible, 7) SPEC: Spectinomycin, 8) GEN: Gentamicin, 9) AZI: Azithromycin, 10) CLIN: Clindamycin, 11) TEL: Telithromycin, 12) FQ: Fluoroquinolone, 13) ERY: Erythromycin, 14) COL: colistin.

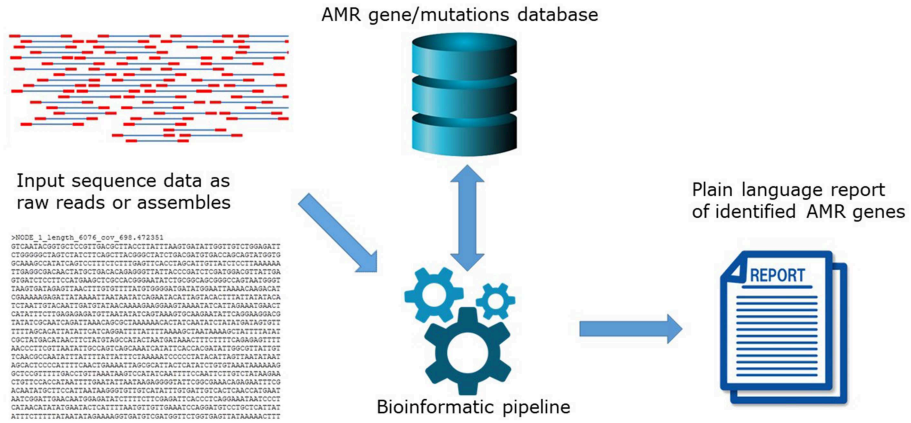


FIGURE 1 | The principle of *in silico* AMR determinant detection using a search algorithm to query input DNA.

Open-access systems are available at public genome data centers such as the Center for Genomic Epidemiology (CGE) http://www.genomicepidemiology.org/ online or downloadable for local install from github (https://github.com/), bitbucket (https://bitbucket.org/account/user/genomicepidemiology/projects/DB) and similar.

The various bioinformatics software can process sequence data either as reads or as assemblies (25). Generally, available resources do not include quality control of input sequence data thus it is the users' responsibility to ensure the quality of submitted sequences or assemblies. When using assembly-based methods, differences among assemblers may compromise comparability of the outcome (15, 26). Following assembly, the most common approaches to compare the input data with the AMR reference databases rely on BLAST and Hidden Markov Model searches, among others. BLAST-based tools can give different outputs based on default settings for gene length and percentage of similarity. This can negatively affect specificity if the settings are too low or too high. Moreover, assembly-based methods are computationally demanding. Despite these caveats, assembly-based methods may have an added value in an AMR surveillance context as they allow analysis of the genetic context of the AMR genes such as their presence on mobilizable potential. Read-based methods may use different tools to align reads to AMR databases, including Bowtie2, BWA, and KMA (25). Recently, the KMA (k-mer alignment) has been develop to map raw reads directly against redundant AMR databases (27). The KMA tool was developed specifically for rapid and accurate bacterial genome analyses in contrast to other mapping methods such as BWA that were developed for large reference genome, such as the human genome and subsequently applied empirically to microbiology (27). KMA uses k-mer seeding to speed-up mapping and the Needleman-Wunsch algorithm to accurately align extensions from k-mer seeds. Multi-mapping reads are resolved using a novel sorting scheme (ConClave scheme) to ensure an accurate selection of templates (27). Read-based methods allow identification of AMR genes present in low abundance which might be overlooked where assemblies are incomplete (25).

Independent of the bioinformatics approach chosen, the performance of *in silico* AMR prediction is critically dependent on the availability of accurate AMR databases. AMR reference databases can be subdivided into solutions specialized for detection of resistance to specific antimicrobials and/or in specific bacterial species or in solutions allowing detection of virtually any possible AMR determinant in any DNA/amino acid sequence. Besides their focus area, AMR reference databases have important differences which users need to acknowledge for choosing the optimal fit-for-purpose database. First, AMR reference databases differ for criteria of inclusion of entries. For example, entries in CARD must have been published in scientific literature. In ResFinder, publication is not a strict requirement. Genes must have a GenBank number and expert review of the GenBank entries. Also, the types of entries differ across databases, with most databases including AMR genes and only a few databases including mutations of chromosomal genes mediating AMR. Finally, the available AMR databases differ regarding the format of the entries (fasta, json, etc.), the possibility of download, and the availability and frequency of curation (**Table 2**).

At present, at least 47 online available resources for *in silico* AMR prediction are published in the scientific literature (13, 26, 28–63) (**Table 2**). They range from basic AMR reference databases that can be embedded in the user's own bioinformatics pipeline, to systems having a well-curated database with integrated search tools. These bioinformatics resources have interfaces of different complexity that require different skills in bioinformatics and microbiology for performing the sequence analyses and interpreting the results (**Table 2**). As the features of these systems differ widely, the outputs obtained by different tools may not be fully comparable. Moreover, employing the same tool for different input formats of the same data (e.g., raw reads vs. assembled sequences, trimmed vs. non-trimmed reads; assemblies obtained by different software, etc.) can produce different results (64). A reliable genomic approach to assaying AMR gene content requires accurate curated reference databases

TABLE 2 | Open-access resources for *in silico* antimicrobial resistance detection in bacteria.

Name	Target	Software		Database		Input sequence		Link	Year of development	Curation (last update)	References
		Type	Downloadable[a]	Source	Downloadable	Type	Format				
ABRES Finder	General AMR	Profile HMM	No	Own	No	Amino acid	FASTA	http://scbt.sastra.edu/ABRES/index.php	2017	Not specified	Unpublished
ABRICATE	General AMR	BLAST	Yes	ResFinder, CARD, ARG-ANNOT, NCBI AMRFinder, EcOH, PlasmidFinder, Ecoli_VF and VFDB	Yes	Nucleotide	FASTA	https://github.com/tseemann/abricate	2016	2019	Unpublished
ARDB	General AMR	BLAST	Yes	Own	Yes	Nucleotide	FASTA	https://ardb.cbcb.umd.edu/	2009	2009	(28)
ARG-ANNOT	General AMR	–	–	Own	Yes	–	–	Discontinued	2014	2018	(29)
ARIBA	General AMR (single isolate sequences)	Minimap, Bowtie2	Yes	Derived from ARG-ANNOT, CARD, PlasmidFinder, ResFinder, VFDB[b]; customizable	No	Nucleotide	FASTQ	https://github.com/sanger-pathogens/ariba	2017	2019	(30)
CARD	General AMR	BLAST, RGI	Yes	Own	Yes	Nucleotide, amino acid	FASTA	https://card.mcmaster.ca/home	2013	2019	(31)
IRIDA plugin AMR detection	General AMR	RGI, staramr	Yes	CARD, PointFinder, PlasmidFinder and ResFinder	Yes	Nucleotide	FASTQ	https://github.com/phac-nml/irida-plugin-amr-detection	2019	2019	Unpublished
Kmer resistance	General AMR	KMA	Yes	ResFinder	Yes	Nucleotide	FASTA, FASTQ	https://cge.cbs.dtu.dk/services/KmerResistance-2.2/	2016	2019	(26)
MEGARes (AMRplusplus)	General AMR	BWA	Yes	Derived from ARG-ANNOT, CARD, NCBI Lahey Clinic beta-lactamase archive, ResFinder[b]	Yes	Nucleotide	FASTQ	https://megares.meglab.org/	2016	2016	(32)

(Continued)

TABLE 2 | Continued

Name	Target	Software		Database		Input sequence		Link	Year of development	Curation (last update)	References
		Type	Downloadable[a]	Source	Downloadable	Type	Format				
NCBI AMRFinder	General AMR	BLAST, HMMER	Yes	Own	Yes	Nucleotide, amino acid	FASTA, GFF	https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/AMRFinder/	2017	2019	(33)
Noradab	General AMR	BLAST	No	Derived from ARDB and CARD[b]	Yes	Nucleotide, amino acid	FASTA	http://noradab.bi.up.ac.za/	2018	Not specified	(34)
Patric	General AMR	BLAST	Yes	Own	Yes	Nucleotide, amino acid	FASTA	https://www.patricbrc.org/	2004	2019	(35)
ResFinder	General AMR	BLAST, KMA	Yes	Own	Yes	Nucleotide	FASTA, FASTQ	https://cge.cbs.dtu.dk/services/ResFinder/	2012	2019	(36)
SRST2	General AMR	BOWTIE2	Yes	Derived from ARG-ANNOT[b]	Yes	Nucleotide	FASTA, FASTQ and any other format readable by BOWTIE2	https://github.com/katholt/srst2	2014	2019	(37)
SSTAR	General AMR	BLAST	Yes	Derived from ARG-ANNOT and Resfinder[b]	Yes	Nucleotide	FASTA	https://github.com/tomdeman-bio/Sequence-Search-Tool-for-Antimicrobial-Resistance-SSTAR-	2015	2018	(38)
INTEGRALL	AMR genes and associated integrons	BLAST	No	Own	Yes	Nucleotide	FASTA	http://integrall.bio.ua.pt/?	2008	2019	(39)
MvirDB	AMR genes, protein toxins and virulence factors for bio-defense applications	BLAST	No	Derived from Tox-Prot, SCORPION, the PRINTS virulence factors, VFDB, TVFac, Islander, ARGO and a subset of VIDA[b]	Yes	Nucleotide, amino acid	FASTA	Discontinued (http://mvirdb.llnl.gov/)	2007	Not specified	(40)
BacMet	Biocide and metal resistance	BLAST	No	Own	Yes	Nucleotide, amino acid	FASTA	http://bacmet.biomedicine.gu.se/	2013	2018	(41)
ResCap	Antibiotic, heavy metal and biocide resistance	BLAST, Bowtie2	Yes	Derived from ARG-ANNOT, CARD, RED-DB, ResFinder, Bacmet[b]	Yes	Nucleotide	FASTA, FASTQ	https://github.com/valflanza/ResCap	2017	2017	(42)

(Continued)

TABLE 2 | Continued

Name	Target	Software		Database		Input sequence		Link	Year of development	Curation (last update)	References
		Type	Downloadable[a]	Source	Downloadable	Type	Format				
ARGO	Beta-lactam and vancomycin resistance	–	–	Own	–	–	–	Discontinued (http://bioinformatics.org/argo/beta/antibioticresistance.php)	2005	–	(43)
RED-DB	Beta-lactam, glycopeptide, aminoglycoside, tetracycline, sulphonamide, macrolide, lincosamide, streptogramin b, oxazolidinone and quinolone resistance	BLAST	No	Own	Yes	Nucleotide, amino acid	FASTA	http://www.fibim.unisi.it/REDDB/	2007-2013	Not specified	Unpublished
Tetracycline MLS nomenclature	Macrolide, lincosamide, streptogramin and tetracycline resistance	–	–	Own	Yes	–	–	https://faculty.washington.edu/marilynr/	Not specified	2019	Unpublished
β-lactamases Database	β-lactamases	–	–	Own	Yes	–	–	http://ifr48.timone.univ-mrs.fr/beta-lactamase/public/	Not specified	Not specified	Unpublished
BLAD	β-lactamases	–	–	Own	No	Nucleotide, amino acid	FASTA	http://www.blad.co.in/	2012	Not specified	Unpublished
BLDB	β-lactamases	BLAST	No	Own	Yes	Nucleotide, amino acid	FASTA	http://bldb.eu/	2017	2019	(44)
CBMAR	β-lactamases	BLAST	No	Own	Yes	Nucleotide, amino acid	FASTA	http://proteininformatics.org/mkumar/lactamasedb/	2014	2014	(45)
LacED	β-lactamases	BLAST	No	Own	Yes	Amino acid	FASTA	http://www.laced.uni-stuttgart.de/	2009	Not specified[c]	(46)
AMRtime	AMR genes in metagenomic data	DIAMOND	Yes	CARD	Yes	Nucleotide	FASTQ	https://github.com/beiko-lab/AMRtime	2017	2019	(47)
DeepARG	AMR genes in metagenomic data	BLAST, DIAMOND	Yes	Derived from RDB, CARD, UNIPROT[b]	Yes	Nucleotide, amino acid	FASTA, FASTQ	https://bench.cs.vt.edu/deeparg	2017	2019	(48)
GROOT	AMR genes in metagenomic data	LSH Forest indexing	Yes	Derived from ARG-ANNOT, CARD, Resfinder	Yes	Nucleotide	FASTQ	https://github.com/will-rowe/groot	2018	2019	(49)
SARG (ARGs-OAP; ARGpore)	AMR genes in metagenomic data	BLAST, HMMER, UBLAST	Yes	Derived from ARDB and CARD[b]	Yes	Nucleotide	any format is supported	https://smile.hku.hk/SARGs	2016	2019	(50)
SEAR	AMR genes in metagenomic data	BLAST, BWA-MEM	Yes	ARG-ANNOT	Yes	Nucleotide	FASTQ	Discontinued (https://github.com/will-rowe/SEAR)	2015	2018	(51)

(Continued)

TABLE 2 | Continued

Name	Target	Software		Database		Input sequence		Link	Year of development	Curation (last update)	References
		Type	Downloadable[a]	Source	Downloadable	Type	Format				
ShortBRED	AMR genes in metagenomic data	BLAST, USEARCH	Yes	Derived from ARDB and CARD[b]	Yes	Amino acid	FASTA	http://huttenhower.sph.harvard.edu/shortbred	2015	2019	(52)
Mustard	AMR determinants in the human gut microbiota	BLAST	No	Derived from Resfinder, ARG-ANNOT, the Lahey Clinic (http://www.lahey.org/studies/), RED-DB (http://www.fibim.unisi.it/REDDB/), Marilyn Roberts' website for macrolides and tetracycline resistance (http://faculty.washington.edu/marilynr/) and different functional metagenomics studies[b]	Yes	Nucleotide, amino acid	FASTA	http://mgps.eu/Mustard/	2017	2017	(53)
FARMEDB	AMR genes discovered by functional metagenomics	BLAST	No	Own	Yes	Nucleotide, amino acid	FASTA	http://staff.washington.edu/jwallace/farme/index.html	2016	Not specified[c]	Unpublished
ResFams	AMR genes discovered by functional metagenomics	–	–	Derived from CARD, LacED, Lahey beta-lactamases (now at NCBI)[b]	Yes	–	–	http://www.dantaslab.org/resfams	2014	2018	(54)
ResFinderFG	AMR genes discovered by functional metagenomics	BLAST	Yes	Own	No	Nucleotide	FASTA, FASTQ	https://cge.cbs.dtu.dk/services/ResFinderFG-1.0/	2016	Not specified	Unpublished
Galileo AMR (MARA, RAC)	AMR genes in Gram-negative bacteria	BLAST (ATTACCA)	Yes	Own	Yes	Nucleotide	FASTA	https://galileoamr.arcbio.com/mara/	2017	Not specified3	(55)
LREfinder	Linezolid resistance in enterococci	KMA	Yes	Own	Yes	Nucleotide	FASTA, FASTQ	https://cge.cbs.dtu.dk/services/LRE-finder/	2019	2019	(56)

(Continued)

TABLE 2 | Continued

Name	Target	Software		Database		Input sequence		Link	Year of development	Curation (last update)	References
		Type	Downloadable[a]	Source	Downloadable	Type	Format				
MUBII-TB-DB	AMR mutations in Mycobacterium tuberculosis	BLAST	No	Own	No	Nucleotide	FASTA	https://umr5558-bibiserv.univ-lyon1.fr/mubii/mubii-select.cgi	2013	Not specified	(57)
Mykrobe	AMR in Mycobacterium tuberculosis and Staphylococcus aureus	Own (based on de Bruijn graph)	Yes	Own	Yes	Nucleotide	FASTQ	http://www.mykrobe.com/	2015	2019	(58)
TBDReaM	AMR in Mycobacterium tuberculosis	–	–	Own	Yes	–	–	https://tbdreamdb.ki.se/Info/	2009	2014	(59)
PointFinder	Selected mutations in chromosomal genes of Escherichia coli, Salmonella sp., Campylobacter sp., Staphylococcus aureus, Enterococcus sp., Mycobacterium tuberculosis, Neisseria gonorrhoeae	BLAST, KMA	Yes	Own	Yes	Nucleotide	FASTA, FASTQ	https://cge.cbs.dtu.dk/services/ResFinder/	2017	2019	(13)
SCCmec Finder	SCCmec elements in Staphylococcus aureus	BLAST, KMA	Yes	Own	Yes	Nucleotide	FASTA, FASTQ	https://cge.cbs.dtu.dk/services/SCCmecFinder/	2016	2018	(60)
U-CARE	AMR in Escherichia coli	BLAST	No	Own	Yes	Amino acid	FASTA	http://www.e-bioinformatics.net/ucare/	2013	Not specified	(61)
ARGDIT	Toolkit for validation and integration of AMR gene database	–	Yes	–	–	Nucleotide, amino acid	FASTA	https://github.com/phglab/ARGDIT	2018	2019	(62)
ARG-miner	Robust and comprehensive curation of AMR gene databases	–	–	Derived from ARDB, ARG-ANNOT, CARD, DeepARG-DB, MEGARes, NDARO, ResFinder, SARG, UniProt[b]	Yes	–	–	https://bench.cs.vt.edu/argminer/#/home	2018	2019 (crowd-curation)	(48)

[a] *Yes, standalone version is available (usually in Bitbucket or in GitHub) either with or without a corresponding web version; no, only web version is available.*

[b] *Curation to avoid redundancies and remove selected sequences (see respective references for details).*

[c] *Active, based on authors' knowledge; discontinued databases may still be available for download via WayBack Machine.*

that should be synchronized and harmonized in a way to ensure comparable outputs worldwide. Once that is achieved, the bioinformatics method of monitoring will undeniably lead to a paradigm shift in the way that we conduct AMR surveillance and compare results internationally. Importantly, the currently available tools may detect new gene variants, but they are not presently equipped to detect new AMR genes. Identifying novel resistance elements from genomic data is being pursued using iterative kmer-based analytics and other machine learning schemes but these strategies still require well-characterized reference genomes with phenotypic data for training (11, 19–21).

BENCHMARKING OF BIOINFORMATICS TOOLS TO DETECT ANTIMICROBIAL RESISTANCE DETERMINANTS

Benchmarking exercises are important to assess the performance, and reliability of the available bioinformatics tools which have different complexity in design and function.

Designing and executing a benchmarking trial offers several challenges. At a recent meeting (October 2017) organized by the European Commission Joint Research Center, the challenges of designing a benchmarking strategy for assessing bioinformatics tools to detect AMR determinants was discussed (65). Here, several challenges were identified, and considerations discussed which included: (1) the origin of the dataset tested; (2) sustainable reference datasets; (3) quality of the test genomes; (4) what determinants to include in a dataset; (5) the, expected result; and (6) performance thresholds. The sequence dataset could either be real or artificially composed. In both cases, this will have implications for accurate benchmarking. A real dataset needs to be properly characterized and the true reference result defined. Furthermore, a real dataset may be biased in content for certain resistance determinants, such as mutations in the *ampC* promoter of *E. coli,* and thereby affect some bioinformatics tools more than others (26). In contrast, a simulated dataset needs to be accurate and correct but also contain a variety of different determinants or mechanisms. Ideally, a combination could be applied designing a desired benchmarking dataset to represent real-life scenarios aligned with the test objective (e.g., only focused on extended spectrum β-lactamases). The scope of bacterial species represented can also influence the results (65).

The quality and type of sequence data are also important factors. This also needs to mimic a real-life scenario where genomes will differ in error rates, read lengths, and read quality and may be raw reads or assemblies. The robustness of bioinformatics tools will differ in performance when dealing with low quality genomes and assemblies compared to optimal conditions (26, 65).

Prior to executing a benchmarking exercise, the reference AMR classes need to be determined as to whether all known or acquired determinants will be included, or only specific mechanisms such as certain enzymes, efflux pumps, mutations/single nucleotide polymorphisms (SNPs), upregulated or downregulated genes or porins. Ideally, the bioinformatics tools should enable the detection of all known determinants if used for surveillance or guiding clinical treatment unless the scope is different and agreed upon (65).

Since the main objective of a benchmarking exercise is to assess the ability of the bioinformatics tool to provide reliable analysis of AMR gene content, it is vital that the concordance is high between the reference result and the expected outcome (65). The sensitivity is especially important as the misidentification of a resistant strain is more consequential than the finding of silent resistance genes in phenotypically susceptible isolates. As previously mentioned, discrepancies observed between phenotypic reference result and the expected genomic outcome is often due to incorrect phenotypic antimicrobial susceptibility test data.

Assessing the performance of bioinformatics tools is often based on a comparison between the genotypic and phenotypic results and a calculation of the specificity, sensitivity, positive predictive (PPV) and negative predictive values (NPV), accuracy [Simple Matching Coefficient (SMC)] and performance [Matthew's Correlation Coefficient (MCC)] followed by a comparison of these parameter's between the different bioinformatics tools (26, 66).

Surprisingly, only a few studies have benchmarked bioinformatics tools against each other to detect AMR determinants. 24 used two previously published pair-end Miseq datasets (7, 8) of 196 genomes of four species and 143 genomes from two species (five species in total), respectively. Phenotypic susceptibility test data was used as the reference result in predicting AMR determinants when benchmarking the KmerResistance vers 1.0 (target only enzymes) (70% identity and 10% depth corr (co-occurrence of K-mers), ResFinder vers. 2.0 (target only enzymes) [98% identity and 60 coverage (assembly/BLAST)], and SRST2 (90% identity 90% coverage) (clustering/Bowtie2). To further challenge the sensitivity, the datasets were down-sampled to 1% of the reads and re-analyzed. Overall, the three bioinformatics tools performed equally well with almost the same accuracy, SMC and performance, MCC testing the two datasets; SMC and MCC were app. 96% and 0.90 for the Stoesser et al. collection, respectively whereas the SMC and MCC ranged from 98 to 100% and 0.91 to 0.99 for the Zankari et al. collection, respectively with the lowest performance by SRST2 and the highest by KmerResistance (26). The KmerResisance tool performed significant better than the two others when data were contaminated or down-sampled to contain a few reads—all bioinformatics tools performed best using raw reads input data (26).

Another study (ENGAGE) (66) evaluated the Public Health England's GeneFinder tool, which targets enzymes and some chromosomal point mutations for fluoroquinolone resistance using two HiSeq datasets, 125 *Salmonella* genomes and 164 *E. coli* genomes of which a large proportion harbored upregulated *ampC*-mediated resistance to extended spectrum cephalosporins. ResFinder provided the highest accuracy, SMC and performance, MCC predicting resistance in the *E. coli* genomes and GeneFinder for *Salmonella* genomes. The correlation to phenotypic susceptibility testing was for *Salmonella* spp. Ninety percent for all bioinformatics tools but higher for GeneFinder specifically for fluoroquinolones. The

accuracy, SMC revealed to be lower in *E. coli* than testing *Salmonella* for all bioinformatics tools due to the bias of the *E. coli* dataset containing a high number of upregulated *ampC* genotypes not predicted by any of the bioinformatics tools (66). Hunt et al. similarly benchmarked the same bioinformatics tools as in Clausen et al. including also the ARIBA tool (30). The ARIBA tool contain in addition to enzymes also chromosomal point mutations thus, outperforming both KmerResistance (26) and SRST2 (37).

Following the benchmarking described above, both the ResFinder and the KmerResistence bioinformatics tools have been updated. Thus, the Resfinder tool now includes a number of chromosomal point mutations such as those to detect resistance to colistin, fluoroquinolones, etc. Overall, the benchmarking exercises revealed that all bioinformatics tools evaluated performed almost similarly good but were affected by the type and quality of input data.

In an assessment of the accuracy of NCBI's AMRFinder, a 2018 study by Feldgarden et al compared it with a 2017 version of ResFinder (33). AMRFinder was evaluated first using a set of 6,242 genomes with 87,679 AST data points for 14 antimicrobial drugs. Overall, 98.4% were consistent with predictions. When compared with ResFinder, most gene calls were identical. While there were 1,229 gene symbol differences, 81% were attributed to differences in database composition. AMRFinder and ResFinder use HMM- and BLAST-based approaches, respectively, and are the commonly used resources for genome-based AMR tracking. Synchronized harmonization of the databases, as is done globally with genomic sequence databases, is needed to minimize inconsistent outputs due to algorithmic differences.

ENSURING HIGH QUALITY GENOMIC DATA BY PROFICIENCY TESTING

Standardization of WGS procedures from DNA preparation to the final genome is paramount to ensure reliable prediction of AMR determinants for surveillance and clinical purposes. To ensure the production of reliable high quality genomic data, laboratories routinely performing WGS should participate in laboratory proficiency testing (PT) or external quality assurance systems (EQAS) (67, 68). For decades, global and regional EQAS in phenotypic AST of foodborne pathogens has been conducted to ensure the quality of performed dilution and diffusion AST (69–71). There is an urgent need to also establish a mechanism to provide a global proficiency testing in the area of WGS to establish standardization in the field (68). This goal is part of the charter of the Global Microbial Identifier (GMI), launched in 2011, to help establish a "global system of DNA genome databases for microbial and infectious disease identification and diagnostics" (https://www.globalmicrobialidentifier.org/).

In 2014, GMI launched its first pilot PT in WGS lead by the DTU and US FDA to trial test the WGS platforms, procedures, test material and the functionality of the assessment pipeline (72). In 2015, a full roll-out of the pilot was delivered by GMI to a global audience. The GMI continued to provide proficiency testing in 2016 and 2017. Cultures and pure DNA for library construction were provided to participating laboratories for DNA purification, library preparation, and WGS followed by *in silico* prediction of wgMLST and AMR determinants. The genomes and analysis were submitted to DTU for quality control assessment using closed genomes of the test strains as a reference. The quality control assessment was facilitated by an in-house developed PT QC pipeline measuring a large number of parameters. These included the numbers of reads after trimming, unmapped reads, map to the total reference DNA, reference chromosome, reference plasmids; proportion of reads that map to reference chromosome; coverage of the reference chromosome and reference plasmids; depth of coverage of total DNA, reference chromosome, and reference plasmids; Phred quality score (Q score), total size and proportion of assembly map to the reference DNA, number of contigs including above a length above 200 bp, N50, and NG50. Underperformance was observed and reported in each trial mainly caused by laboratory contamination or poor performance.

DATA SHARING—PUBLIC/PRIVATE

An important element of genomics as a tool for AMR surveillance and diagnostics is that, once data quality standards are met, the data set is platform-independent, discrete and portable. The analytical outputs and data sharing then become the most important considerations (**Figure 2**). A plethora of international and governmental position papers have stressed the need for global cooperation and data sharing to combat infectious diseases and worsening antimicrobial resistance (73–82). Countries have different levels of legal restriction on the sharing of medical information and biological material with potential commercial value or compliance to the EU General Data Protection Regulation. While the legal issue may be more intractable, the public health advantages to global data sharing are obvious. In the US, where fewer restrictions are in place, WGS data from national surveillance systems are continuously placed in the public domain both for public health purposes, and for exploitation by innovators to develop and update new technologies. This permits global access to information on common microbiological threats, something that will become more important as travel and trade increase and as new threats arise.

ONLINE REPOSITORIES TO HOST AND LINK GENOME AND ANTIMICROBIAL SUSCEPTIBILITY DATA

Concurrently with the vast amount of genomic data being produced, traditional antimicrobial susceptibility testing is being conducted in parallel on a large scale. Up until recently, it was only possible to submit and store DNA sequence data in the International Nucleotide Sequence Database Collaboration (INSDC), whereas all AST data was stored separately in closed local or national repositories. Furthermore, not all genomic data is submitted to the online open genomic repositories of INSDC and shared globally due to difficulties to submit, a lack of appreciation for its value, access to local or national repositories,

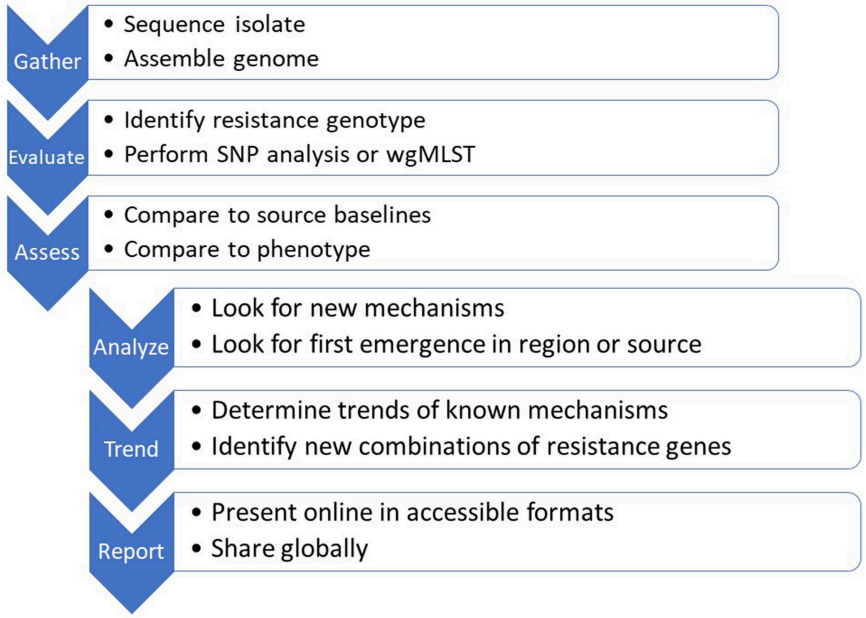


FIGURE 2 | The sequence-based monitoring approach to track global antimicrobial resistance using bioinformatics tools.

fear of being data being published by others, or privacy of the data (83). Nonetheless, today the NCBI and the European Bioinformatics Institute (EMBL-EBI) can accommodate AST data along with the WGS information, to facilitate a global monitoring of AMR in bacteria to strengthen global public health (84, 85).

EUROPEAN NUCLEOTIDE ARCHIVE REPOSITORY

At European Nucleotide Archive (ENA), a mechanism to host and link submitted genomic and AST data has been developed by the EU COMPARE partners and EMBL-EBI (85). Briefly, the EMBL-EBI system allows submitted genomes and associated metadata in the ENA to be stored as open access or privately in a secured login protected repository with named data hubs (86). The system is designed to accommodate submission of susceptibility data from both dilution or diffusion methods. Novel software has been developed to validate conformity of the AST data to ensure harmonization of the data (85). The submitted genomic and AST data could be analyzed by using existing bioinformatics infrastructure and implemented cloud-based bioinformatics workflows in specific an extended version of the Bacterial Analysis Pipeline consisting of ContigAnalyzer-1.0, KmerFinder-2.1, MLST-1.6, ResFinder-2.1, VirulenceFinder-1.2, PlasmidFinder-1.2, pMLST-1.4 (87) with the inclusion of also the cgMLSTFinder 1.0. The submitted data could be queried and downloaded in multiple ways including via the Pathogen Data Portal for surveillance, identification, and investigation https://www.ebi.ac.uk/ena/pathogens/home. Subsequently, the data could be visualized by using a developed Notebook tool integrated the Pathogen Data Portal to query and display all typing data including distribution of the phenotypic AST data enable a potential real time monitoring of AMR (85). The advantage of the data hub model and similar embassy cloud system is the possibility for privacy to control own data having restricted access to only owners or collaborators while analyzing or publishing the data or await less political sensitivity due to GDPR which all a major barriers in data sharing (88–90).

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION REPOSITORY

The National Center for Biotechnology Information (NCBI) is the US member of the INSDC and part of the United States National Institutes of Health, and houses hundreds of thousands of bacterial genomes from around the world. Sequences are submitted from global research studies, but the majority are from national public health surveillance programs with systematic sampling schema. With the expansion of WGS capacity, the number of genome submission is expected to rise soon to over 100,000 annually from US sources alone.

To help make these large datasets accessible, the NCBI Pathogens page (https://www.ncbi.nlm.nih.gov/pathogens/) was developed. This resource is designed for exploring the genomic features of various bacterial pathogens. These include major foodborne and zoonotic pathogens, such as *Salmonella enterica*, *Escherichia coli*, and *Campylobacter* spp. Included in these datasets is a variety of metadata, including strain ID, source, date collected, geographical location, antimicrobial resistance, and more. This page was established in collaboration with GenomeTrakr, an international consortium of laboratories organized by the U.S. Food and Drug Administration (FDA) that

collect and sequence bacterial strains from a variety of food and environmental sources (91).

A major feature of the Pathogens page is the phylogenetic trees, as genomes are arranged into clusters based on relatedness according to SNPs. These allow users to explore and interpret the relatedness of bacterial strains. These have provided a robust database of bacterial species that can be used for genomic comparisons with isolates collected from human patients. This information can be used to help identify foodborne disease outbreaks and support regulatory actions by the FDA.

Another major aspect of the Pathogens page is the AMR reference gene database mentioned above (33). The tool, AMRFinder is automatically run on all genomes submitted to NCBI, resulting in AMR genotype outputs that identify resistance genes from each sequence (33). This, combined with the phylogenetic tree outputs, allows for identification and potential prioritization of investigations into resistant outbreaks of pathogenic organisms.

The NCBI Pathogens web portal also contains phenotypic information, when submitters of these data choose to include it. Over 7,000 isolates now have phenotypic MIC data associated with them, allowing users to interrogate the data for various resistance phenotypes, including those conferred by mutations not tracked presently by the genotypic outputs of AMRFinder (33).

To help make the resistance information accessible, the US Food and Drug Administration developed a tool called ResistomeTracker (https://www.fda.gov/animal-veterinary/national-antimicrobial-resistance-monitoring-system/global-salmonella-resistome-data). This suite of data dashboards is focused exclusively on analysis and visualization of AMR genes extracted from the complete genomes at the NCBI. ResistomeTracker was developed for the U.S. National Antimicrobial Resistance Monitoring System (NARMS) to better understand the epidemiological aspects of resistance by making the large amounts of resistome data accessible to a broad user audience. This includes the identification of new resistance determinants, differences in the prevalence of resistance genes among various food commodities, and geographical spread over time. Additionally, continuous updates to ResistomeTracker enable users to detect early resistance threats. ResistomeTracker allows for user-directed queries of the data that are informative for individual interests. Because it is linked directly to the NCBI pathogen database, it allows the user to begin a query with a specific resistance allele, and end with a phylogenetic analysis of related strains. It currently is focused on foodborne bacteria, but can be modified to exploit and genome for resistance gene content.

USING WGS IN AMR SURVEILLANCE

In the United States, national laboratory capacity for AMR monitoring and WGS is growing. It consists of federally coordinated networks operated by State public health laboratories and Universities. The Centers for Disease Control and Prevention (CDC) coordinates the Antibiotic Resistance Laboratory Network (ARLN) to rapidly detect emerging resistance threats in healthcare, food and the community. Among many activities, this comprehensive network performs WGS for numerous pathogens, including all isolates of *Mycobacterium tuberculosis*. WGS is used also as a routine method to characterize *Neisseria gonorrhoeae,* and other major pathogens, including those involved in outbreaks.

The National Antimicrobial Resistance Monitoring System (NARMS) is a long-standing program focused on bacteria transmitted commonly through food (92). NARMS is a partnership of the CDC, the FDA and United States Department of Agriculture Food Safety and Inspection Service (FSIS); it is focused on tracking resistance in enteric bacteria from humans, retail meats and food animals, respectively. NARMS began systematic WGS of *Salmonella* in 2013 and has incorporated WGS data for *Salmonella* and *Campylobacter* in its reports since 2014. Online tools enable users to examine resistance trends at the genetic level using various query filters. These tools provide graphical visualizations of the genotypes behind changing resistance patterns over time by source and serotype.

As national resistance surveillance matures to better fit the One Health model, animal pathogens and environmental testing are beginning. In the US, the Department of Agriculture National Animal Health Laboratory Network (NAHLN) and the FDA Veterinary Laboratory Investigation and Response Network (Vet-LIRN) are starting to gather resistance information and WGS data on pathogens from food animals and companion animals, respectively. The US Environmental Protection Agency (EPA) conducts periodic water surveys that includes detection of resistance genes. While in the early stages, national public health surveillance programs using DNA sequence information will continue to expand and permit new associations to be inferred from resistomic analyses of the data.

In Europe, its mandatory by law, Directive 2003/99/EC (https://eur-lex.europa.eu/eli/dir/2003/99/oj) for Member States (MSs) to monitor AMR phenotypically by MIC determination in *Salmonella*, *Campylobacter*, and *E. coli* obtained from healthy food-producing animals and from food. The monitoring also include a specific monitoring of extended-spectrum beta-lactamase (ESBL)-, AmpC- and carbapenemase-producing *Salmonella* and indicator commensal *E. coli* stipulated in the Commission Implementing Decision 2013/652/EU of 12 November 2013 (http://data.europa.eu/eli/dec_impl/2013/652/oj). The data collection on human diseases including AMR from MSs is optimal and based on either MIC or disk diffusion and conducted in accordance with Decision 1082/2013/EU (http://data.europa.eu/eli/dec/2013/1082/oj).

A number of MSs providing data for the specific monitoring of AmpC- and carbapenemase-producing *Salmonella* and indicator commensal *E. coli* from healthy food-producing animals and from food, has expressed an interest to replace the mandatory phenotypic MIC determination with WGS due to this already been implemented locally in the specific MSs. Thus, in the preparatory work of updating the Commission Implementing Decision 2013/652/EU coming into force in 2021, the preliminary draft of the technical specifications on

harmonized monitoring of resistance in zoonotic and indicator bacteria from food-producing animals and food from EFSA suggested to allow replacing MIC determination with WGS combined with using the CGE ResFinder tool till 2025 (36). From 2025, the using of WGS combined with using the CGE ResFinder tool will be mandatory for the specific monitoring of AmpC- and carbapenemase-producing *Salmonella* and indicator commensal *E. coli* from healthy food-producing animals and from food and considered to be expended replacing all phenotypic MIC determinations as well as species identification. The resulting AMR determinant profile will be submitted to EFSA and used to predict the phenotype which will be reported in the European Union summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food. It will be optional for the individual MSs to also submit the DNA sequences and metadata data to ENA. It's believed that all MSs by 2015 have acquired WGS and conducing bioinformatics analysis of DNA sequences of single isolates for monitoring purposes.

AMR SURVEILLANCE USING METAGENOMICS

Current AMR surveillance often focuses on few pathogens mainly based on passive reporting of phenotypic laboratory results for a few selected specific pathogens as in the Danish monitoring system, DANMAP https://www.danmap.org/, leading to a narrow pathogen spectrum that does not capture all relevant AMR genes. The majority of AMR genes may be present in the commensal bacterial flora of healthy humans and animals or the environment.

Metagenomics techniques, using short-read next-generation sequencing data, benefit from the ability to quantify thousands of especially transmissible resistance genes in a single sample without any prior selection of which genes to look for. Moreover, it can provide additional information about the presence of bacterial species, pathogens and virulence genes and the data can be re-analyzed, if novel genes of interest are identified.

It was recently shown that metagenomics is superior to conventional methods for AMR surveillance in pig herds (93), useful for comparing AMR across livestock in Europe (94), as well as investigations related to epidemiological data (95). The utility for surveillance of global AMR gene dissemination through international flights (96) and using urban sewage to determine the local and global resistome has also been proven (97, 98).

Metagenomics will sequence all DNA present in the sample including food and host DNA, which may result in low sensitivity. Quantitative PCR procedures, including large scale capture PCR methodologies have been developed, likely providing higher sensitivity (42). However, these methodologies have not been compared with respect to sensitivity and specificity.

In the future the application of metagenomics directly on samples from healthy and clinical ill individuals and animals as well as potential reservoir might results in the ultimate One Health surveillance of AMR allowing determination of all resistance genes and their context in all reservoirs. However, as for single isolates different pipelines and databases are also used for such metagenomics studies and there is a need for global standardization.

PERSPECTIVES

An important advantage of using WGS technologies in detecting and tracking AMR is the opportunity to expand it to align with a One Health surveillance framework and allowing for exact comparisons across reservoirs. This cannot be done using WGS only on the phenotypic antimicrobial class level, but at the exact genetic mechanism level. This One Health goal has so far been impeded by the high cost of testing animal and environmental samples using classical methods based on metabolic and biochemical characterization. As the NGS technology becomes more affordable, it will become more common to use metagenomics to explore the potential role of different environments in the ecology of resistance. Thus, One Health monitoring is now poised to evolve into nucleotide surveillance of complex microbial ecosystems. And to the extent that the data can be generated and reported without delay, it appears that something analogous to a "weather map" of infectious diseases and resistance is possible. This was not practicable in the past, where *ad hoc* gene detection was the norm and PFGE was the typing tool of choice.

CONCLUSION

The advancement in whole genome sequencing and the application of online tools for real-time detection of AMR determinants is essential for control and prevention strategies to combat the increasing threat of AMR. We identified a number of accessible tools available in the prediction of AMR determinants to support expanding to establish global pathogen surveillance and AMR tracking based on genomics. In addition, we identified a number of preceding requirements for a successful transition such as curated AMR databases ensuring a high concordance between pheno- and genotypes, benchmarking designs, PT schemes, sharing options etc. There is however, a vital need for standardization of pipelines and databases as well as phenotypic predictions based on the genomic data.

AUTHOR CONTRIBUTIONS

RH and PM conceived, outlined and critically revised the manuscript. All authors wrote, read and accepted the manuscript.

FUNDING

This study has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 643476 (COMPARE) and from the Novo Nordisk Foundation (NNF16OC0021856: Global Surveillance of Antimicrobial Resistance).

REFERENCES

1. Didelot X, Bowden R, Wilson DJ, Peto TE, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet.* (2012) 13:601–12. doi: 10.1038/nrg3226
2. Fricke WF, Rasko DA. Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions. *Nat Rev Genet.* (2014) 15:49–55. doi: 10.1038/nrg3624
3. Koser CU, Ellington MJ, Cartwright EJ, Gillespie SH, Brown NM, Farrington M, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. *PLoS Pathog.* (2012) 8:e1002824. doi: 10.1371/journal.ppat.1002824
4. Allard MW, Stevens EL, Brown EW. All for one and one for all: the true potential of whole-genome sequencing. *Lancet Infect Dis.* (2019) 19:683–4. doi: 10.1016/S1473-3099(19)30172-0
5. O'Neill J. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations. The Review on Antimicrobial Resistance* London, UK: Wellcome Trust and the UK Department of Health (2016).
6. Cartwright EJ, Patel MK, Mbopi-Keou FX, Ayers T, Haenke B, Wagenaar BH, et al. Recurrent epidemic cholera with high mortality in Cameroon: persistent challenges 40 years into the seventh pandemic. *Epidemiol Infect.* (2013) 141:2083–93. doi: 10.1017/S0950268812002932
7. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agerso Y, Lund O, et al. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. *J Antimicrob Chemother.* (2013) 68:771–7. doi: 10.1093/jac/dks496
8. Stoesser N, Batty EM, Eyre DW, Morgan M, Wyllie DH, Del Ojo EC, et al. Predicting antimicrobial susceptibilities for Escherichia coli and Klebsiella pneumoniae isolates using whole genomic sequence data. *J Antimicrob Chemother.* (2013) 68:2234–44. doi: 10.1093/jac/dkt180
9. Gordon NC, Price JR, Cole K, Everitt R, Morgan M, Finney J, et al. Prediction of Staphylococcus aureus antimicrobial resistance by whole-genome sequencing. *J Clin Microbiol.* (2014) 52:1182–91. doi: 10.1128/JCM.03117-13
10. Zhao S, Tyson GH, Chen Y, Li C, Mukherjee S, Young S, et al. Whole-genome sequencing analysis accurately predicts antimicrobial resistance phenotypes in *Campylobacter* spp. *Appl Environ Microbiol.* (2015) 82:459–66. doi: 10.1128/AEM.02873-15
11. McDermott PF, Tyson GH, Kabera C, Chen Y, Li C, Folster JP, et al. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal salmonella. *Antimicrob Agents Chemother.* (2016) 60:5515–20. doi: 10.1128/AAC.01030-16
12. Shelburne SA, Kim J, Munita JM, Sahasrabhojane P, Shields RK, Press EG, et al. Whole-genome sequencing accurately identifies resistance to extended-spectrum beta-lactams for major gram-negative bacterial pathogens. *Clin Infect Dis.* (2017) 65:738–45. doi: 10.1093/cid/cix417
13. Zankari E, Allesoe R, Joensen KG, Cavaco LM, Lund O, Aarestrup FM. PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. *J Antimicrob Chemother.* (2017) 72:2764–8. doi: 10.1093/jac/dkx217
14. Tyson GH, Sabo JL, Rice-Trujillo C, Hernandez J, McDermott PF. Whole-genome sequencing based characterization of antimicrobial resistance in *Enterococcus*. *Pathog Dis.* (2018) 76:4931055. doi: 10.1093/femspd/fty018
15. Mason A, Foster D, Bradley P, Golubchik T, Doumith M, Gordon NC, et al. Accuracy of different bioinformatics methods in detecting antibiotic resistance and virulence factors from *Staphylococcus aureus* whole-genome sequences. *J Clin Microbiol.* (2018) 56:e01815-17. doi: 10.1128/JCM.01815-17
16. CRyPTIC Consortium and the 100,000 Genomes Project, Allix-Béguec C, Arandjelovic I, Bi L, Beckert P, Bonnet M, et al. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. *N Engl J Med.* (2018) 379:1403–15. doi: 10.1056/NEJMoa1800474
17. Lauener FN, Imkamp F, Lehours P, Buissonniere A, Benejat L, Zbinden R, et al. Genetic determinants and prediction of antibiotic resistance phenotypes in *Helicobacter pylori*. *J Clin Med.* (2019) 8:E53. doi: 10.3390/jcm8010053
18. Su M, Satola SW, Read TD. Genome-based prediction of bacterial antibiotic resistance. *J Clin Microbiol.* (2019) 57:e01405-18. doi: 10.1128/JCM.01405-18
19. Eyre DW, De SD, Cole K, Peters J, Cole MJ, Grad YH, et al. WGS to predict antibiotic MICs for Neisseria gonorrhoeae. *J Antimicrob Chemother.* (2017) 72:1937–47. doi: 10.1093/jac/dkx067
20. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal salmonella. *J Clin Microbiol.* (2019) 57:e01260-18. doi: 10.1128/JCM.01260-18
21. Nguyen M, Brettin T, Long SW, Musser JM, Olsen RJ, Olson R, et al. Developing an *in silico* minimum inhibitory concentration panel test for *Klebsiella pneumoniae*. *Sci Rep.* (2018) 8:421–18972. doi: 10.1038/s41598-017-18972-w
22. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, 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. doi: 10.1016/j.cmi.2016.11.012
23. Hoffmann M, Zhao S, Pettengill J, Luo Y, Monday SR, Abbott J, et al. Comparative genomic analysis and virulence differences in closely related salmonella enterica serotype heidelberg isolates from humans, retail meats, and animals. *Genome Biol Evol.* (2014) 6:1046–68. doi: 10.1093/gbe/evu079
24. Magiorakos AP, Srinivasan A, Carey RB, Carmeli Y, Falagas ME, Giske CG, et al. Multidrug-resistant, extensively drug-resistant and pandrug-resistant bacteria: an international expert proposal for interim standard definitions for acquired resistance. *Clin Microbiol Infect.* (2012) 18:268–81. doi: 10.1111/j.1469-0691.2011.03570.x
25. Boolchandani M, D'Souza AW, Dantas G. Sequencing-based methods and resources to study antimicrobial resistance. *Nat Rev Genet.* (2019) 20:356–70. doi: 10.1038/s41576-019-0108-4
26. Clausen PT, Zankari E, Aarestrup FM, Lund O. Benchmarking of methods for identification of antimicrobial resistance genes in bacterial whole genome data. *J Antimicrob Chemother.* (2016) 71:2484–8. doi: 10.1093/jac/dkw184
27. Clausen PTLC, Aarestrup FM, Lund O. Rapid and precise alignment of raw reads against redundant databases with KMA. *BMC Bioinformatics.* (2018) 19:307–2336. doi: 10.1186/s12859-018-2336-6
28. Liu B, Pop M. ARDB–antibiotic resistance genes database. *Nucleic Acids Res.* (2009) 37:D4437. doi: 10.1093/nar/gkn656
29. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother.* (2014) 58:212–20. doi: 10.1128/AAC.01310-13
30. Hunt M, Mather AE, Sanchez-Buso L, Page AJ, Parkhill J, Keane JA, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genom.* (2017) 3:e000131. doi: 10.1099/mgen.0.000131
31. Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* (2017) 45:D566–D573. doi: 10.1093/nar/gkw1004
32. Lakin SM, Dean C, Noyes NR, Dettenwanger A, Ross AS, Doster E, et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* (2017) 45:D574–80. doi: 10.1093/nar/gkw1009
33. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Using the NCBI AMRFinder tool to determine antimicrobial resistance genotype-phenotype correlations within a collection of NARMS isolates. *BioRxiv [Preprint].* (2019). doi: 10.1101/550707
34. Van Goethem MW, Pierneef R, Bezuidt OKI, Van De PY, Cowan DA, Makhalanyane TP. A reservoir of 'historical' antibiotic resistance genes in remote pristine Antarctic soils. *Microbiome.* (2018) 6:40–0424. doi: 10.1186/s40168-018-0424-5
35. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res.* (2017) 45:D535–D542. doi: 10.1093/nar/gkw1017
36. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* (2012) 67:2640–4. doi: 10.1093/jac/dks261
37. Inouye M, Dashnow H, Raven LA, Schultz MB, Pope BJ, Tomita T, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* (2014) 6:90–0090. doi: 10.1186/s13073-014-0090-6

38. de Man TJ, Limbago BM. SSTAR, a stand-alone easy-to-use antimicrobial resistance gene predictor. *mSphere.* (2016) 1:e00050-15. doi: 10.1128/mSphere.00050-15
39. Moura A, Soares M, Pereira C, Leitao N, Henriques I, Correia A. INTEGRALL: a database and search engine for integrons, integrases and gene cassettes. *Bioinformatics.* (2009) 25:1096–8. doi: 10.1093/bioinformatics/btp105
40. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T. MvirDB–a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res.* (2007) 35:D391–4. doi: 10.1093/nar/gkl791
41. Pal C, gtsson-Palme J, Rensing C, Kristiansson E, Larsson DG. BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids Res.* (2014) 42:D737–43. doi: 10.1093/nar/gkt1252
42. Lanza VF, Baquero F, Martinez JL, Ramos-Ruiz R, Gonzalez-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted metagenomics. *Microbiome.* (2018) 6:11–0387. doi: 10.1186/s40168-017-0387-y
43. Scaria J, Chandramouli U, Verma SK. Antibiotic Resistance Genes Online (ARGO): a Database on vancomycin and beta-lactam resistance genes. *Bioinformation.* (2005) 1:5–7. doi: 10.6026/97320630001005
44. Naas T, Oueslati S, Bonnin RA, Dabos ML, Zavala A, Dortet L, et al. Beta-lactamase database (BLDB) - structure and function. *J Enzyme Inhib Med Chem.* (2017) 32:917–9. doi: 10.1080/14756366.2017.1344235
45. Srivastava A, Singhal N, Goel M, Virdi JS, Kumar M. CBMAR: a comprehensive beta-lactamase molecular annotation resource. *Database.* (2014) 2014:bau111. doi: 10.1093/database/bau111
46. Thai QK, Bos F, Pleiss J. The lactamase engineering database: a critical survey of TEM sequences in public databases. *BMC Genomics.* (2009) 10:390. doi: 10.1186/1471-2164-10-390
47. Maguire F, Alcock B, Brinkman FS, McArthur AG, Beiko RG. AMRtime: Rapid Accurate Identification of Antimicrobial Resistance Determinants from Metagenomic Data. In: *Third American Society for Microbiology Meeting on Rapid Applied Microbial Next-Generation Sequencing and Bioinformatics Pipelines*. Washington, DC (2018).
48. Arango-Argoty G, Garner E, Pruden A, Heath LS, Vikesland P, Zhang L. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome.* (2018) 6:23–0401. doi: 10.1186/s40168-018-0401-z
49. Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics.* (2018) 34:3601–8. doi: 10.1093/bioinformatics/bty387
50. Yin X, Jiang XT, Chai B, Li L, Yang Y, Cole JR, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. *Bioinformatics.* (2018) 34:2263–70. doi: 10.1093/bioinformatics/bty053
51. Rowe W, Baker KS, Verner-Jeffreys D, Baker-Austin C, Ryan JJ, Maskell D, et al. Search engine for antimicrobial resistance: a cloud compatible pipeline and web interface for rapidly detecting antimicrobial resistance genes directly from sequence data. *PLoS ONE.* (2015) 10:e0133492. doi: 10.1371/journal.pone.0133492
52. Kaminski J, Gibson MK, Franzosa EA, Segata N, Dantas G, Huttenhower C. High-specificity targeted functional profiling in microbial communities with ShortBRED. *PLoS Comput Biol.* (2015) 11:e1004557. doi: 10.1371/journal.pcbi.1004557
53. Ruppe E, Ghozlane A, Tap J, Pons N, Alvarez AS, Maziers N, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat Microbiol.* (2019) 4:112–23. doi: 10.1038/s41564-018-0292-6
54. Gibson MK, Forsberg KJ, Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* (2015) 9:207–16. doi: 10.1038/ismej.2014.106
55. Partridge SR, Tsafnat G. Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: the Multiple Antibiotic Resistance Annotator (MARA) and database. *J Antimicrob Chemother.* (2018) 73:883–90. doi: 10.1093/jac/dkx513
56. Hasman H, Clausen PTLC, Kaya H, Hansen F, Knudsen JD, Wang M, et al. LRE-Finder, a Web tool for detection of the 23S rRNA mutations and the optrA, cfr, cfr(B) and poxtA genes encoding linezolid resistance in enterococci from whole-genome sequences. *J Antimicrob Chemother.* (2019) 74:1473–6. doi: 10.1093/jac/dkz092
57. Flandrois JP, Lina G, Dumitrescu O. MUBII-TB-DB: a database of mutations associated with antibiotic resistance in *Mycobacterium tuberculosis*. *BMC Bioinformatics.* (2014) 15:107. doi: 10.1186/1471-2105-15-107
58. Bradley P, Gordon NC, Walker TM, Dunn L, Heys S, Huang B, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. *Nat Commun.* (2015) 6:10063. doi: 10.1038/ncomms10063
59. Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. *PLoS Med.* (2009) 6:e2. doi: 10.1371/journal.pmed.1000002
60. Kaya H, Hasman H, Larsen J, Stegger M, Johannesen TB, Allesoe RL, et al. SCCmecFinder, a web-based tool for typing of staphylococcal cassette chromosome mec in *Staphylococcus aureus* using whole-genome sequence data. *mSphere.* (2018) 3:e00612-17. doi: 10.1128/mSphere.00612-17
61. Saha SB, Uttam V, Verma V. u-CARE: user-friendly comprehensive antibiotic resistance repository of *Escherichia coli*. *J Clin Pathol.* (2015) 68:648–51. doi: 10.1136/jclinpath-2015-202927
62. Chiu JKH, Ong RT. ARGDIT: a validation and integration toolkit for antimicrobial resistance gene databases. *Bioinformatics.* (2019) 35:2466–74. doi: 10.1093/bioinformatics/bty987
63. Arango-Argoty GA, Guron GKP, Garner E, Riquelme MV, Heath LS, Pruden A, et al. ARGminer: a web platform for crowdsourcing-based curation of antibiotic resistance genes. *bioRxiv [Preprint].* (2019). doi: 10.1101/274282
64. Xavier BB, Das AJ, Cochrane G, De GS, Kumar-Singh S, Aarestrup FM, et al. Consolidating and exploring antibiotic resistance gene data resources. *J Clin Microbiol.* (2016) 54:851–9. doi: 10.1128/JCM.02717-15
65. Angers-Loustau A, Petrillo M, Bengtsson-Palme J, Berendonk T, Blais B, Chan KG, et al. The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. *F1000Res.* (2018) 7:459. doi: 10.12688/f1000research.14509.1
66. Technical University of Denmark - National Food Institute. *Final report of ENGAGE - Establishing Next Generation sequencing Ability for Genomic analysis in Europe.* Istituto Zooprofilattico Sperimentale del Lazio e della Toscana; Federal Institute for Risk Assessment; National Institute of Public Health - National Institute of Hygiene; National Veterinary Research Institute; Public Health England; Animal and Plant Health Agency, and Istituto Zooprofilattico Sperimentale delle Venezie. EN-1431 (2018). 252 p.
67. Gargis AS, Kalman L, Berry MW, Bick DP, Dimmock DP, Hambuch T, et al. Assuring the quality of next-generation sequencing in clinical laboratory practice. *Nat Biotechnol.* (2012) 30:1033–6. doi: 10.1038/nbt.2403
68. Rossen JWA, Friedrich AW, Moran-Gilad J. Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect.* (2018) 24:355–60. doi: 10.1016/j.cmi.2017.11.001
69. Hendriksen RS, Seyfarth AM, Jensen AB, Whichard J, Karlsmose S, Joyce K, et al. Results of use of WHO Global Salm-Surv external quality assurance system for antimicrobial susceptibility testing of Salmonella isolates from 2000 to 2007. *J Clin Microbiol.* (2009) 47:79–85. doi: 10.1128/JCM.00894-08
70. Lo Fo Wong DM, Hendriksen RS, Mevius DJ, Veldman KT, Aarestrup FM. External quality assurance system for antibiotic resistance in bacteria of animal origin in Europe (ARBAO-II) 2003. *Vet Microbiol.* (2006) 115:128–39. doi: 10.1016/j.vetmic.2005.12.016
71. Pedersen SK, Wagenaar JA, Vigre H, Roer L, Mikoleit M, idara-Kane A, et al. Proficiency of WHO global foodborne infections network external quality assurance system participants in identification and susceptibility testing of thermotolerant *Campylobacter* spp. from 2003 to 2012. *J Clin Microbiol.* (2018) 56:e01066-18. doi: 10.1128/JCM.01066-18
72. Deng X, den Bakker HC, Hendriksen RS. Genomic epidemiology: whole-genome-sequencing-powered surveillance and outbreak investigation of foodborne bacterial pathogens. *Annu Rev Food Sci Technol.* (2016) 7:353–74. doi: 10.1146/annurev-food-041715-033259
73. Aarestrup FM, Koopmans MG. Sharing data for global infectious disease surveillance and outbreak detection. *Trends Microbiol.* (2016) 24:241–5. doi: 10.1016/j.tim.2016.01.009

74. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature.* (2017) 544:309–15. doi: 10.1038/nature22040
75. Kaye J, Heeney C., Hawkins N, de Vries J, Boddington, P. Data sharing in genomics–re-shaping scientific practice. *Nat Rev Genet.* (2009) 10:331–5. doi: 10.1038/nrg2573
76. McArthur AG, Tsang KK. Antimicrobial resistance surveillance in the genomic age. *Ann N Y Acad Sci.* (2017) 1388:78–91. doi: 10.1111/nyas.13289
77. Sane JEM. *Overcoming Barriers to Data Sharing in Public Health A Global Perspective.* (2015). London, UK: Chatham House, the Royal Institute of International Affairs.
78. Wielinga PR, Hendriksen RS, Aarestrup FM, Lund O, Smits SL, Koopmans MPG, et al. Global microbial identifier. In: Deng X, den Bakker HC, Hendriksen RS, editors. *Applied Genomics of Foodborne Pathogens*. Springer International Publishing Switzerland; Food Microbiology and Food Safety (2017). p. 13–32. doi: 10.1007/978-3-319-43751-4_2
79. World Health Organization. *Whole Genome Sequencing for Foodborne Disease Surveillance: Landscape Paper*. Geneva (2018). Available online at: https://apps.who.int/iris/handle/10665/272430.
80. Yozwiak NL, Schaffner SF, Sabeti PC. Data sharing: make outbreak research open access. *Nature.* (2015) 518:477–9. doi: 10.1038/518477a
81. Zhang J, Chiodini R, Badr A, Zhang G. The impact of next-generation sequencing on genomics. *J Genet Genomics.* (2011) 38:95–109. doi: 10.1016/j.jgg.2011.02.003
82. Ribeiro DS, van de Burgwal LHM, Regeer BJ. Overcoming challenges for designing and implementing the One Health approach: a systematic review of the literature. *One Health.* (2019) 7:100085. doi: 10.1016/j.onehlt.2019.100085
83. Ribeiro DS, Koopmans MP, Haringhuizen GB. Threats to timely sharing of pathogen sequence data. *Science.* (2018) 362:404–6. doi: 10.1126/science.aau5229
84. Otto M. Next-generation sequencing to monitor the spread of antimicrobial resistance. *Genome Med.* (2017) 9:68–0461. doi: 10.1186/s13073-017-0461-x
85. Matamoros S, Hendriksen RS, Pataki B, Pakseresht N, Rossello M, Silvester N, et al. Accelerating surveillance and research of antimicrobial resistance - an online repository for sharing of antimicrobial susceptibility data associated with whole genome sequences. *BioRxiv [Preprint].* (2019). doi: 10.1101/532267
86. Amid C, Pakseresht N, Silvester N, Jayathilaka S, Lund O, Dynovski LD, et al. The COMPARE data hubs. *bioRxiv [Preprint].* (2019). doi: 10.1101/555938
87. Thomsen MC, Ahrenfeldt J, Cisneros JL, Jurtz V, Larsen MV, Hasman H, et al. A bacterial analysis platform: an integrated system for analysing bacterial whole genome sequencing data for clinical diagnostics and surveillance. *PLoS ONE.* (2016) 11:e0157718. doi: 10.1371/journal.pone.0157718
88. Ribeiro CDS, van Roode MY, Haringhuizen GB, Koopmans MP, Claassen E, van de Burgwal LHM. How ownership rights over microorganisms affect infectious disease control and innovation: a root-cause analysis of barriers to data sharing as experienced by key stakeholders. *PLoS ONE.* (2018) 13:e0195885. doi: 10.1371/journal.pone.0195885
89. Contreras JL. NIH's genomic data sharing policy: timing and tradeoffs. *Trends Genet.* (2015) 31:55–7. doi: 10.1016/j.tig.2014.12.006
90. Shabani M, Bezuidenhout L, Borry P. Attitudes of research participants and the general public towards genomic data sharing: a systematic literature review. *Expert Rev Mol Diagn.* (2014) 14:1053–65. doi: 10.1586/14737159.2014.961917
91. Timme RE, Sanchez Leon M, Allard MW. Utilizing the public GenomeTrakr database for foodborne pathogen traceback. *Methods Mol Biol.* (2019) 1918:201–12. doi: 10.1007/978-1-4939-9000-9_17
92. Karp BE, Tate H, Plumblee JR, Dessai U, Whichard JM, Thacker EL, et al. National antimicrobial resistance monitoring system: two decades of advancing public health through integrated surveillance of antimicrobial resistance. *Foodborne Pathog Dis.* (2017) 14:545–57. doi: 10.1089/fpd.2017.2283
93. Munk P, Andersen VD, de Knegt L, Jensen MS, Knudsen BE, Lukjancenko O, et al. A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. *J Antimicrob Chemother.* (2017) 72:385–92. doi: 10.1093/jac/dkw415
94. Munk P, Knudsen BE, Lukjancenko O, Duarte ASR, Van GL, Luiken REC, et al. Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nat Microbiol.* (2018) 3:898–908. doi: 10.1038/s41564-018-0192-9
95. Van GL, Luiken REC, Sarrazin S, Munk P, Knudsen BE, Hansen RB, et al. The antimicrobial resistome in relation to antimicrobial use and biosecurity in pig farming, a metagenome-wide association study in nine European countries. *J Antimicrob Chemother.* (2019) 74:865–76. doi: 10.1093/jac/dky518
96. Nordahl PT, Rasmussen S, Hasman H, Caroe C, Baelum J, Schultz AC, et al. Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci Rep.* (2015) 5:11444. doi: 10.1038/srep11444
97. Hendriksen RS, Munk P, Njage P, van Bunnik B., McNally, L., Lukjancenko, O., et al. (2019) Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. *Nat Commun.* 10, 1124-08853. doi: 10.1038/s41467-019-08853-3
98. Pehrsson EC, Tsukayama P, Patel S, Mejia-Bautista M, Sosa-Soto G, Navarrete KM, et al. Interconnected microbiomes and resistomes in low-income human habitats. *Nature.* (2016) 533:212–6. doi: 10.1038/nature17672

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

MINI REVIEW
published: 24 October 2019
doi: 10.3389/fpubh.2019.00309

Development and Implementation of Whole Genome Sequencing-Based Typing Schemes for *Clostridioides difficile*

Sandra Janezic [1,2]* *and Maja Rupnik* [1,2]

[1] National Laboratory for Health, Environment and Food, Maribor, Slovenia, [2] Medical Faculty, University of Maribor, Maribor, Slovenia

Clostridioides difficile is an important nosocomial pathogen increasingly observed in the community and in different non-human reservoirs. The epidemiology and transmissibility of *C. difficile* has been studied using a variety of typing methods, including more recently developed whole-genome sequence (WGS) analysis that is becoming used routinely for bacterial typing worldwide. Here we review the schemes for WGS-based typing methods available for *C. difficile* and their applications in the field of human *C. difficile* infection (CDI). The two main approaches to discover genomic variations are single nucleotide variant (SNV) analysis and methods based on gene-by-gene comparisons (frequently called core genome or whole genome MLST, cgMLST, or wgMLST). SNV analysis currently provides the ultimate resolution, however, typing nomenclature and standardized methodology are missing. On the other hand, gene-by-gene approaches allow portability and standardized nomenclature, and are therefore becoming increasingly popular in bacterial epidemiology and outbreak investigation. Two commercial software packages (BioNumerics and Ridom SeqSphere+) and an open source database (EnteroBase) for allele and sequence type determination for *C. difficile* are currently available. Proof-of-concept WGS studies have already enabled advances in the investigation of the population structure of *C. difficile* species, microevolution within the epidemic strains, intercontinental transmission over time and in tracking of transmission events. WGS of clinical *C. difficile* isolates demonstrated a considerable genetic diversity suggesting diverse reservoirs for CDI. WGS was also shown to aid in resolving relapses and reinfections in recurrent CDI and has potential for use as a tool for assessing hospital infection prevention and control performance.

Keywords: ***Clostridioides (Clostridium) difficile*, wgMLST, cgMLST, typing, CDI, SNV**

OPEN ACCESS

Edited by:
Vitali Sintchenko, University of Sydney, Australia

Reviewed by:
Valter Viana Andrade-Neto, Oswaldo Cruz Foundation (Fiocruz), Brazil
Daniel Raymond Knight, Murdoch University, Australia

***Correspondence:**
Sandra Janezic
sandra.janezic@nlzoh.si

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *31 May 2019*
Accepted: *08 October 2019*
Published: *24 October 2019*

Citation:
Janezic S and Rupnik M (2019) Development and Implementation of Whole Genome Sequencing-Based Typing Schemes for Clostridioides difficile. Front. Public Health 7:309. doi: 10.3389/fpubh.2019.00309

INTRODUCTION

Clostridioides (Clostridium) difficile is currently one of the most important human pathogens (1). The majority of *C. difficile* infections (CDI) is still identified or associated with the healthcare environment, though the incidence of community CDI is rapidly increasing. Because of its importance as a nosocomial pathogen, the development of different typing methods was needed to identify and control hospital transmissions and outbreaks. Several typing schemes were introduced

for *C. difficile*; among early phenotypic methods serotyping was used widely, but subsequently replaced by pulsed-field gel electrophoresis (PFGE) and finally by PCR ribotyping which is the current gold standard for *C. difficile* typing (1–3). However, apart from multi locus sequence typing (MLST), standardization of all established typing methods has been difficult and inter-laboratory comparisons hampered (2).

Although these methods have contributed greatly to understanding of the epidemiology of CDI, they usually do not have sufficient discriminatory power to distinguish between closely related stains needed for outbreak investigations and to understand transmission events. With development of new sequencing methodologies, there is now the possibility to sequence and compare whole bacterial genomes and not rely only on a single or a few genomic loci to address the genetic relatedness of strains. Therefore, the genome-wide sequence analysis is now frequently used for molecular typing to provide accurate and reproducible investigation of the relatedness of isolates with the highest level of genetic resolution (4).

Here we will review studies on the development and implementation of typing methods based on whole genome sequencing (WGS) and their applications, focusing mainly on healthcare-associated CDI. Proof-of-concept studies have already demonstrated the general applicability of WGS-based typing for investigation of global and national surveillance of *C. difficile* epidemiology, and have expanded our understanding of transmission dynamics and recurrent infections. All these aspects will be reviewed here. However, use of WGS for strain characterization such as analysis of virulence and resistance gene pool and evolutionary aspects will not be covered in this review. *C. difficile* is commonly isolated also from animals and the environment and the paper by Knight and Riley in this special issue (5) will cover applications of comparative genomics from this perspective.

COMPARATIVE GENOMICS AND TWO DIFFERENT APPROACHES FOR WGS TYPING

For the principles of next-generation sequencing technologies and bioinformatic processes, from the raw sequence data to the genomes, the reader is referred to other recent reviews (4, 6).

To determine the genomic similarities and differences between investigated isolates (e.g., to determine which strains could be clonal) different comparative genomics approaches are available. They differ mainly in methodologies used, easiness of data sharing and their discriminatory power. Below we will briefly describe the two of most commonly used approaches for typing of isolates for epidemiological surveillance purposes. The first one is based on comparison of differences in single nucleotide polymorphisms (SNPs), also called single nucleotide variant (SNV) sites. The second approach is based on analysis of multiple genes across the whole genome, so called gene-by gene or allele-based approaches. This is also designated core genome (cg) or whole genome (wg) multi locus sequence typing (cgMLST or wgMLST) (**Figure 1**).

SNV Approach—When Are Two Strains Clonal?

Strain typing based on core genome SNVs (cgSNVs) is currently considered as a method with very high discriminatory power, since it allows us to distinguish between isolates if their genomes differ in a single nucleotide (7). In this approach, short reads (data generated from sequencing of short genomic fragments) or assembled contigs (longer contiguous sequences of overlapping reads) are mapped against the genome of a reference strain to identify differences in coding and non-coding regions. This process is named variant calling (8). The pipeline that has been widely used for SNV analysis of *C. difficile* includes mapping of short reads to a reference genome, variant calling, filtering of high quality SNVs, and identification and removal of putative recombination regions. The result is a concatenated set of high quality SNVs present in the core genome (part of genome that is common to all comparing isolates). The number of SNVs is subsequently used to asses genetic relatedness of isolates (9–11). Relationships between isolates can be visualized by constructing phylogenetic trees to help us understand transmission networks.

The choice of the reference strain can have significant impact, especially on the resolution of SNV-based approaches. The reference strain should be closely related to the isolates included in the comparison since only the regions present in the reference strain will be used for variant calling. Therefore, the more distant the reference sequence the more regions will be omitted from the analysis. Also, a standardized nomenclature would be difficult to adopt since there are multiple algorithms used to analyze SNVs. For this reason, SNV calling is a very useful method for local transmission analysis but not as appropriate for global strain comparisons, unless the genome sequences are made publicly available. However, in this case, the genomes still need to be (re)analyzed locally (8).

The commonly adopted way to determine relatedness of strains in the SNV approach is to count the number of SNV differences between two sequences (SNV threshold). However, it is important to note that proposed criteria of SNV relatedness should not be taken as the absolute rules but should be considered as a guide (8). To determine the SNV threshold, it is important to know the evolutionary rates, i.e., the rate at which the particular bacterial genome evolves (12). This can be estimated from longitudinal sampling from infected individuals and then assessing the number of accumulated substitutions in the genome over time (9).

By comparing genome sequences of the first and the last isolate obtained from individual patient (samples were collected at a median interval of 51 days), an evolutionary rate of 0.74 SNVs (95% confidence interval, 0.22–1.40) per genome per year and a mean within-host diversity of 0.30 SNVs (95% CI, 0.13–0.42 SNVs) were determined, in the study by Eyre et al. (10). Similar estimations of *C. difficile* evolutionary rates were obtained in other studies, either by using serial samples from the patients with recurrent or on-going CDI and/or in *in vitro* gut models (9, 10, 13). By using this prediction of evolutionary rate, the guideline for two isolates being clonal, or genetically related (are most probably a result of direct transmission), is that there are $\leq$2 SNVs between their sequenced genomes. For

FIGURE 1 | Comparison of allele-based and SNV-based typing approaches. cgMLST, wgMLST, and SNV approaches are based on the genome-wide analysis and MLST includes only seven housekeeping genes. Note that strains B, C, and D are identical in MLST approach (same allelic profile in seven genes) and both are the same MLST-ST; but they would differ in cgMLST, strain C and D having identical cgMLST allelic profile, and strain B differs from C and D in three additional genes. In the SNP-based approach, short reads are aligned to a reference genome and the nucleotide differences in both coding (light blue boxes) and non-coding regions (excluding horizontally acquired elements and putative recombination regions) are determined. The number of SNV differences between the pairs of isolates is presented in the matrix on the right.

genetically unrelated isolates ≥10 SNVs are expected (10). This SNV relatedness criterion has now been widely accepted for transmission networks and outbreak investigations, and used in several studies that will be presented later in this review.

Gene-by-Gene Comparison, cgMLST, and wgMLST

Cg- or wgMLST typing works on the same principles as the classical MLST, described by Maiden et al. (14), a comparison of strains based on sequence differences in a pre-defined set of housekeeping genes/loci. Usually seven housekeeping genes/loci are included in MLST schemes for most bacteria, including *C. difficile* (15). For each of the seven loci, the different sequences are assigned distinct allele numbers and the alleles at each genes are described as the allelic profile (**Figure 1**). Finally, for each allelic profile (the series of seven allele numbers) a unique sequence type (ST) is determined (14).

Because only a small number of genes are included in the analysis, MLST usually does not have sufficient discriminatory power to differentiate between closely related strains, e.g., strains that belong to the same PCR ribotypes, which makes this method insufficient for investigations of transmission events. To overcome this, an extension of MLST using a genome-wide gene-by-gene allele calling of hundreds or thousands of loci, so-called cgMLST and wgMLST was developed (16). The cgMLST scheme utilizes comparison of the non-repetitive genes that are conserved in all the members of a species, so called core genes. On the other hand, wgMLST examines a greater number of loci, and includes accessory genes as well as the core genes; these are genes that are variably present across a species (**Figure 1**), including also repetitive genes and pseudogenes (4).

Available Schemes for *C. difficile* WGS-Based Typing

For *C. difficile*, three publicly available schemes are available for cg- and/or wgMLST typing, and analysis can be performed either by using commercial software (BioNumerics, Applied Maths or SeqSphere+, Ridom) or by a freely accessible online resource (EnteroBase). Additional new schemes are being developed (https://www.biorxiv.org/content/10.1101/686212v1?rss=1). The cgMLST scheme for *C. difficile* include 2270 loci (60.4% of the genes present in strain 630; SeqSphere; https://www.cgmlst.org/ncs/schema/3560802/) (17). wgMLST is available in Enterobase and in BioNumerics where, together with 1,999

core genes, another 6,713 accessory genes are included in the analysis http://www.applied-maths.com/sites/default/files/extra/Release-Note-Clostridium-difficile-schema.pdf.

The advantage of cgMLST and MLST is that sequences and allelic profiles of strains can be compared via the internet with central databases enabling uniform typing nomenclature that facilitate international comparability of typing data (16, 17). On the other hand, wgMLST might offer greater resolution between closely related strains, but the nomenclature is not standardized. However, EnteroBase contains all publically available genomic sequences (uploaded from public archives and assembled into annotated draft genomes), and therefore wgMLST data can be compared to all previously published *C. difficile* genomes and interpreted within a global context (18).

In contrast to SNV, the allele-based approaches do not need the genome of a closely related reference strain for the initial alignment of reads or contigs. Also, in the allele-based approach, both mutation (usually resulting in a single SNV) and recombination (that is more likely to introduce multiple deletions or insertions within allele) are counted as a single evolutionary event, meaning that there is no need to apply additional steps to identify and remove putative recombination regions (9, 19).

To test the discriminatory power and applicability of cgMLST to differentiate closely related strains, Bletz et al. (17) reanalyzed data from published outbreak investigations. With cgMLST they were able to differentiate among epidemiologically related strains and the conclusions were in concordance with the published SNV analysis. By re-analyzing two different outbreak investigations and considering the guide for number of SNV expected in genetically unrelated and related isolates ($\geq$10 SNV and $<$2 SNVs, respectively) (10), the authors proposed a threshold of $\geq$7 alleles difference for strains being unrelated and $\leq$6 alleles for strains that are likely to be clonal. With this threshold, the cgMLST predicted the same clusters of related strains as SNV analysis. All strains within the defined threshold were assigned to the same cgMLST cluster type (CT) (17).

CURRENT IMPLEMENTATION OF WGS TYPING IN HUMAN CDI

The feasibility of using WGS of *C. difficile* genomes on benchtop sequencing platforms for transmission investigation to rapidly distinguish between outbreak and non-outbreak cases in a clinically relevant timescale was first demonstrated in 2012 in a pilot study conducted by Eyre et al. (20). Since then SNV-based analysis has been widely adopted for CDI surveillance and has revealed some novel understandings about transmission dynamics and recurrent infections (**Table 1**).

Source Identification for Hospital CDI Cases

Traditionally, most cases of CDI have been thought to be acquired within the hospital environment, where transmissions occurs by horizontal spread from symptomatic patients (39, 40). However, assessment of CDI transmission in hospital settings by classical genotyping approaches was hampered by the low discriminatory power of used methods and by the number of patients that carry endemic genotypes, either PCR ribotypes or STs (9).

To assess the role of symptomatic patients in the transmission of *C. difficile* in the hospital environment Eyre et al. (10), sequenced genomes of *C. difficile* isolates from 1,223 patients with CDI. In this study, only 35% ($n = 333$) of isolates could be genetically linked (had $\leq$2 SNV) to at least one other isolate from a symptomatic patient and for 36% ($n = 120$) of these cases no plausible epidemiological link could be identified. Isolates from almost half (45%) the patients were genetically unrelated ($\geq$10 SNPs) to any other previous case, meaning that these patients had likely acquired *C. difficile* from sources other than symptomatic patients. These findings suggest that there are rather diverse reservoirs of *C. difficile* and that transmissions other that those occurring between symptomatic patients within the hospital settings should be considered (e.g., asymptomatic patients, animals, households, and environmental sources) (10).

The role of asymptomatic patients in the transmission of *C. difficile* was explored by WGS in another study conducted by Eyre et al. (22), which demonstrated that although asymptomatic carriage is common, transmission from asymptomatic carriers is likely to be infrequent. In a similar Canadian study, slightly higher linkage rates were reported, where 46 and 52% of CDI cases could be linked to previous symptomatic and infected or colonized patients, respectively (36).

A study conducted in a single hospital demonstrated that a diverse set of isolates can be found also among children with CDI and that *C. difficile* transmissions, direct or indirect, between children with CDI are even less frequent (12.5%) than transmissions among adult CDI patients (35).

Several other studies have also addressed the questions of importance of other non-hospital reservoirs in *C. difficile* transmission and are reviewed in more details by Knight and Riley in this issue (5).

Use of WGS for Study of CDI Recurrences

Within 2 months after treatment of an initial CDI episode, up to 25% of patients develop recurrent infection (41). Recurrent infection can be due to reinfection (CDI caused by newly acquired strain) or relapse (CDI caused by the original strain). Discrimination between relapses and reinfection usually does not have direct clinical implications and will not affect treatment. However, it might be important for controlling CDI, either through interventions to manage *C. difficile* transmission, or treatment policies (25). Several studies have already demonstrated usefulness of WGS comparisons in understanding the epidemiology of CDI recurrences (23–26). In these studies, the authors used similar approaches as described for transmission studies. In case of reinfections, isolates from the initial and following episodes were expected to be genetically unrelated, differing $\geq$10 SNVs, and in case of relapses, the isolates would be clonal, differing in $\leq$2 SNVs (23). All studies that explored the source of recurrent infection demonstrated

TABLE 1 | WGS-based studies of *C. difficile* transmissions, outbreaks, or recurrences.

References	Aim	Country	Description
Didelot et al. (9)	Transmission	UK	Microevolutionary analysis of *C. difficile* (assessment of within-host evolutionary rate) and use of whole-genome sequencing for studying *C. difficile* transmission.
Eyre et al. (20)	Transmission	UK	A proof-of-principle study to investigate potentials of benchtop sequencers in routine clinical practice to investigate transmissions. Example of small cluster of genetically (MLST) identical *C. difficile* strains that could be differentiated with WGS.
Eyre et al. (10)	Transmission	UK	Investigating the role of symptomatic patients in the transmission of *C. difficile*. Study also demonstrates that in the settings with standard infection control most cases of infections are acquired from other sources, not symptomatic cases.
Eyre et al. (21)	Mixed infections	UK	Describing new algorithm for detection of mixed CDI in clinical samples from whole genome sequencing data.
Eyre et al. (22)	Transmission	UK	Investigating the role of asymptomatic patients in the transmission of *C. difficile*.
Eyre et al. (23)	Recurrence	UK	Use of WGS to determine if the reductions in recurrence of CDI observed with fidaxomicin occurred by preventing relapse, reinfection or both. Study demonstrated that fidaxomicin was superior to vancomycin in treating recurrent CDI.
Mac Aoga'in et al. (24)	Recurrence	Ireland	Use of WGS of *C. difficile* to discriminate between relapses and reinfections, and putative patient-patient transmission events in Ireland.
Kumar et al. (25)	Transmission	UK	A WGS to track the transmission of *C. difficile* PCR ribotype 027 within single hospital in UK, and to distinguish between the relapses and reinfections.
Sim et al. (26)	Recurrence	USA	Use of WGS to determine the rate of relapse and reinfection in patients with recurrent CDI.
Mawer et al. (27)	Transmission	UK	Exploring the role of symptomatic patients that are toxigenic strain positive but fecal toxin negative in transmissions of *C. difficile*.
Eyre et al. (28)	Transmission	UK	Use of WGS as surveillance tool to assess infection control effectiveness in hospitals by identifying the extent of hospital-acquired CDI transmissions within hospitals.
Stoesser et al. (29)	Transmission	UK	Investigation of genetic overlap of infant and regional *C. difficile* strains in Oxfordshire.
Donskey et al. (30)	Transmission	USA	Transmission of *C. difficile* from colonized or infected long-term care facility residents.
Endres et al. (31)	Outbreak	USA	Environmental transmission of *C. difficile* PCR ribotype 027 at a long-term care facility.
Eyre et al. (32)	Transmission	UK	WGS to analyze distinct patterns of *C. difficile* PCR ribotype spread across Europe.
Halstead et al. (33)	Transmission	UK	WGS to investigate if asymptomatic carriers contribute to nosocomial CDI.
Isidro et al. (34)	Outbreak	Portugal	Genomic investigation of *C. difficile* PCR ribotype 017 outbreak strains.
Kociolek et al. (35)	Transmission	USA	Transmission of CDI among symptomatic children.
Kong et al. (36)	Transmission	Canada	Investigation of transmission patterns between infected and colonized patients.
Williamson et al. (37)	Transmission	USA	Transmission of PCR ribotype 027 within healthcare facility and comparison to global collection of ribotype 027 isolates.
García-Fernández et al. (38)	Transmission	Spain	Routes and frequencies of transmission of *C. difficile* in a tertiary-care hospital in Madrid.

that the majority of recurrent episodes are caused by primarily infecting strain, meaning that relapses are more common than reinfections (23–26).

BACKWARD COMPATIBILITY BETWEEN WGS AND MLST

Currently, an assortment of classical and WGS-based typing methods is used for investigations of *C. difficile* epidemiology (2). Reverse compatibility of WGS with traditional typing methods is therefore important to compare the genotypes obtained with different approaches and to compare newly sequenced strains to existing and historical strains (42). From WGS data, seven MLST loci can be easily extracted to determine the allelic profile and ST. For ST calling directly from draft genomes a publically available PubMLST.org database can be used (43). SeqSphere and BioNumerics also enable ST determination directly from WGS data.

WHY CAN PCR RIBOTYPE NOT BE DETERMINED WITH WGS

PCR ribotyping has become a method of choice for typing of *C. difficile* in the majority of laboratories (2, 44). The method is based on analysis of banding patterns of PCR-amplified internal transcribed spacers (ITS) located between 16S and 23S rRNA genes in ribosomal operon. In *C. difficile*, as in many other bacteria, the ribosomal operon is present in several copies in the genome and different copies differ in the length of ITS (45) and, due to intraspecific diversity of ITS, PCR ribotyping is a good method for *C. difficile* genotyping (2).

In contrast to MLST-ST, PCR ribotype cannot be directly determined from WGS. Regions that are amplified in PCR ribotyping are repetitive and it is not possible to map short sequence reads generated by NGS correctly to such repetitive and modular regions (45, 46). To assign a PCR ribotype to a new ST or cgMLST cluster type, a representative strain would still need to be PCR ribotyped. But with the advances in NGS technologies (e.g.,

PacBio and Nanopore), read lengths are continually increasing (4). The availability of very long and very precise sequences will ultimately enable the *in silico* PCR ribotyping.

The ability to predict PCR ribotypes from whole genome sequencing data remains controversial. While the genome sequences of strains belonging to the same PCR ribotype mostly group together, it is important to appreciate the differences between a true 'PCR ribotype determination' and ribotype inferred from genome sequencing data. Firstly, while grouping of strains with identical PCR ribotype is to be expected, there are exceptions and similarity of genome sequences of two different PCR ribotypes has been documented (36 and unpublished data). Secondly, due to limitations of short read sequencing explained above, comparison of two genomes shows only similarities in large part of genome, but not necessarily in the regions that are actually used for PCR ribotyping (i.e. ITS). Therefore, it is important to differentiate between ribotypes determined by actual PCR ribotyping and putative PCR ribotypes based on genome similarity, but excluding rDNA regions.

CONCLUSION

WGS-based typing methods offer an excellent platform with high resolution and reproducibility that enable studies of both transmission and epidemiology of CDI, as well as positioning strains within the global population. However, especially for the understanding of global CDI epidemiology, whole genome data availability, either by sharing raw data or allelic profiles through freely accessible databases that support direct comparison of isolates is of paramount importance.

AUTHOR CONTRIBUTIONS

SJ and MR both contributed to the conception and writing of the paper.

FUNDING

This work was in part supported by the Slovenian Research Agency grant J4-8224.

REFERENCES

1. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. *Clostridium difficile* infection. *Nat Rev Dis Primer*. (2016) 2:16020. doi: 10.1038/nrdp.2016.20
2. Huber CA, Foster NF, Riley TV, Paterson DL. Challenges for standardization of *Clostridium difficile* typing methods. *J Clin Microbiol*. (2013) 51:2810–4. doi: 10.1128/JCM.00143-13
3. Knetsch CW, Lawley TD, Hensgens MP, Corver J, Wilcox MW, Kuijper EJ. Current application and future perspectives of molecular typing methods to study *Clostridium difficile* infections. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull*. (2013) 18:20381. doi: 10.2807/ese.18.04.20381-en
4. Quainoo S, Coolen JPM, van Hijum SAFT, Huynen MA, Melchers WJG, van Schaik W, et al. Whole-genome sequencing of bacterial pathogens: the future of nosocomial outbreak analysis. *Clin Microbiol Rev*. (2017) 30:1015–63. doi: 10.1128/CMR.00016-17
5. Knight DR, Riley TV. Genomic delineation of zoonotic origins of *Clostridium difficile*. *Front Public Health*. (2019) 7:164. doi: 10.3389/fpubh.2019.00164
6. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. *Clin Microbiol Infect*. (2018) 24:342–9. doi: 10.1016/j.cmi.2017.12.015
7. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. *Nat Rev Genet*. (2012) 13:601–12. doi: 10.1038/nrg3226
8. Schürch AC, Arredondo-Alonso S, Willems RJL, Goering RV. Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clin Microbiol Infect*. (2018) 24:350–4. doi: 10.1016/j.cmi.2017.12.016
9. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol*. (2012) 13:R118. doi: 10.1186/gb-2012-13-12-r118
10. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, et al. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med*. (2013) 369:1195–205. doi: 10.1056/NEJMoa1216064
11. Knight DR, Squire MM, Collins DA, Riley TV. Genome analysis of *Clostridium difficile* PCR ribotype 014 lineage in Australian pigs and humans reveals a diverse genetic repertoire and signatures of long-range interspecies transmission. *Front Microbiol*. (2017) 7:2138. doi: 10.3389/fmicb.2016.02138
12. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. *Microb Genomics*. (2016) 2:e000094. doi: 10.1099/mgen.0.000094
13. Eyre DW, Walker AS, Freeman J, Baines SD, Fawley WN, Chilton CH, et al. Short-term genome stability of serial *Clostridium difficile* ribotype 027 isolates in an experimental gut model and recurrent human disease. *PLoS ONE*. (2013) 8:e63540. doi: 10.1371/journal.pone.0063540
14. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci USA*. (1998) 95:3140–5. doi: 10.1073/pnas.95.6.3140
15. Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, Fung R, et al. Multilocus sequence typing of *Clostridium difficile*. *J Clin Microbiol*. (2010) 48:770–8. doi: 10.1128/JCM.01796-09
16. Maiden MCJ, Jansen van Rensburg MJ, Bray JE, Earle SG, Ford SA, Jolley KA, et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat Rev Microbiol*. (2013) 11:728–36. doi: 10.1038/nrmicro3093
17. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. *J Clin Microbiol*. (2018) 56:e01987–17. doi: 10.1128/JCM.01987-17
18. Zhou Z, Alikhan N-F, Mohamed K, the Agama Study Group, Achtman M. The user's guide to comparative genomics with EnteroBase. Three case studies: micro-clades within *Salmonella enterica* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all Escherichia. *bioRxiv*. (2019) 613554. doi: 10.1101/613554
19. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, et al. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci USA*. (2010) 107:7527–32. doi: 10.1073/pnas.0914322107
20. Eyre DW, Golubchik T, Gordon NC, Bowden R, Piazza P, Batty EM, et al. A pilot study of rapid benchtop sequencing of *Staphylococcus aureus* and *Clostridium difficile* for outbreak detection and surveillance. *BMJ Open*. (2012) 2:e001124. doi: 10.1136/bmjopen-2012-001124
21. Eyre DW, Cule ML, Griffiths D, Crook DW, Peto TEA, Walker AS, et al. Detection of mixed infection from bacterial whole genome sequence data allows assessment of its role in *Clostridium difficile* transmission. *PLoS Comput Biol*. (2013) 9:e1003059. doi: 10.1371/journal.pcbi.1003059

22. Eyre DW, Griffiths D, Vaughan A, Golubchik T, Acharya M, O'Connor L, et al. Asymptomatic *Clostridium difficile* colonisation and onward transmission. *PLoS ONE*. (2013) 8:e78445. doi: 10.1371/journal.pone.0078445
23. Eyre DW, Babakhani F, Griffiths D, Seddon J, Del Ojo Elias C, Gorbach SL, et al. Whole-genome sequencing demonstrates that fidaxomicin is superior to vancomycin for preventing reinfection and relapse of infection with *Clostridium difficile*. *J Infect Dis*. (2014) 209:1446–51. doi: 10.1093/infdis/jit598
24. Mac Aogáin M, Moloney G, Kilkenny S, Kelleher M, Kelleghan M, Boyle B, et al. Whole-genome sequencing improves discrimination of relapse from reinfection and identifies transmission events among patients with recurrent *Clostridium difficile* infections. *J Hosp Infect*. (2015) 90:108–16. doi: 10.1016/j.jhin.2015.01.021
25. Kumar N, Miyajima F, He M, Roberts P, Swale A, Ellison L, et al. Genome-based infection tracking reveals dynamics of *Clostridium difficile* transmission and disease recurrence. *Clin Infect Dis Off Publ Infect Dis Soc Am*. (2016) 62:746–52. doi: 10.1093/cid/civ1031
26. Sim JHC, Truong C, Minot SS, Greenfield N, Budvytiene I, Lohith A, et al. Determining the cause of recurrent *Clostridium difficile* infection using whole genome sequencing. *Diagn Microbiol Infect Dis*. (2017) 87:11–6. doi: 10.1016/j.diagmicrobio.2016.09.023
27. Mawer DPC, Eyre DW, Griffiths D, Fawley WN, Martin JSH, Quan TP, et al. Contribution to *Clostridium Difficile* transmission of symptomatic patients with toxigenic strains who are fecal toxin negative. *Clin Infect Dis Off Publ Infect Dis Soc Am*. (2017) 64:1163–70. doi: 10.1093/cid/cix079
28. Eyre DW, Fawley WN, Rajgopal A, Settle C, Mortimer K, Goldenberg SD, et al. Comparison of control of *Clostridium difficile* infection in six English hospitals using whole-genome sequencing. *Clin Infect Dis Off Publ Infect Dis Soc Am*. (2017) 65:433–41. doi: 10.1093/cid/cix338
29. Stoesser N, Eyre DW, Quan TP, Godwin H, Pill G, Mbuvi E, et al. Epidemiology of *Clostridium difficile* in infants in Oxfordshire, UK: Risk factors for colonization and carriage, and genetic overlap with regional *C. difficile* infection strains. *PLoS ONE*. (2017) 12:e0182307. doi: 10.1371/journal.pone.0182307
30. Donskey CJ, Sunkesula VCK, Stone ND, Gould CV, McDonald LC, Samore M, et al. Transmission of *Clostridium difficile* from asymptomatically colonized or infected long-term care facility residents. *Infect Control Hosp Epidemiol*. (2018) 39:909–16. doi: 10.1017/ice.2018.106
31. Endres BT, Dotson KM, Poblete K, McPherson J, Lancaster C, Bassères E, et al. Environmental transmission of *Clostridioides difficile* ribotype 027 at a long-term care facility; an outbreak investigation guided by whole genome sequencing. *Infect Control Hosp Epidemiol*. (2018) 39:1322–9. doi: 10.1017/ice.2018.230
32. Eyre DW, Davies KA, Davis G, Fawley WN, Dingle KE, De Maio N, et al. Two distinct patterns of *Clostridium difficile* diversity across Europe indicating contrasting routes of spread. *Clin Infect Dis Off Publ Infect Dis Soc Am*. (2018) 67:1035–44. doi: 10.1093/cid/ciy252
33. Halstead FD, Ravi A, Thomson N, Nuur M, Hughes K, Brailey M, Oppenheim BA. Whole genome sequencing of toxigenic *Clostridium difficile* in asymptomatic carriers: insights into possible role in transmission. *J Hosp Infect*. (2018) 102:125–34. doi: 10.1016/j.jhin.2018.10.012
34. Isidro J, Menezes J, Serrano M, Borges V, Paixão P, Mimoso M, et al. Genomic study of a *Clostridium difficile* multidrug resistant outbreak-related clone reveals novel determinants of resistance. *Front Microbiol*. (2018) 9:2994. doi: 10.3389/fmicb.2018.02994
35. Kociolek LK, Gerding DN, Espinosa RO, Patel SJ, Shulman ST, Ozer EA. *Clostridium difficile* whole genome sequencing reveals limited transmission among symptomatic children: a single-center analysis. *Clin Infect Dis*. (2018) 67:229–34. doi: 10.1093/cid/ciy060
36. Kong LY, Eyre DW, Corbeil J, Raymond F, Walker AS, Wilcox MH, et al. *Clostridium difficile*: investigating transmission patterns between infected and colonized patients using whole genome sequencing. *Clin Infect Dis*. (2019) 68:204–9. doi: 10.1093/cid/ciy457
37. Williamson CHD, Stone NE, Nunnally AE, Hornstra HM, Wagner DM, Roe CC, et al. A global to local genomics analysis of *Clostridioides difficile* ST1/RT027 identifies cryptic transmission events in a northern Arizona healthcare network. *Microb Genom*. (2019) 5:e000271. doi: 10.1099/mgen.0.000271
38. García-Fernández S, Frentrup M, Steglich M, Gonzaga A, Cobo M, López-Fresneña N, et al. Whole-genome sequencing reveals nosocomial *Clostridioides difficile* transmission and a previously unsuspected epidemic scenario. *Sci Rep*. (2019) 9:6959. doi: 10.1038/s41598-019-43464-4
39. Vonberg R-P, Kuijper EJ, Wilcox MH, Barbut F, Tüll P, Gastmeier P, et al. Infection control measures to limit the spread of Clostridium difficile. *Clin Microbiol Infect*. (2008) 14:2–20. doi: 10.1111/j.1469-0691.2008.01992.x
40. Khanna S, Pardi DS. *Clostridium difficile* infection: new insights into management. *Mayo Clin Proc*. (2012) 87:1106–17. doi: 10.1016/j.mayocp.2012.07.016
41. Kelly CP. Can we identify patients at high risk of recurrent *Clostridium difficile* infection? *Clin Microbiol Infect*. (2012) 18:21–7. doi: 10.1111/1469-0691.12046
42. Rossen JWA, Friedrich AW, Moran-Gilad J, ESCMID Study Group for Genomic and Molecular Diagnostics (ESGMD). Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. (2018) 24:355–60. doi: 10.1016/j.cmi.2017.11.001
43. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*. (2018) 3:124. doi: 10.12688/wellcomeopenres.14826.1
44. Fawley WN, Knetsch CW, MacCannell DR, Harmanus C, Du T, Mulvey MR, et al. Development and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis PCR-ribotyping protocol for *Clostridium difficile*. *PLoS ONE*. (2015) 10:e0118150. doi: 10.1371/journal.pone.0118150
45. Sadeghifard N, Gürtler V, Beer M, Seviour RJ. The mosaic nature of intergenic 16S-23S rRNA spacer regions suggests rRNA operon copy number variation in *Clostridium difficile* strains. *Appl Environ Microbiol*. (2006) 72:7311–23. doi: 10.1128/AEM.01179-06
46. Janezic S, Indra A, Rattei T, Weinmaier T, Rupnik M. Recombination drives evolution of the *Clostridium difficile* 16S-23S rRNA intergenic spacer region. *PLoS ONE*. (2014) 9:e106545. doi: 10.1371/journal.pone.0106545

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

ORIGINAL RESEARCH
published: 21 November 2019
doi: 10.3389/fpubh.2019.00317

The Transformation of Reference Microbiology Methods and Surveillance for *Salmonella* With the Use of Whole Genome Sequencing in England and Wales

Marie Anne Chattaway[1], Timothy J. Dallman[1], Lesley Larkin[2], Satheesh Nair[1], Jacquelyn McCormick[2], Amy Mikhail[2], Hassan Hartman[1], Gauri Godbole[1], David Powell[1], Martin Day[1], Robert Smith[3] and Kathie Grant[1]*

[1] Gastrointestinal Bacteria Reference Unit, Public Health England, London, United Kingdom, [2] Tuberculosis, Acute Respiratory, Gastrointestinal, Emerging/Zoonotic Infections, and Travel Health and IHR Division (T.A.R.G.E.T.), Public Health England, London, United Kingdom, [3] Public Health Wales, Cardiff, United Kingdom

OPEN ACCESS

Edited by:
Vitali Sintchenko, University of Sydney, Australia

Reviewed by:
Craig Hedberg, University of Minnesota, United States
Pimlapas Leekitcharoenphon, Technical University of Denmark, Denmark

***Correspondence:**
Marie Anne Chattaway
marie.chattaway@phe.gov.uk

Specialty section:
This article was submitted to Infectious Diseases – Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *27 March 2019*
Accepted: *15 October 2019*
Published: *21 November 2019*

Citation:
Chattaway MA, Dallman TJ, Larkin L, Nair S, McCormick J, Mikhail A, Hartman H, Godbole G, Powell D, Day M, Smith R and Grant K (2019) The Transformation of Reference Microbiology Methods and Surveillance for Salmonella With the Use of Whole Genome Sequencing in England and Wales. Front. Public Health 7:317. doi: 10.3389/fpubh.2019.00317

The use of whole genome sequencing (WGS) as a method for supporting outbreak investigations, studying *Salmonella* microbial populations and improving understanding of pathogenicity has been well-described (1–3). However, performing WGS on a discrete dataset does not pose the same challenges as implementing WGS as a routine, reference microbiology service for public health surveillance. Challenges include translating WGS data into a useable format for laboratory reporting, clinical case management, *Salmonella* surveillance, and outbreak investigation as well as meeting the requirement to communicate that information in an understandable and universal language for clinical and public health action. Public Health England have been routinely sequencing all referred presumptive *Salmonella* isolates since 2014 which has transformed our approach to reference microbiology and surveillance. Here we describe an overview of the integrated methods for cross-disciplinary working, describe the challenges and provide a perspective on how WGS has impacted the laboratory and surveillance processes in England and Wales.

Keywords: WGS, genomic typing, molecular epidemiology, *Salmonella*, SNP typing

INTRODUCTION

Public Health England's (PHE) Gastrointestinal Bacterial Reference Unit (GBRU) receives approximately 10,000 presumptive *Salmonella* isolates each year from diagnostic microbiology laboratories, private laboratories and food, water and environmental laboratories for confirmation of identity and typing. Of the average 8,500 individual case reports of salmonellosis in England and Wales annually, ∼95% of clinical diagnostic isolates are sent to the reference laboratory for confirmation and further typing. The reporting of *Salmonella* isolated from human clinical diagnostic samples in public health laboratories is mandatory under national legislation (4, 5).

Prior to the introduction of WGS, presumptive *Salmonella* isolates were identified and characterized using a variety of methods including assaying biochemical properties (6), real-time PCR (7), phenotypic microarrays (Omnilog), and serology (8, 9). Further discrimination for select

serovars was routinely carried out using phage-typing (PT) (10) and suspected outbreak isolates were reactively subjected to pulsed-field gel electrophoresis (PFGE) (11) or multi-locus variable number of tandem repeats analysis (MLVA) (12). The approach of using multiple laboratory techniques for the characterization of *Salmonella* was highly specialized, laborious, time consuming and open to interpretation error. When the option of using a Whole Genome Sequencing (WGS) approach to streamline laboratory processes, reduce processing time, improve the fine typing discriminatory power for surveillance and outbreak detection in real-time became available, PHE utilized the opportunity to assess its potential in a public health setting.

In 2014, GBRU began evaluating and validating WGS methods as a replacement for conventional confirmation and further characterization methods for *Salmonella* spp and began reporting results derived from WGS analysis routinely for surveillance purposes from April 2015 (13). The implementation of this methodology has required a change in how we approach our testing processes, the reporting of microbiological data, the integration with epidemiological data and application of cross-disciplinary working encompassing microbiological, bioinformatics and epidemiological expertise. Here, following 4 full years of implementation in England and Wales, we describe an overview of our experiences to date, provide a perspective on our approach to maximize the utility and benefits, present on overview of WGS data generated between April 2016 and March 2018 and describe some of the limitations and challenges in implementing WGS for routine *Salmonella* surveillance.

PHES WGS IMPLEMENTATION APPROACH

Identification of *Salmonella* and the Bioinformatics Pipeline Process

Presumptive *Salmonella* isolates are submitted by frontline testing laboratories to the *Salmonella* Reference Service for confirmation and further characterization (**Figure 1**). On receipt the DNA is extracted using the Qiasymphony automated DNA extraction machine [Qiagen, UK] and sequenced using the Illumina HiSeq 2500 platform in rapid run mode (2 × 100 bp reads). The samples are batched with other pathogen isolates received for sequencing for the maximum capacity of 96 isolates per lane, per flowcell. The quality of raw FASTQ files is evaluated using an in-house program, qa_and_trim, which determines the metric yield of the sample (where yields of data from an isolate are below 150 Mb and are repeated) and trims the files using Trimmomatic (14) (using the parameters LEADING:30, TRAILING:30, SLIDINGWINDOW:10:20, and MINLEN:50). All subsequent analysis is carried out on the trimmed files. As previously described, the PHE KmerID pipeline (https://github.com/phe-bioinformatics/kmerid) is used to compare the sequenced reads with published genomes to identify the bacterial species and *Salmonella* subspecies (13). The quality of the sample is further evaluated by MLST using the Achtman seven gene scheme (15) (MOST, https://github.com/phe-bioinformatics/MOST) (16). Each sample is assigned a "traffic light" color depending on its coverage metrics: *Green*-maximum percentage non-consensus depth <15%, minimum consensus depth >2, percentage coverage = 100%, and that the ST determination has not failed; *amber*-maximum non-percentage consensus depth is ≥15% or minimum consensus depth is between 0 and 2 (inclusive); *red*-percentage coverage <100% or the ST determination has failed.

Salmonella serovar determination is predicted based on the *Salmonella* eBURST group (eBG) or Sequence Type (ST) (15) and checked against a validated PHE database (13). Validation of eBG and ST for inferring serovar is an ongoing process and currently requires a minimum of three isolates within that group to have been validated with the SeqSero profile (17) and confirmed with full phenotypic serology of both the somatic and flagella antigens (8, 9). Partial phenotypic serology is also currently performed when STs contain more than one serovar (polymorphic) or where referring primary diagnostic laboratories refer mixed cultures or they indicate conflicting serology results on the request form. To ensure reports are kept within TAT, where there are novel STs, the isolate is assigned an internal temporary ST until it has been submitted to a public repository and assigned a standard ST. The temporary ST is then overwritten with the new ST.

Microbial fine typing is achieved by utilizing the high discriminatory power of single nucleotide polymorphisms (SNP). A bioinformatics application, SnapperDB has been developed to quantify SNP relatedness and derive an isolate level nomenclature termed the "SNP Address" (18). This applies multi-threshold single linkage clustering to describe an isolate's position in the population structure of a given Salmonella eBG. Single-linkage clustering is performed at seven descending thresholds of SNP distance; 250, 100, 50, 25, 10, 5, and 0. This clustering results in a discrete seven-digit code where each number represents the cluster membership at each descending SNP distance threshold. Maximum likelihood phylogenies of selected strains of interest are constructed based on SNPs extracted from SnapperDB using RaxML v8.2.8 (19).

Turnaround times (TATs) before WGS averaged around 20 days from isolate receipt to reporting of validated results; Biochemistry—5–28 days, Serotyping—3–21 days, PT—3–10 days, PFGE—7–10 days. The average TAT for results utilizing WGS is now 10 days but these reports can be issued in as little as 6 days and can replace all of the previous methods. The reduced TAT and improvement of laboratory typing data has improved the outbreak investigation process since data is received quicker for analysis and case definitions have been refined and based on the enhanced granularity of the typing. The validation process for reporting laboratory results has remained the same with a two stage process involving the technical and medical validator checking the validity and quality metrics (such as the yield) of the WGS data and other performed tests for *Salmonella* identification. Participation in External Quality Assessment (EQA) schemes remain the same with the addition of specific EQAs now in place for cluster detection via genomic methods.

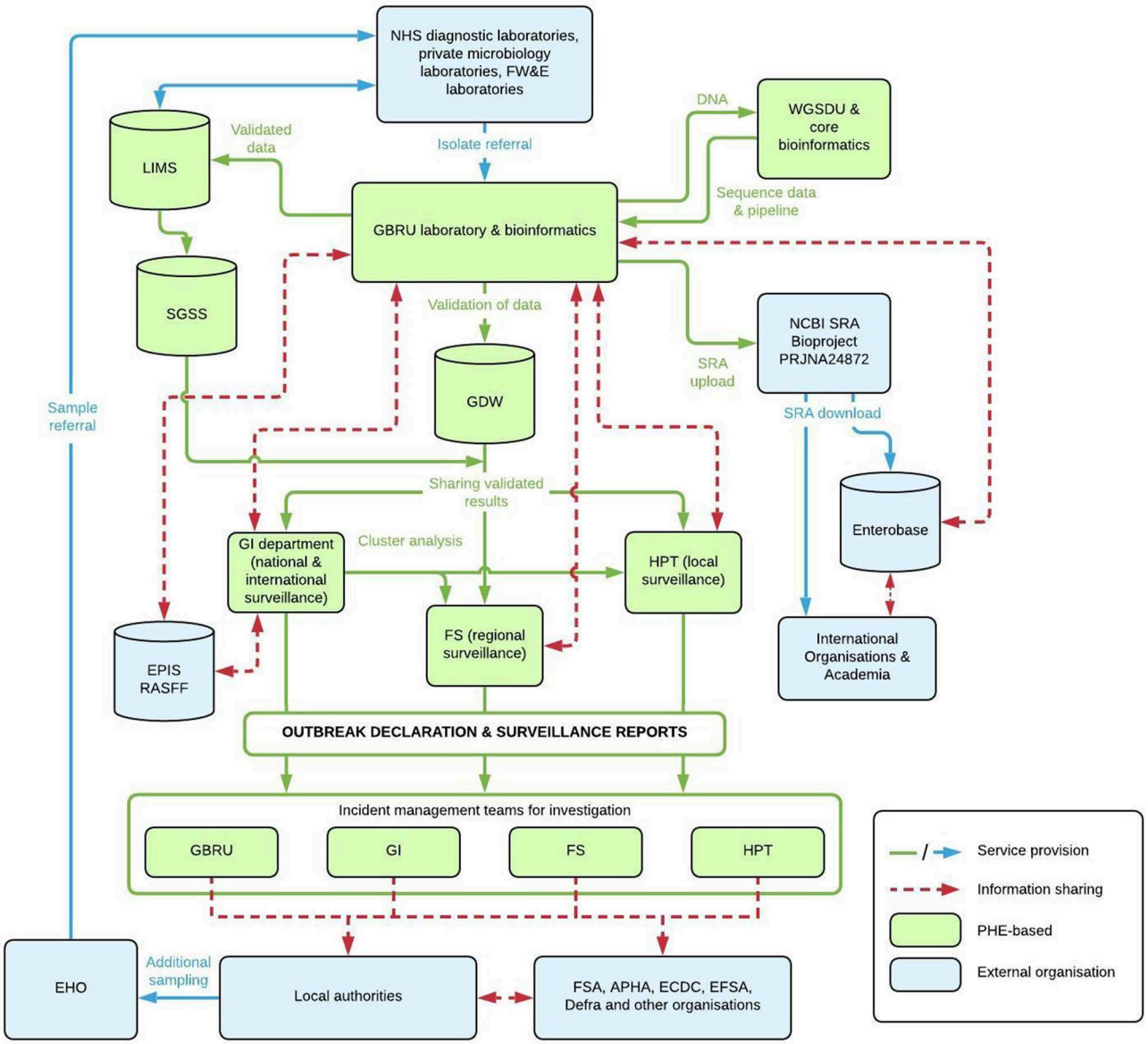


FIGURE 1 | Flow Chart of Service Provision and information workflows between PHE and external organizations for *Salmonella* reference microbiology and surveillance. NHS, National Health Service; FW&E, Food, Water and Environmental; PHE, Public Health England; GBRU, Gastrointestinal Bacteria Reference Unit; WGSDU, Whole Genome Sequencing Delivery Unit; LIMS, Laboratory Information Management System; GDW, Gastro Data Warehouse; SGSS, Second Generation Surveillance System; NCBI, National Center for Biotechnology Information; SRA, Short Read Archive; GI, Gastrointestinal; FS, Field Services; HPT, Health Protection Team; EPIS, Epidemic Intelligence Information System; RASFF, Rapid Alert System for Food and Feed; EHO, Environmental Health Officers; FSA, Food Standards Agency; APHA, Animal and Plant Health Agency; ECDC, European Center for Disease Prevention and Control; EFSA, European Food Safety Authority; DEFRA, Department for Environment, Food and Rural Affairs. Databases/Platforms include GDW, LIMS, EPIS, RASFF, and Enterobase.

Antimicrobial Resistance and Clinical Interpretation

Using WGS data, genetic antimicrobial resistance (AMR) determinants are sought using reference mapping approaches as previously described (20, 21). Resistance genes are identified by comparison to an in-house curated library collated from publicly accessible databases (PRJNA313047) (22, 23). Known chromosomal mutations, acquired resistance genes and resistance-conferring mutations relevant to β-lactams (including carbapenems), fluoroquinolones, aminoglycosides, chloramphenicol, macrolides, sulphonamides, tetracyclines, trimethoprim, and fosfomycin and acquired genes associated with colistin resistance are included in the reference database. Genotypic markers to infer phenotypic antimicrobial resistance have been recently validated (20, 21) but further work is required to translate this into a clinically useful format

(24). Phenotypic antimicrobial sensitivity testing (AST) are carried out to provide minimal inhibitory concentrations (MICs) (according to EUCAST guidelines http://www.eucast.org/clinical_breakpoints/). These are provided for clinical management where requested by diagnostic laboratories and a percentage of *Salmonella* are routinely phenotypically tested to check clinically important (e.g., bacteraemia or treatment failure cases) isolates and for horizon scanning purposes to detect novel and /or emerging mechanisms of resistance.

Reporting Results and Integrated Analysis of the Data

Frontline diagnostic laboratories report the isolation of *Salmonella* spp to PHE via the Second Generation Surveillance System (SGSS), a database that stores and manages data on laboratory isolates and results, and is the preferred method for capturing routine laboratory surveillance data on all infectious diseases and antimicrobial resistance from laboratories across England (25). This data is used for the monitoring of the overall number of *Salmonella* isolated at frontline laboratories and the number of isolates referred to GBRU. WGS results (ST, eBG, serovar, and SNP address) populate a Laboratory Information Management System (LIMS) at the *Salmonella* reference laboratory, where they are validated and reported to the sending clinician (**Figure 1**). The WGS data are currently only available via a restricted access web-based system, the Gastro Data Warehouse (GDW), a secure, encrypted, rationalized database containing results on all isolates processed by GBRU (**Figure 1**). PHE staff access data for cases within their region(s) on GDW via a web-enabled interface through which line-listings of case epidemiological data and sequencing results can be extracted based on case demographic and/or sequencing results, such as inferred serovar, ST, or SNP address. GDW also contains a cluster extraction functionality which allows users to search for SNP clusters based on desired temporal, size, and SNP distance level thresholds. This allows real-time surveillance of microbiological clusters by regional and national teams in line with the TAT stated above.

Routine surveillance and monitoring of *Salmonella* trends for general surveillance and risk assessment purposes is still carried out at the serovar level. SNP typing is routinely undertaken for the most commonly reported eBGs, and new eBGs/STs can be added to the routine pipeline as necessary; currently 86% of isolates received undergo SNP typing in real time. For those eBG not subject to SNP typing, the exceedance algorithm applied on the SGSS data is still used for outbreak detection at the serovar level (26). Where a potential outbreak event is detected, retrospective SNP typing of all the isolates within the ST/eBG is undertaken to refine outbreak detection and prospective SNP typing becomes routine. The SNP address is now utilized by PHE epidemiologists and microbiologists as the primary method for identifying microbiological clusters of gastrointestinal infections in England to detect potential outbreak events. Case isolates that fall within a 5-SNP single linkage cluster are considered likely to be exposed to a common source of contamination. The number of SNPs within a 5-SNP linkage cluster will vary depending on the size, type, source, and length of the outbreak. For example an international outbreak of *S.* Enteritidis, spanning over 3 years, had two distinct 5-SNP single linkage clusters even though they were from the same source of eggs from Poland. Cluster 1 had a maximum SNP distance of 18 SNPs whereas Cluster 2 had 37 SNPs (27). Validation studies (28) and prospective use in outbreak investigations (29, 30) indicate that the 5-SNP level is suitable for detection of salmonellosis cases that are likely to be epidemiologically linked and share a common exposure or source of infection.

In order to analyze and act on the data in real time in a systematic manner and manage the high volume of data generated by WGS, an automated reporting system, the "SNP Cluster Tool," has been developed using the statistical software R (31). The tool identifies and extracts epidemiological and sequencing data for clusters of two or more cases which cluster at the 5-SNP level where at least one case has been reported in the preceding week. Clusters are automatically summarized by rule-based categories in terms of case demographics (age, sex, geographic distribution, and travel history) and cluster-level characteristics (size, period of time since the first case was reported and cluster growth rate). The resultant summary tables are distributed on a weekly basis to microbiologists and epidemiologists working on *Salmonella* surveillance at the national and at the regional level. This automated approach facilitates rapid cluster assessment and prioritization of clusters requiring further investigation. The 5-SNP level is used primarily as an initial cluster extraction and assessment threshold but subsequent analysis of the cluster epidemiology and phylogeny may result in this threshold being extended as guided by the epidemiology. Where warranted this may even lead to the subsequent selection of more than one epidemiologically or phylogenetically related 5-SNP cluster to define the case definition for an outbreak investigation (29, 32). A key difference in defining SNP-clusters both microbiologically and epidemiologically compared to previous typing methods and epidemiological approaches is that the microbiological characterization is considered sufficiently discriminatory that clusters are usually defined independently of time. Therefore, in most national outbreaks we apply non time-limited, phylogeny-based case definitions and, in addition, no longer apply some traditional exclusion criteria such as travel history.

Phylogenetic trees are generated for clusters which have been prioritized for further assessment. Phylogenetic analysis provides insight into the genetic relationship between outbreak isolates which may reveal underlying epidemiological processes or sampling dynamics (33). In addition, phylogenetic context determined through assessing available epidemiological data for isolates related at a wider genetic threshold may assist hypothesis generation may assist hypothesis generation in terms of geographical origin or potential source. Phylodynamic reconstruction using Bayesian evolutionary analysis (34) may also be deployed in outbreak settings to estimate the temporal origin of the outbreak strain and to identify changes in population size over time. These approaches can be particularly valuable for outbreaks with long durations and where the assessment of the success of interventions is needed (27).

PHE also make validated FASTQ sequences publically available (**Figure 1**) by routinely uploading *Salmonella* sequence data to NCBI BioProject PRJNA248792 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA248792). Basic metadata is provided including the Month/Year, Country, Isolation source (e.g., human, animal, food), serovar and ST. As of 20th March 2019, 45,413 SRA experiments are available for analysis. Data from NCBI is routinely imported to Enterobase, so that other organizations can utilize its online tools such as analyzing population structures (**Figure 2**) or utilizing cgMLST tools and compare PHE genomes with their own data in outbreak detection. This enables any user to have access to the data for comparison analysis and has enabled real-time comparison of outbreaks at the international level.

Experiences and Outputs of WGS Implementation at PHE 2016–2018

WGS has not yet fully replaced traditional typing methods, a review of the 17,899 confirmed *Salmonella* laboratory results reported between April 2016 and March 2018 indicated that 89.1% of *Salmonella* serovars were reported by eBG/ST inference alone while the other 10.9% were reported on the antigenic phenotype (**Figure 3**).

Of the 17,899 reports, a total of 4,096 (22.8%) isolates required further microbiological tests including serology and PCR (**Figure 3**). The main reasons for additional serological testing included novel STs, mixed cultures referred by the sending laboratory and polymorphic *Salmonella* (more than one serovar within a ST) (**Figure 3**).

Out of the 17,899 isolates reported between April 2016- March 2018, 2,128 (11.8%) were tested phenotypically for AST (**Table 1**). There were no resistant *Salmonella* detected using phenotypic methods that were missed using WGS surveillance during this period, although results continue to show that genotypic AMR mutations do not always express phenotypically (20, 21). The use of WGS has enabled real-time, high throughput, routine surveillance of resistance determinants to detect emerging threats, such as the confirmation of the first ESBL *S.* Typhi case in the UK (35). A useful benefit of genotypic characterization of AMR determinants is the ability to rapidly add additional gene targets to the database, enabling rapid screening of thousands of isolates in a short period of time. In 2015, PHE demonstrated the use of WGS for rapid screening of the genomes of ∼24,000 *Salmonella enterica*, *E. coli*, *Klebsiella* spp., *Enterobacter* spp., *Campylobacter* spp. and *Shigella* spp. to identify novel transmissible colistin resistance (mcr-1) in 15 human and food isolates (36). Another example of utilizing WGS AMR data has been monitoring of emerging resistance to a first-line antibiotic azithromycin in *Salmonella* spp (37).

Since implementing WGS methods in April 2014, *Salmonella* reporting trends in England and Wales have been generally consistent with previous years. However, assessing laboratory data using eBG rather than serovar has shown that analysis of the data at the serovar level doesn't optimally reflect the incidence of genetically related groups. Assessment of eBGs reported between April 2016 and March 2018 shows that eBG 4 (*S.* Enteritidis, 4,866 isolates), eBG 1 (*S.* Typhimurium, 3,025 isolates) and eBG31 (*S.* Infantis,469 isolates) constitute the main burden of salmonellosis in England and Wales (**Figure 2**) as also reflected in analysis at the serovar level (5,240, 3,649, and 540 serovar reports, respectively). However, for polyphyletic serovars (serovars found in multiple eBGs), for example S. Newport, "rank" in terms of number of reports varies substantially when comparing the traditional serovar (671 isolates) to the multiple eBGs of which it is comprised. *S.* Newport was the third most commonly reported serovar between April 2016 and March 2018, however is comprised of multiple eBGs (eBG 2,3,7,35), with the most commonly reported *S.* Newport eBG (eBG3) being the 14th most commonly reported eBG (244 isolates) overall (**Figure 2**).

Of the 17,899 isolates reported from April 2016 to March 2018, 13,948 *Salmonella* isolates clustered with at least one other isolate at the 5-SNP level. These formed 2,007 clusters, distributed across 46 eBGs (**Table 2**). This time period was selected to identify the number of active clusters (i.e., the number of clusters with at least one new case added), however cluster statistics were analyzed using all cases with membership in the cluster regardless of when the result was reported. The majority of reported clusters were small, with only 29% of clusters constituting five or more cases (range: 2–423 cases, median: 3 cases). When these clusters were analyzed including all cases in the 5-SNP cluster, including those prior to March 2018, fifty-eight percent of clusters contained cases reported over a period of time exceeding 3 months (range: 0.03–115 months [linked to historical cases in these clusters], median: 6 months). Clusters of eBG4 (*S.* Enteritidis) constitute the majority of the longest duration clusters, and there is evidence gained from retrospective sequencing and analysis of isolates from 2008 to 2015 that an outbreak linked to feeder mice has persisted have persisted for over 10 years to date (38).

DISCUSSION

Improvement in Reference Services Including Diagnostics

Implementation of WGS has transformed reference microbiology services both in terms of improved accuracy of results (13), and reduced turnaround times by ∼50%. Further reduction of TATs is possible but we are currently limited by the requirement to batch process samples and the continuation of additional phenotypic work. As routine WGS is implemented for more organisms across PHE, the increase in numbers will enable increased sequencing runs and hence a reduction in TATs. The simplification of sample processing also reduces the potential for laboratory errors and minimizes staff exposure to pathogens thereby improving safety practices. In addition, we have utilized the sequence data generated through routine testing to develop specific, rapid real-time PCR tests to assist in the management of patients including for the rapid differential diagnoses of typhoidal from non-typhoidal *Salmonella* (39) and to detect azithromycin resistant infections (in house assay). This has had a direct clinical impact as same day testing can be provided for urgent clinical cases. It is also worth noting the rapidly developing technology of desktop and nanopore

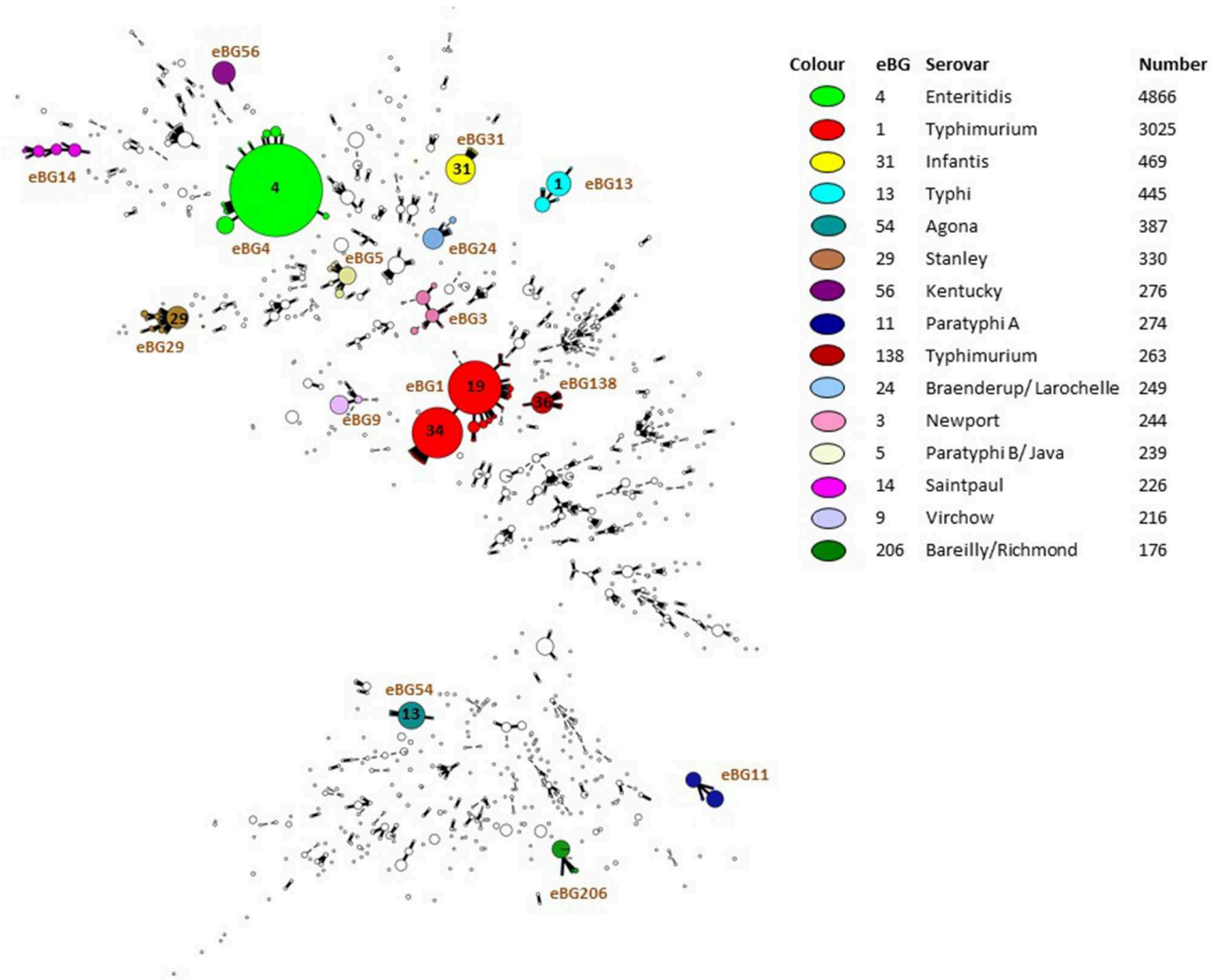


FIGURE 2 | Population structure of 16,854 *Salmonella* isolated from humans and submitted to PHE from local and regional hospital laboratories in England and Wales between April 2016 and March 2018.

sequencing becoming available to clinical laboratories. As these technologies become more affordable and common in clinical practice, real-time diagnostic sequencing will be able to identify pathogens, detect virulence factors and drug resistance markers to support clinical treatment. Currently local laboratories are legally required to notify PHE of the isolation of *Salmonella* sp. from a human sample; although further characterization is not mandated in the current legislation (4, 5). Fortunately, the majority (>95%) of isolated Salmonellae are currently sent to the reference laboratory for further typing to enable a robust national surveillance system. A move to sequencing occurring locally could pose a risk to a cohesive, representative national data set due to the lack of legal basis for such, though we think it likely that a system for sequence sharing would be set up to address this. However, even with the implementation of PCR which has been in place for over a decade, not all frontline laboratories use this technology. Benchtop sequencing is unlikely to have a large impact on the current reference services model in the short term with the current infrastructure in place.

Enhanced Surveillance and Outbreak Investigation

Although published evidence does not yet support the use of WGS-inferred antimicrobial susceptibility to guide clinical management of individual cases (24), studies have shown WGS to be an extremely rapid, robust, accurate tool for AMR surveillance in food-borne pathogens such as *Salmonella* spp. (20, 21). It is expected that information derived from WGS-based studies will increasingly be used to inform public health interventions aimed at limiting further dissemination of AMR genes in foodborne pathogens.

Considering the variability in eBG for some serovars (**Figure 2**), assessing *Salmonella* trends by eBGs, where available, may be more appropriate than by serovar, as differentiation by

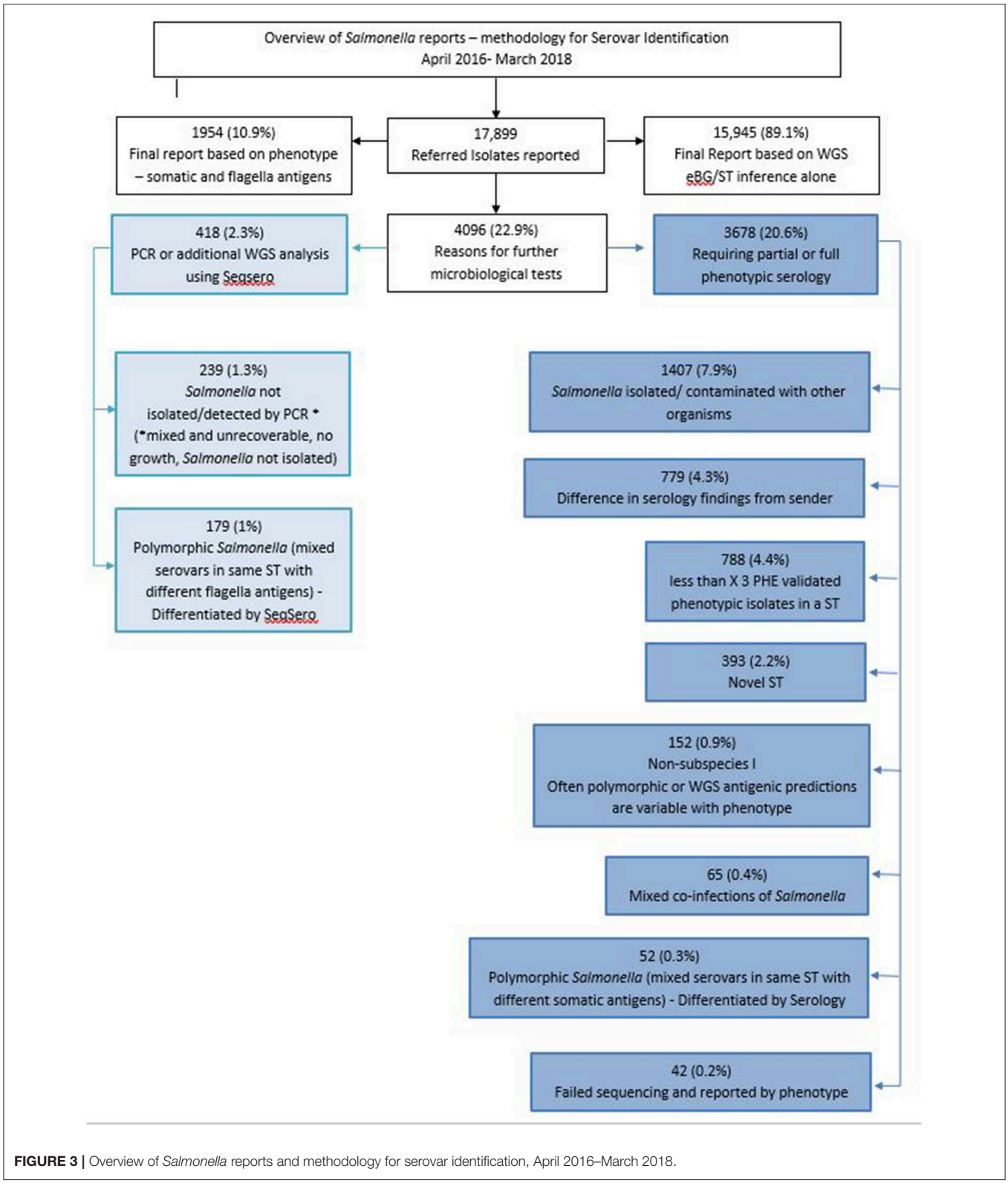


FIGURE 3 | Overview of *Salmonella* reports and methodology for serovar identification, April 2016–March 2018.

serovar does not optimally define the population heterogeneity to the level possible using eBG. Therefore, we are moving more to the use of eBG and in future eBG/ST for general surveillance, trend monitoring and outbreak detection based on exceedance algorithms. This work is still underway to integrate into routine surveillance systems.

TABLE 1 | Current criteria for selection of *Salmonella* isolates for phenotypic antimicrobial sensitivity testing by in-agar dilution.

Criteria	No. isolates tested 1 April 2016–31 March 2018
All *S.* Typhi isolates	457
All *S.* Paratyphi A isolates	284
All *S.* Paratyphi B isolates	36
All *S.* Paratyphi C/Choleraesuis and variants of 6,7:C:1,5 isolates	6
Non-typhoidal *Salmonella* (NTS) Bacteremia's	433
Invasive or complex NTS clinical cases (from patient sources other than feces and blood and by request).	103
Food, animal and environmental *Salmonella* isolates	161
From analysis of *Salmonella* sequencing data: all isolates that genotypically show the presence of one or more extended spectrum ß-lactamase genes	200
From analysis of *Salmonella* sequencing data: all isolates that genotypically show the presence of one or more extended spectrum ß-lactamase genes	163
From analysis of *Salmonella* sequencing data: all isolates that genotypically show the presence of one or more Carbapenamase resistance genes	1
From analysis of *Salmonella* sequencing data: all isolates that genotypically show the presence of two or more macrolide resistance genes	240
From analysis of *Salmonella* sequencing data: all isolates that genotypically show the presence of one or more colistin resistance genes	45
Total No. Isolates	2,128

Table summarizing the current criteria for phenotypic testing of isolates for antimicrobial resistance testing (AST) and the numbers tested for AST of the 17,899 isolates reported between April 2016 and March 2018. Note, not all resistance gene markers will express phenotypically. Numbers are the total tested, not necessarily the number with antimicrobial resistance.

The high-resolution typing provided by WGS for routine surveillance is facilitating the improved detection of smaller and geographically widespread clusters of common serovars such as *S.* Enteritidis and—especially for common strains. In these cases, the detection of a national outbreak would not have been possible without the use of WGS to delineate the outbreak strain from background numbers of commonly reported serovars/serovar and phage type combinations, and WGS can provide a much more refined case definition (38). Previous methods such as PT did not provide information on genotypic relationships and with common PTs, outbreak strains may have been overlooked particularly with ongoing outbreaks involving multiple PTs. In addition, cases have been epidemiologically investigated that were not genetically linked to the outbreak strain (38). Although, PFGE and PulseNet has been the backbone in the detection and sharing of outbreaks (https://www.cdc.gov/pulsenet/pathogens/pfge.html) on a global scale, there have been occasions where PFGE has not always been useful in detecting the same clone (40). The introduction of WGS in PHE and other agencies has enhanced the way we compare outbreak isolates and has facilitated an understanding of sources of outbreaks that would not have been possible with previous typing methods (30, 32, 33).

Data Accessibility and Integration of Cross Disciplinary Working

Key to the integration of epidemiology and phylogenetic information at PHE is data management and real time accessibility via the GDW database (**Figure 1**), as well as the SNP address nomenclature. The use of WGS generates a huge volume of data that requires further assessment by epidemiologists to determine if there is a need for action/outbreak investigation. The large amount of sequencing data generated for analysis each week necessitated the development of automated data extraction and analysis tools that have the capacity to deal with large amounts of data to aid rapid assessment and prioritization for further investigation. The sharing of the summary outputs of clusters and access to the WGS results integrated with basic case epidemiological data in a single database accessible by microbiologists, bioinformaticians and epidemiologists at the local, regional and national level means that local, regional and national teams are able to interpret fine typing microbiological data together with epidemiological data as part of routine surveillance, and target their investigations/resources where cases are most likely linked to a common source of contamination. A welcome consequence of implementing WGS has been closer working between public health infectious disease experts resulting in an enhanced, multidisciplinary approach to GI surveillance and outbreak investigation (**Figure 1**).

Inter-agency sharing and comparisons of microbiological, epidemiological, and food chain analysis results is necessary for effective food safety and control of zoonotic diseases at the UK and at the international level. The comparison of WGS results enhances effective assessment of cross-border threats and participation in multi-country outbreak investigations. Sharing raw sequence data, along with utilizing international information platforms supported by European Center for Disease Prevention and Control (ECDC) for the sharing of microbiological and epidemiological information, has proved successful for collaborative multi-agency, multi-country outbreak investigations (32, 33, 41).

Gaps, Limitations, and Future Work

As with any new system, there are limitations and there is room for improvement. A robust microbiological surveillance system depends upon high isolate referral rates, so, while there is currently high coverage for human diagnostic samples, there are laboratories (particularly in the private sector) that do not refer food isolates for further characterization. Consequently, crucial information from the food chain that could help inform hypothesis generation and target outbreak investigation and food chain analysis is being missed. Currently there is no system in place for routine sharing of animal data outside of outbreak investigations but PHE are addressing this together with the Animal and Plant Health Agency (APHA). In addition, the potential move to culture-independent diagnostic tests for GI pathogens by hospital laboratories threatens to reduce the

TABLE 2 | Characteristics of *Salmonella* WGS clusters, England, April 2016–March 2018.

eBG*	Serovar	Clusters	Cluster size			Cluster duration (months)			Cluster cases per week
			Median	Min	Max	Median	Min	Max	Max
4	Enteritidis	616	3.0	2	423	10.00	0.03	115.00	9.0
1	Typhimurium	606	3.0	2	165	4.00	0.03	74.00	5.0
31	Infantis	75	2.0	2	61	4.00	0.03	41.00	3.0
13	Typhi	67	3.0	2	112	12.00	0.10	72.00	2.0
11	Paratyphi A	59	3.0	2	36	15.00	0.13	64.00	2.0
29	Stanley	52	2.0	2	9	3.00	0.03	39.00	3.0
54	Agona	51	2.0	1	72	5.00	0.03	89.00	2.0
5	Java	45	2.0	2	13	5.00	0.03	47.00	2.0
56	Kentucky	37	2.0	2	28	8.00	0.03	40.00	2.0
9	Virchow	35	2.0	2	38	13.00	0.07	46.00	2.0
22	Hadar	34	2.5	2	17	5.00	0.07	41.00	3.0
138	Typhimurium	34	2.0	2	8	0.73	0.03	28.00	3.5
24	Braenderup	30	2.5	2	69	7.00	0.03	49.00	3.5
206	Bareilly	30	2.0	2	27	5.00	0.03	39.00	2.0
3	Newport	25	2.0	2	20	2.00	0.03	38.00	2.0
7	Newport	22	2.5	2	16	1.50	0.03	34.00	2.0
34	Bovis morbificans	18	2.5	2	24	6.00	0.10	38.00	2.5
62	Mbandaka	17	2.0	2	5	7.00	0.03	35.00	2.0
247	Mikawasima	16	3.0	2	16	1.00	0.16	24.00	4.0
44	Oranienburg	13	4.0	2	19	15.00	0.49	43.00	1.0
49	Chester	13	3.0	2	32	19.00	0.03	42.00	2.0
2	Newport	12	3.0	2	45	7.00	0.03	42.00	2.0
35	Newport	10	2.0	2	5	3.00	0.20	27.00	1.5
41	Oranienburg	10	3.0	2	29	2.00	0.03	17.00	2.0
65	Anatum	9	2.0	2	3	0.72	0.03	15.00	2.0
205	Weltevreden	7	2.0	2	7	0.66	0.03	13.00	2.0
12	Brandenburg	6	3.0	2	5	4.00	0.16	10.00	2.0
64	Kottbus	6	2.5	2	7	0.64	0.20	12.00	1.5
17	Javiana	5	3.0	2	4	7.00	0.03	35.00	2.0
61	Litchfield	5	2.0	2	4	5.00	1.00	37.00	1.0
70	Virchow	5	3.0	2	7	1.00	0.10	2.00	2.0
164	Kentucky	5	2.0	2	3	12.00	0.16	29.00	1.0
26	Heidelberg	4	2.0	2	3	0.71	0.03	3.00	2.0
32	Java	4	2.0	2	3	0.29	0.03	2.00	1.0
67	Give	4	9.0	2	17	10.00	0.03	17.00	1.5
271	Indiana	4	7.0	2	28	9.00	0.03	17.00	2.0
291	Kedougou	4	3.0	2	50	15.50	0.36	27.00	2.0
421	Adjame	3	5.0	4	7	0.69	0.23	1.00	1.0
270	Liverpool	2	4.0	3	5	6.34	0.69	12.00	2.0
292	Agbeni	2	3.5	2	5	10.00	1.00	19.00	1.0
57	Derby	1	2.0	2	2	12.00	12.00	12.00	1.0
244	Derby	1	20.0	20	20	40.00	40.00	40.00	1.0
264	Derby	1	5.0	5	5	30.00	30.00	30.00	1.0
1483	Abony	1	4.0	4	4	28.00	28.00	28.00	1.0
1992	Carno	1	12.0	12	12	5.00	5.00	5.00	1.0

*Table describing the characteristics of Salmonella whole sequencing genome clusters in order of decreasing cluster burden by eBG (*where the eBG is not defined, the ST will be specified). Each eBG is characterized in terms of the number of clusters detected during the surveillance period (with the data for each cluster including isolates falling into the clusters outside the study period), the number of cases within clusters, the age of the cluster in months and the number of cases detected per cluster per week. The maximum cluster duration will date back to any historical strains that have been sequenced and fall into the cluster.*

representativeness of WGS data as isolates would not always available for sequencing.

Although a small number of isolates are still being fully phenotypically serotyped due to validation of novel STs (**Figure 3**), *in silico* serotyping methods such as SeqSero (17) or SISTR (42) hold great promise in providing a direct replacement for prediction of individual somatic and flagella antigens, as currently defined by the Kaufmann-White-Le Minor scheme. It should be noted however that genotypic prediction does not always correlate to phenotypic expression which is problematic for defining novel *Salmonella* strains. We recognize that continuing to perform phenotypic serology routinely is not desirable or sustainable and we aim to cease all traditional serotyping methods in future.

Additional limitations include the necessity of pure cultures required for DNA extraction as contamination will interfere with bioinformatic outputs including accurate sequence typing, fine typing results of SNP analysis and correct calling of AMR gene determinants. Batch processing of samples is still required for sequencing to improve efficiency and maintain cost-effective operations; as a result, TATs are typically in excess of 7 days and in urgent typhoidal cases, PCR (39) is still required to provide a preliminary identification.

Recent publications (20, 21) have demonstrated the utility of WGS-inferred antimicrobial susceptibility for clinical management, rapid surveillance initiatives and monitoring of emerging resistance. It is acknowledged that novel mechanisms of resistance could be missed using genotypic determination of AMR and how the presence of AMR determinants relates to MICs is as yet still not fully understood, therefore a certain level of phenotypic testing is still required. MIC prediction by WGS and machine learning is currently being investigated (43), where the observed MIC is underpinned by genetic factors encoded in the DNA, prediction should be possible and a potential model for the future. It is crucial to perform active curation of the resistance gene databases to maintain the high sensitivity of genotypic prediction especially due to novel, emerging resistance mechanisms. Our in-house pipeline, for instance, does not detect impermeability or efflux pumps as these mechanisms are not always encoded by a single gene that can be easily detected.

The SNP address derived from the PHE pipeline has been utilized to identify microbiologically linked cases through collaborative working and sharing of sequence data in international outbreak investigations. However, there are multiple different pipelines and nomenclatures used in different organizations, so WGS results may not always be easily communicated between agencies using different systems in the initial stages of detection and assessment of threats. Real-time multi-country comparison of WGS data remains challenging, and the future use of harmonized typing schemes and supporting infrastructure is welcomed (44, 45) and validation studies have already begun (46). One example is the NCBI Pathogen Detection Portal (https://www.ncbi.nlm.nih.gov/pathogens) and is a working example of close to real-time comparison system for surveillance of bacterial pathogens using WGS. There are multiple caveats, such as making the data public and being able to interpret phylogenetic trees but this approach does work and an open framework for all to access.

The high volume of clusters detected each week and longevity of some clusters due to persistent sources of contamination can be challenging in terms of consistent resource allocation. A high-level of expertise is required to interpret WGS data in combination with epidemiological evidence.

CONCLUSION

The Whole Is More Than the Sum of Its Parts

The integration of routine WGS as a replacement for traditional microbiological methods has revolutionized reference microbiology and impacted real-time surveillance of gastrointestinal pathogens for improved public health outcomes. PHE have now implemented routine WGS methods for *Salmonella* (13), *Shigella* (47, 48), *Campylobacter*, *Escherichia* (48, 49), *Listeria* (50), *Vibrio* (51), and *Yersinia* species (52). It is envisioned that WGS methods will be implemented for all gastrointestinal bacterial pathogens services at PHE within the next few years.

The large volume of data generated by the use of WGS has required additional tools be developed to facilitate surveillance, cluster assessment and prioritization, and outbreak detection; using these tools these processes have become more discriminatory and can occur in near real-time compared to previous typing methodologies. This has improved outbreak detection, hypothesis generation, and source attribution in ways not previously possible.

The posting of sequences on a publicly accessible database means other countries can compare with their in-house databases and has facilitated substantial international collaboration that would not have possible if all data was only kept in-house.

International harmonization of WGS typing methods for surveillance is crucial and still in the development phase. Close collaboration between epidemiologists, bioinformaticians, microbiologists, clinicians and food safety experts is essential to maximize the public health potential provided by WGS.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article. In addition, raw sequence data described in this article is publically available on NCBI, PHE Salmonella Bioproject: PRJNA248792.

AUTHOR CONTRIBUTIONS

MC, SN, and MD implemented the wet lab WGS pipelines, performed analysis, and identification. MD,

MC, and GG performed AST identification and reporting. TD and HH performed bioinformatic analysis. LL, JM, AM, and RS performed cluster analysis and epidemiological investigations. DP performed data analysis. MC and KG wrote the manuscript. TD, LL, SN, JM, AM, HH, GG, DP, MD, and RS contributed to the manuscript.

FUNDING

This study was funded by Public Health England and the National Institute for Health Research Health Protection Research Unit (NIHR HPRU) (NIHR HPRU-2012-10038). The views expressed are those of the author (s) and not necessarily those of the NHS, the NIHR, the Department of Health or Public Health England.

ACKNOWLEDGMENTS

All of the diagnostic laboratories and the food, water and environmental laboratories of isolating and referring isolates to PHE. Thank you to Clare Maguire, Andrew Levy, and Anais Painset from GBRU for their support with the reference service. We would also like to thanks Cath Arnold and her team at the Whole Genome Sequencing Delivery Unit and Jonathan Green and his team at the Bioinformatics Unit at PHE for their support.

REFERENCES

1. Taylor AJ, Lappi V, Wolfgang WJ, Lapierre P, Palumbo MJ, Medus C, et al. Characterization of foodborne outbreaks of *Salmonella enterica* serovar enteritidis with whole-genome sequencing single nucleotide polymorphism-based analysis for surveillance and outbreak detection. *J Clin Microbiol.* (2015) 53:3334–40. doi: 10.1128/JCM.01280-15
2. Wuyts V, Denayer S, Roosens NH, Mattheus W, Bertrand S, Marchal K, et al. Whole genome sequence analysis of *Salmonella* Enteritidis PT4 outbreaks from a national reference laboratory's viewpoint. *PLoS Curr.* (2015) 7:1–14. doi: 10.1371/currents.outbreaks.aa5372d90826e6cb0136ff66bb7a62fc
3. Thomas M, Fenske GJ, Antony L, Ghimire S, Welsh R, Ramachandran A, et al. Whole genome sequencing-based detection of antimicrobial resistance and virulence in non-typhoidal *Salmonella enterica* isolated from wildlife. *Gut Pathog.* (2017) 9:66. doi: 10.1186/s13099-017-0213-x
4. England PH. *The Health Protection (Notification) Regulations 2010.* Public Health England, The Stationery Office Limited (2010). p. 659.
5. Wales PH. *The Health Protection (Notification) (Wales) Regulations 2010.* Public Health England, The Stationery Office Limited (2010). p. 1546.
6. Cowan S, Steel T, Barrow GI, Feltham RKA. *Cowan and Steel's Manual for the Identification of Medical Bacteria*. Cambridge: Cambridge University Press (1993).
7. Hopkins KL, Peters TM, Lawson AJ, Owen RJ. Rapid identification of *Salmonella enterica* subsp. arizonae and *S enterica* subsp diarizonae by real-time polymerase chain reaction. *Diagn Microbiol Infect Dis.* (2009) 64:452–4. doi: 10.1016/j.diagmicrobio.2009.03.022
8. Grimont PADWFX. *Antigentic Formulae of the Salmonella Serovars*. Institut Pasteur: WHO Collaborating Centre for Reference and Research on Salmonella (2008).
9. Guibourdenche M, Roggentin P, Mikoleit M, Fields PI, Bockemuhl J, Grimont PA, et al. Supplement 2003-2007 (No. 47) to the White-Kauffmann-Le Minor scheme. *Res Microbiol.* (2010) 161:26–29. doi: 10.1016/j.resmic.2009.10.002
10. Callow BR. A new phage-typing scheme for *Salmonella* Typhi-murium. *J Hyg.* (1959) 57:346–59. doi: 10.1017/S0022172400020209
11. Peters TM. Pulsed-field gel electrophoresis for molecular epidemiology of food pathogens. *Methods Mol Biol.* (2009) 551:59–70. doi: 10.1007/978-1-60327-999-4_6
12. Hopkins KL, Peters TM, de Pinna E, Wain J. Standardisation of multilocus variable-number tandem-repeat analysis (MLVA) for subtyping of *Salmonella enterica* serovar Enteritidis. *Euro Surveill.* (2011) 16:19942. doi: 10.2807/ese.16.32.19942-en
13. Ashton PM, Nair S, Peters TM, Bale JA, Powell DG, Painset A, et al. Identification of *Salmonella* for public health surveillance using whole genome sequencing. *PeerJ.* (2016) 4:e1752. doi: 10.7717/peerj.1752
14. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* (2014) 30:2114–20. doi: 10.1093/bioinformatics/btu170
15. Achtman M, Wain J, Weill FX, Nair S, Zhou Z, Sangal V, et al. Multilocus sequence typing as a replacement for serotyping in *Salmonella enterica*. *PLoS Pathog.* (2012) 8:e1002776. doi: 10.1371/journal.ppat.1002776
16. Tewolde R, Dallman T, Schaefer U, Sheppard CL, Ashton P, Pichon B, et al. MOST: a modified MLST typing tool based on short read sequencing. *PeerJ.* (2016) 4:e2308. doi: 10.7717/peerj.2308
17. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, et al. *Salmonella* serotype determination utilizing high-throughput genome sequencing data. *J Clin Microbiol.* (2015) 53:1685–92. doi: 10.1128/JCM.00323-15
18. Dallman T, Ashton P, Schafer U, Jironkin A, Painset A, Shaaban S, et al. SnapperDB: a database solution for routine sequencing analysis of bacterial isolates. *Bioinformatics.* (2018) 34:3028–29. doi: 10.1093/bioinformatics/bty212
19. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* (2014) 30:1312–3. doi: 10.1093/bioinformatics/btu033
20. Day MR, Doumith M, Do Nascimento V, Nair S, Ashton PM, Jenkins C, et al. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Salmonella* enterica serovars Typhi and Paratyphi. *J Antimicrob Chemother.* (2017) 73:365–72. doi: 10.1093/jac/dkx379
21. Neuert S, Nair S, Day MR, Ashton PM, Mellor KC, Jenkins C, et al. Prediction of phenotypic antimicrobial resistance profiles from whole genome sequences of non-typhoidal *Salmonella* enterica. *Front Microbiol.* (2018) 9:592. doi: 10.3389/fmicb.2018.00592
22. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. *J Antimicrob Chemother.* (2012) 67:2640–4. doi: 10.1093/jac/dks261
23. Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M, Peto T, et al. A curated dataset of complete *Enterobacteriaceae* plasmids compiled from the NCBI nucleotide database. *Data Brief.* (2017) 12:423–6. doi: 10.1016/j.dib.2017.04.024
24. Ellington MJ, Ekelund O, Aarestrup FM, Canton R, Doumith M, Giske C, 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. doi: 10.1016/j.cmi.2016.11.012
25. England PH. *Laboratory Reporting to Public Health England: A Guide for Diagnostic Laboratories*. PHE publications gateway number: 2016137. London: Public Health England (2016).
26. Noufaily A, Farrington P, Garthwaite P, Enki DG, Andrews N, Charlett A. *Detection of Infectious Disease Outbreaks From Laboratory Data With Reporting Delays*. Open University (2016).
27. Pijnacker R, Dallman TJ, Tijsma ASL, Hawkins G, Larkin L, Kotila SM, et al. An international outbreak of Salmonella enterica serotype Enteritidis linked to eggs from Poland: a microbiological and epidemiological study. *Lancet Infect Dis.* (2019) 19:778–86. doi: 10.1016/S1473-3099(19)30047-7
28. Waldram A, Dolan G, Ashton PM, Jenkins C, Dallman TJ. Epidemiological analysis of *Salmonella* clusters identified by whole genome

sequencing, England and Wales 2014. *Food Microbiol.* (2018) 71:39–45. doi: 10.1016/j.fm.2017.02.012
29. EFSA Ea. *European Centre for Disease Prevention and Control and European Food Safety Authority: Multi-Country Outbreak of Salmonella Enteritidis Phage Type 8, MLVA Profile 2-9-7-3-2 and 2-9-6-3-2 Infections*. Stockholm; Parma (2017).
30. Inns T, Ashton PM, Herrera-Leon S, Lighthill J, Foulkes S, Jombart T, et al. Prospective use of whole genome sequencing (WGS) detected a multi-country outbreak of *Salmonella* Enteritidis. *Epidemiol Infect.* (2017) 145:289–98. doi: 10.1017/S0950268816001941
31. Team RC. *A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing (2017).
32. Inns T, Lane C, Peters T, Dallman T, Chatt C, McFarland N, et al. A multi-country *Salmonella* Enteritidis phage type 14b outbreak associated with eggs from a German producer: 'near real-time' application of whole genome sequencing and food chain investigations, United Kingdom, May to September 2014. *Euro Surveill.* (2015) 20:21098. doi: 10.2807/1560-7917.ES2015.20.16.21098
33. Dallman T, Inns T, Jombart T, Ashton P, Loman N, Chatt C, et al. Phylogenetic structure of European *Salmonella* Enteritidis outbreak correlates with national and international egg distribution network. *Microb Genom.* (2016) 2:e000070. doi: 10.1099/mgen.0.000070
34. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol.* (2012) 29:1969–73. doi: 10.1093/molbev/mss075
35. Godbole GS, Day MR, Murthy S, Chattaway MA, Nair S. First report of CTX-M-15 *Salmonella* Typhi from England. *Clin Infect Dis.* (2018) 66:1976–77. doi: 10.1093/cid/ciy032
36. Doumith M, Godbole G, Ashton P, Larkin L, Dallman T, Day M, et al. Detection of the plasmid-mediated mcr-1 gene conferring colistin resistance in human and food isolates of Salmonella enterica and *Escherichia coli* in England and Wales. *J Antimicrob Chemother.* (2016) 71:2300–5. doi: 10.1093/jac/dkw093
37. Nair S, Ashton P, Doumith M, Connell S, Painset A, Mwaigwisya S, et al. WGS for surveillance of antimicrobial resistance: a pilot study to detect the prevalence and mechanism of resistance to azithromycin in a UK population of non-typhoidal Salmonella. *J Antimicrob Chemother.* (2016) 71:3400–8. doi: 10.1093/jac/dkw318
38. Kanagarajah S, Waldram A, Dolan G, Jenkins C, Ashton PM, Carrion Martin AI, et al. Whole genome sequencing reveals an outbreak of Salmonella Enteritidis associated with reptile feeder mice in the United Kingdom, 2012-2015. *Food Microbiol.* (2018) 71:32–8. doi: 10.1016/j.fm.2017.04.005
39. Nair S, Patel V, Hickey T, Maguire C, Greig DR, Lee W, et al. Real-time PCR assay for differentiation of typhoidal and nontyphoidal salmonella. *J Clin Microbiol.* (2019) 57:e00167-19. doi: 10.1128/JCM.00167-19
40. Scaltriti E, Sassera D, Comandatore F, Morganti M, Mandalari C, Gaiarsa S, et al. Differential single nucleotide polymorphism-based analysis of an outbreak caused by Salmonella enterica serovar Manhattan reveals epidemiological details missed by standard pulsed-field gel electrophoresis. *J Clin Microbiol.* (2015) 53:1227–38. doi: 10.1128/JCM.02930-14
41. Authority EFS, Prevention ECFD, Control. *Multi-Country Outbreak of Salmonella Enteritidis Infections Linked to Polish Eggs*. EFSA Supporting Publications (2017). p. 1353E
42. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VP, Nash JH, et al. 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. doi: 10.1371/journal.pone.0147101
43. Nguyen M, Long SW, McDermott PF, Olsen RJ, Olson R, Stevens RL, et al. Using machine learning to predict antimicrobial MICs and associated genomic features for nontyphoidal *Salmonella*. *J Clin Microbiol.* (2019) 57:e01260-18. doi: 10.1128/JCM.01260-18
44. Nadon C, Van Walle I, Gerner-Smidt P, Campos J, Chinen I, Concepcion-Acevedo J, et al. PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Eurosurveillance.* (2017) 22:30544. doi: 10.2807/1560-7917.ES.2017.22.23.30544
45. Alikhan NF, Zhou Z, Sergeant MJ, Achtman M. A genomic overview of the population structure of Salmonella. *PLoS Genet.* (2018) 14:e1007261. doi: 10.1371/journal.pgen.1007261
46. Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MCJ. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. *Int J Food Microbiol.* (2018) 274:1–11. doi: 10.1016/j.ijfoodmicro.2018.02.023
47. Chattaway MA, Greig DR, Gentle A, Hartman HB, Dallman TJ, Jenkins C. Whole-genome sequencing for national surveillance of *Shigella flexneri*. *Front Microbiol.* (2017) 8:1700. doi: 10.3389/fmicb.2017.01700
48. Chattaway MA, Schaefer U, Tewolde R, Dallman TJ, Jenkins C. Identification of *Escherichia coli and Shigella*. species from whole-genome sequences. *J Clin Microbiol.* (2017) 55:616–23. doi: 10.1128/JCM.01790-16
49. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, et al. Whole genome sequencing for public health surveillance of Shiga toxin-producing *Escherichia coli* other than serogroup O157. *Front Microbiol.* (2016) 7:258. doi: 10.3389/fmicb.2016.00258
50. Elson R, Awofisayo-Okuyelu A, Greener T, Swift C, Painset A, Amar CFL, et al. Utility of whole genome sequencing to describe the persistence and evolution of *Listeria monocytogenes* strains within crabmeat processing environments linked to two outbreaks of listeriosis. *J Food Prot.* (2019) 82:30–8. doi: 10.4315/0362-028X.JFP-18-206
51. Greig DR, Schaefer U, Octavia S, Hunter E, Chattaway MA, Dallman TJ, et al. Evaluation of whole-genome sequencing for identification and typing of *Vibrio cholerae*. *J Clin Microbiol.* (2018) 56:e00831-18. doi: 10.1128/JCM.00831-18
52. Inns T, Flanagan S, Greig DR, Jenkins C, Seddon K, Chin T, et al. First use of whole-genome sequencing to investigate a cluster of *Yersinia enterocolitica*, Liverpool, United Kingdom, 2017. *J Med Microbiol.* (2018) 67:1747–52. doi: 10.1099/jmm.0.000856

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

REVIEW
published: 11 December 2019
doi: 10.3389/fpubh.2019.00360

Direct Sequencing of *Cryptosporidium* in Stool Samples for Public Health

Arthur Morris[1†], *Guy Robinson*[2,3†], *Martin T. Swain*[1] *and Rachel M. Chalmers*[2,3*]

[1] *Institute of Biological, Environmental & Rural Sciences, Aberystwyth University, Aberystwyth, United Kingdom,*
[2] *Cryptosporidium Reference Unit, Public Health Wales Microbiology, Singleton Hospital, Swansea, United Kingdom,*
[3] *Swansea University Medical School, Swansea, United Kingdom*

The protozoan parasite *Cryptosporidium* is an important cause of diarrheal disease (cryptosporidiosis) in humans and animals, with significant morbidity and mortality especially in severely immunocompromised people and in young children in low-resource settings. Due to the sexual life cycle of the parasite, transmission is complex. There are no restrictions on sexual recombination between sub-populations, meaning that large-scale genetic recombination may occur within a host, potentially confounding epidemiological analysis. To clarify the relationships between infections in different hosts, it is first necessary to correctly identify species and genotypes, but these differentiations are not made by standard diagnostic tests and more sophisticated molecular methods have been developed. For instance, multilocus genotyping has been utilized to differentiate isolates within the major human pathogens, *Cryptosporidium parvum* and *Cryptosporidium hominis*. This has allowed mixed populations with multiple alleles to be identified: recombination events are considered to be the driving force of increased variation and the emergence of new subtypes. As yet, whole genome sequencing (WGS) is having limited impact on public health investigations, due in part to insufficient numbers of oocysts and purity of DNA derived from clinical samples. Moreover, because public health agencies have not prioritized parasites, validation has not been performed on user-friendly data analysis pipelines suitable for public health practitioners. Nonetheless, since the first whole genome assembly in 2004 there are now numerous genomes of human and animal-derived cryptosporidia publically available, spanning nine species. It has also been demonstrated that WGS from very low numbers of oocysts is possible, through the use of amplification procedures. These data and approaches are providing new insights into host-adapted infectivity, the presence and frequency of multiple sub-populations of *Cryptosporidium* spp. within single clinical samples, and transmission of infection. Analyses show that although whole genome sequences do indeed contain many alleles, they are invariably dominated by a single highly abundant allele. These insights are helping to better understand population structures within hosts, which will be important to develop novel prevention strategies in the fight against cryptosporidiosis.

Keywords: cryptosporidium, public health, genotyping, genome, sequencing, multiplicity of infection

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control, Sweden

Reviewed by:
Michael Arrowood, Centers for Disease Control and Prevention, United States
Lihua Xiao, South China Agricultural University, China
Michelle Power, Macquarie University, Australia

***Correspondence:**
Rachel M. Chalmers
Rachel.Chalmers@wales.nhs.uk

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Infectious Diseases - Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: *31 July 2019*
Accepted: *13 November 2019*
Published: *11 December 2019*

Citation:
Morris A, Robinson G, Swain MT and Chalmers RM (2019) Direct Sequencing of Cryptosporidium in Stool Samples for Public Health. Front. Public Health 7:360. doi: 10.3389/fpubh.2019.00360

INTRODUCTION

The parasite *Cryptosporidium* is a protozoan that occurs worldwide, and can cause the diarrheal disease cryptosporidiosis in humans and animals (**Figure 1**). The life cycle of *Cryptosporidium* (**Figure 2a**) (1) is completed within a single host. Both the asexual phase, and the production of thin-walled oocysts that enable autoinfection, mean the numbers of parasites are increased from possibly single figures in the initial infection, to result in clinically significant infections and the shedding of vast numbers of oocysts in feces (2). These shed oocysts have thick walls, conferring protection for the four infective sporozoites contained within, and enabling long-term survival, environmental transmission, and resistance to commonly used disinfectants including chlorine (3, 4). This means that, in addition to the variety of hosts that act as direct sources of infection (**Figure 1**; **Table 1**), contaminated food, water, or environmental vehicles are involved in transmission and need to be considered and investigated for effective disease control and prevention of outbreaks of cryptosporidiosis (5).

Human cryptosporidiosis is usually a gastrointestinal disease, although there is some evidence for respiratory cryptosporidiosis in some populations (6). Symptoms ranging from mild to severe depending upon a number of factors, including the host's age, immune status, nutrition, genetics, and the site of infection, as well as the infecting species and variant of *Cryptosporidium* (7–9). Clinical symptoms include diarrhea, abdominal pain, vomiting, nausea, and low-grade fever, which, although prolonged (2 weeks is not unusual) are generally self-limiting in immune competent hosts. However, infection can be more problematic and even life-threatening in some severely immunocompromised individuals, and in malnourished young children (10). There are few options for treatment or prevention. Recent studies have shown that in some low-resource countries, where access to safe drinking water, sanitation, hygiene, and healthcare is often poor, *Cryptosporidium* is one of the most important causes of moderate-to-severe diarrheal disease and death in young children (11, 12). Furthermore, long-term effects of infection such as malnutrition, growth, and cognitive deficits have been described, highlighting the socio-economic impact on the adverse outcomes of infection (10). A vicious cycle of malnutrition and diarrhea can become established with detrimental effects on these societies (13). For these reasons, *Cryptosporidium* was included in the World Health Organization's Neglected Diseases Initiative in 2004 (14), which served to raise awareness of the need for international and national investments in prevention and control.

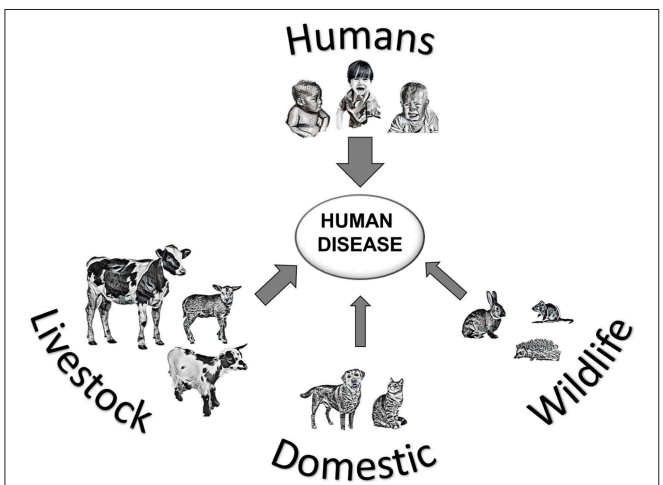


FIGURE 1 | Transmission of *Cryptosporidium* spp. leading to human cryptosporidiosis, arrow thickness represents likely global importance of source hosts.

Thirty-nine species of *Cryptosporidium* have been described at the time of writing (**Table 1**), but not all cause human disease. The vast majority of human cryptosporidiosis is caused by the zoonotic species *Cryptosporidium parvum* or anthroponotic *Cryptosporidium hominis,* with multiple variants that can cause varying severity of symptoms. The diagnostic target of laboratory tests, and those used to detect *Cryptosporidium* in water, is the oocyst, using stained microscopy or immunologically-based assays, or the sporozoite DNA. Routinely applied tests are not able to differentiate species, and molecular methods are needed to investigate true relationships between infections and contaminants and thus elucidate the complex transmission of *Cryptosporidium*. A range of samples need to be investigated, from feces (e.g., stools, diapers, livestock dung, manure, slurry, runoff, and wild life droppings), to contaminated water and food, but these present challenges to detection and genotyping. At present, amplification by culture is not an option in this context, and finding oocyst targets, which may be in low concentration in the sample matrix, can be a hit-and-miss affair. Recent advances in molecular methods generally, and particularly in genomics, have increased the amount of data available particularly on the major pathogenic *Cryptosporidium* species (**Table 1**). Continued generation and accessibility of genomic data will potentially improve the public health response to cryptosporidiosis by identifying new targets for incorporation into diagnostic and genotyping assays (15). Putative virulence and host adaption factors have been proposed (16), and potential chemotherapeutic targets and vaccine candidates are being sought (10, 17) and identified [e.g., (18)].

INTRODUCTION TO *CRYPTOSPORIDIUM* GENOTYPING

To identify *Cryptosporidium* species, genotyping was undertaken initially using conventional PCR combined with either restriction fragment length polymorphism (RFLP) or Sanger sequence analysis, most commonly of the 18S rRNA gene (19). The 18S rRNA gene includes conserved regions interspersed with highly polymorphic regions and is currently considered to provide the definitive sequences for discriminating *Cryptosporidium* species. It is present in multiple copies (5 per sporozoite; 20 per oocyst) facilitating the development of sensitive assays, which is especially important for testing samples such as water where small (but potentially significant) numbers of oocysts

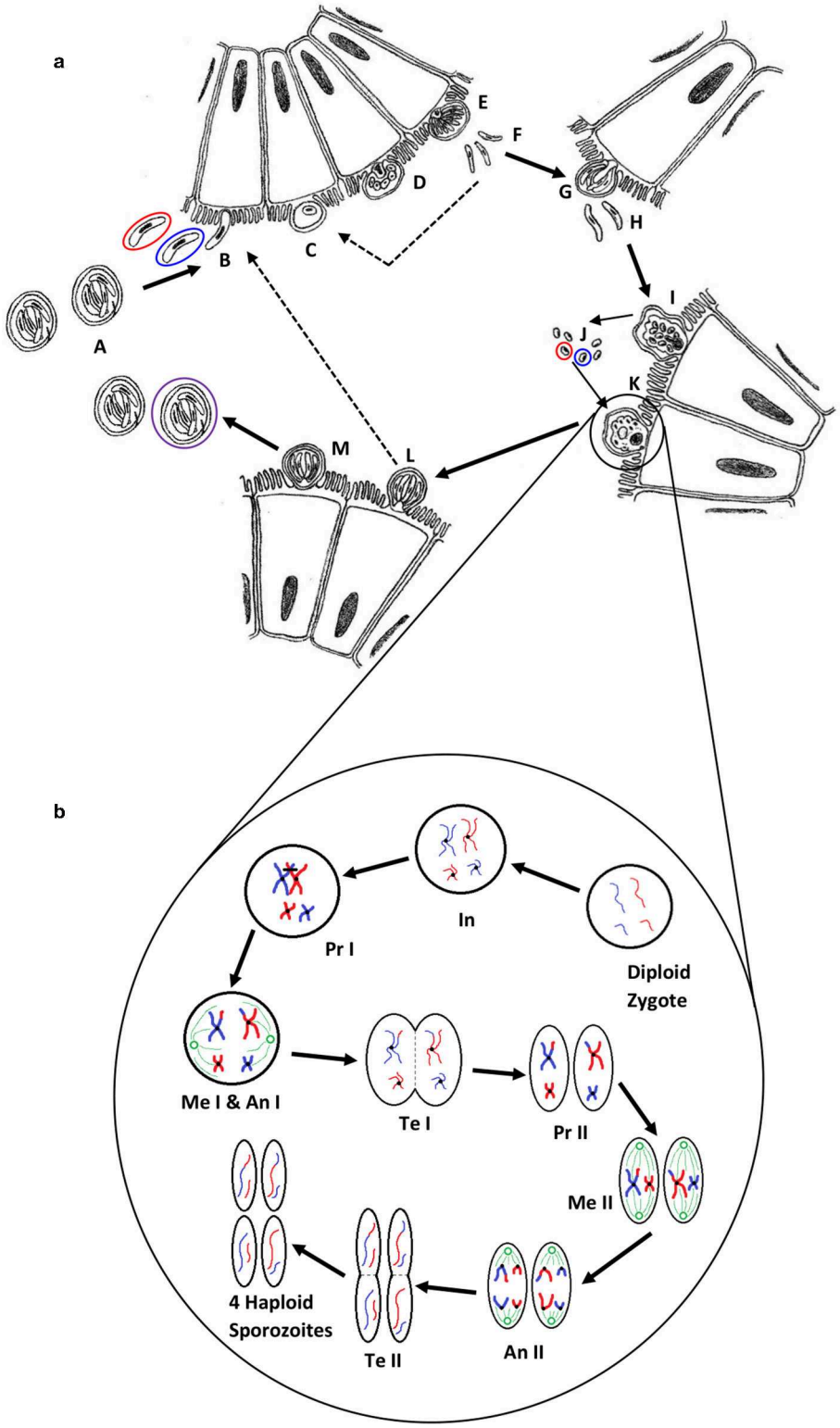


FIGURE 2 | (a) The life cycle of *Cryptosporidium* (1). Oocysts (A) are ingested by the host, most likely as a mixed population of different genotypes; haploid sporozoites (B) (variants are represented by red and blue) excyst and invade the brush border of epithelial cells; each sporozoite develops into a haploid trophozoite with a prominent nucleus (C); the trophozoite undergoes merogony by mitosis to form a type I meront (D,E); up to eight haploid merozoites (F) are released, invade

(Continued)

FIGURE 2 | another cell and undergo merogony again to form either further type I meronts (dotted line) or type II meronts (G), which release four haploid merozoites (H) and form either microgamonts (I) that become multinucleate and mature to form multiple haploid microgametes (J) by mitosis, or a haploid macrogamont (K). Microgamonts are released and potentially each fertilize a macrogamont to form a diploid zygote which undergoes sporogony by meiosis to produce either thin-walled oocysts (L) containing four haploid sporozoites that can autoinfect the host (dotted line), or thick-walled oocysts (M) that are shed in the feces ready to transmit four haploid sporozoites to a new host (the purple circle represents an oocyst that is the product of fertilization between the red and blue genotypes). **(b)** A simplified schematic of genetic recombination in *Cryptosporidium*, potentially generating variation between sporozoites within oocysts. In a mixed infection population, different fertilization scenarios potentially occur—between the same genotypes (resulting in identical daughter sporozoites) or between different genotypes, as in the example shown, that result in a variety of outcomes depending on the random genetic exchange, or lack of, that occurs during meiosis. For simplicity only two example chromosomes are shown with DNA from different genotypes represented by blue and red. The diploid zygote contains duplicate pairs of chromosomes, one set from each parent cell; during interphase (In) the DNA in each chromosome is replicated to produce two identical sister chromatids held together with a centromere; in prophase I (Pr I) the chromosomes start to condense and pair up with the homologous chromosome from the other parent cell, and cross-over can occur resulting in an genetic exchange; during metaphase I (Me I) the paired chromosomes line up along the center of the cell and microtubules connect the centromeres to the centrosomes (shown in green); during anaphase I (An I) each complete set of chromosomes (still paired as sister chromatids) are pulled toward each centrosome—the chromosomes from either parent are randomly combined at this phase introducing a further opportunity for recombination (a blue and a red chromosome are drawn to each centrosome in this example); in telophase I (Te I) the chromosomes start to unravel and cytokinesis starts to split the cell into two, resulting in two haploid cells; in prophase II (Pr II) the chromosomes condense again; during metaphase II (Me II) the chromosomes line up along the center of the cells and microtubules connect the centromeres to the centrosomes; this time during anaphase II (An II) the sister chromatids are separated and pulled apart toward the centrosomes, creating new daughter chromosomes; finally in telophase II (Te II) the chromosomes unravel and cytokinesis starts to split the cells, which in the case of this example due to the crossover event in prophase I, results in four genetically different haploid sporozoites. Depending upon whether random genetic exchanges take place between chromosomes from different genotype parents (either in prophase I or anaphase I) the resulting haploid sporozoites can either be all different, two pairs of identical sporozoites that are different from each parent, or two pairs of identical sporozoites that are the same as the two parents.

may be present. Species-level genotyping has provided improved understanding of human epidemiology in some countries, streamlined by the use of real-time PCR (see below). DNA extraction methods from stool and gene targets have been reviewed in detail by Khan et al. (17).

Beyond the species-level, Sanger sequencing part of the *gp60* gene is most commonly used for further discriminating some *Cryptosporidium* species, including *C. parvum* and *C. hominis* (19–21). The *gp60* gene is hypervariable both between and within *Cryptosporidium* species, and the presence of a highly variable serine repeat region in most species enables further discrimination (19). For nomenclature of *gp60* subtypes, the reader is referred to a review of molecular epidemiologic tools by Xiao and Feng (19). The use of this locus as a subtyping marker has been questioned as it is associated with host cell invasion, and therefore can be considered a virulence factor under selective pressure. Nevertheless, as shown below, it may still be an appropriate target for interrogation as a phenotype determining biomarker. Another issue arises from the use of a single locus; this may not be appropriate due to the genetic recombination that occurs within *Cryptosporidium* populations during the sexual stage of the life-cycle (**Figure 2b**). Whilst not likely or expected between different species, this may occur in populations of mixed subtypes of the same species (22–25). This necessitates the investigation of multiple loci to reveal a more accurate estimate of diversity and population structure (19, 26), and would confer greater discrimination for characterization of isolates (26, 27).

The reality is that genotyping tools are not currently widespread in their application for public health purposes and in most countries *Cryptosporidium* is under-diagnosed and isolates are not characterized (28). In low-resource countries where surveillance data are lacking, research studies have found that *C. hominis* or human-adapted *C. parvum* subtypes predominate (29, 30). *C. parvum* can also be the main species detected in some urban settings with no animals close to residences, further suggesting anthroponotic rather than zoonotic transmission (29). These findings indicate that measures to improve sanitation and hygiene would have greatest impact in these settings. Not only is there a high prevalence of *Cryptosporidium* in these populations, but there is also greater diversity within these species, especially noticeable in *C. hominis*, than is seen in industrialized countries (17, 31).

Genotyping in *Cryptosporidium* Surveillance and Outbreaks

The aim of genotyping in the public health context is to understand transmission and to improve the detection resolution, investigation, and interpretation of waterborne, zoonotic, person-to-person, and foodborne outbreaks. The potential impact lies in:

- Identifying the *Cryptosporidium* species and subtypes that most commonly cause human cryptosporidiosis, and their demographic and temporal-spatial distribution
- Monitoring for the emergence of new species and subtypes in human infection
- Improving detection, investigation, and interpretation of outbreaks
- Increasing the sensitivity of epidemiological investigations to identify links and risk factors, and identify the source of outbreaks and contamination.

In most countries, routine surveillance captures *Cryptosporidium* as an organism, but not species. Where genotyping is used to inform public health, it is mainly in industrialized countries but the framework varies. For example, in England and Wales, clinical diagnostic laboratories have been sending *Cryptosporidium*-positive stools for genotyping for many years, both for molecular surveillance and for outbreak investigations, and most diagnostic stools are genotyped (5, 32). In France, testing for *Cryptosporidium* is not part of routine diagnostic parasitological testing, but a national network of sentinel laboratories was established to test for and genotype new and outbreak cases of cryptosporidiosis (ANOFEL Cryptosporidium

TABLE 1 | *Cryptosporidium* species, their major hosts, oocyst dimensions, reported human infectivity and availability of genome data.

Cryptosporidium species	Mean oocyst dimensions (μm)	Major host(s)	Infections reported in humans	Genomes available (accession number)
C. alticolis	5.4 × 4.9	Voles	No	No
C. apodemi	4.2 × 4.0	Mice	No	No
C. andersoni	7.4 × 5.5	Cattle	Yes (rarely)	PRJNA354069
C. avium	6.3 × 4.9	Birds	No	No
C. baileyi	6.2 × 4.6	Birds	No	PRJNA222835
C. bovis	4.9 × 4.6	Cattle	Yes (rarely)	No
C. canis	5.0 × 4.7	Canids	Yes (occasionally)	No
C. cuniculus	5.6 × 5.4	Lagomorphs, Humans	Yes (occasionally)	PRJNA315496
C. ditrichi	4.7 × 4.2	Mice	Yes (rarely)	No
C. ducismarci	5.0 × 4.8	Tortoises	No	No
C. erinacei	4.9 × 4.4	Hedgehogs	Yes (rarely)	No
C. fayeri	4.9 × 4.3	Marsupials	Yes (rarely)	No
C. felis	4.6 × 4.0	Felids	Yes (occasionally)	No
C. fragile	6.2 × 5.5	Toads	No	No
C. galli	8.3 × 6.3	Birds	No	No
C. homai	Data not available	Guinea Pigs	No	No
C. hominis	4.9 × 5.2	Humans	Yes (commonly)	PRJEB10000 PRJNA13200 PRJNA252787 PRJNA222836 PRJNA222837 PRJNA307563 PRJNA253838 PRJNA253839 PRJNA253834
C. huwi	4.6 × 4.4	Fish	No	No
C. macropodum	5.4 × 4.9	Marsupials	No	No
C. meleagridis	5.2 × 4.6	Birds, mammals	Yes (occasionally)	PRJNA222838 PRJNA315503 PRJNA315502
C. microti	4.3 × 4.1	Voles	No	No
C. molnari	4.7 × 4.5	Fish	No	No
C. muris	7.0 × 5.0	Rodents	Yes (rarely)	PRJNA32283 PRJNA19553
C. occultus	5.2 × 4.9	Rodents	Yes (rarely)	No
C. parvum	5.0 × 4.5	Mammals	Yes (commonly)	PRJNA144 PRJNA320419 PRJNA439211 PRJNA253848 PRJNA253843 PRJNA253845 PRJNA253836 PRJNA253840 PRJNA253846 PRJNA253847 PRJNA320419 PRJNA315506 PRJNA437480 PRJNA315504 PRJNA315508 PRJNA315507 PRJNA315505 PRJNA13873
C. proliferans	7.7 × 5.3	Rodents, maybe Equids	No	No
C. proventriculi	7.4 × 5.7	Birds	No	No
C. rubeyi	4.7 × 4.3	Squirrels	No	No
C. ryanae	3.7 × 3.2	Cattle	No	No
C. scrofarum	5.2 × 4.8	Pigs	Yes (rarely)	No
C. serpentis	6.2 × 5.3	Reptiles	No	No
C. suis	4.6 × 4.2	Pigs	Yes (rarely)	No
C. testudinis	6.4 × 5.9	Tortoises	No	No
C. tyzzeri	4.6 × 4.2	Rodents	Yes (rarely)	No
C. ubiquitum	5.0 × 4.7	Mammals	Yes (occasionally)	PRJNA534291 PRJNA315509 PRJNA315510
C. varanii	4.8 × 4.7	Reptiles	No	No
C. viatorum	5.4 × 4.7	Humans, Rodents	Yes (occasionally)	PRJNA492837
C. wrairi	5.4 × 4.6	Guinea Pigs	No	No
C. xiaoi	3.9 × 3.4	Sheep, Goats	No	No

National Network, 2010). The Netherlands, Sweden and Scotland also use sentinel laboratories to provide sporadic and outbreak samples for genotyping in reference laboratories (28). In the USA, the Centers for Disease Control and Prevention is developing CryptoNet, a molecular-based surveillance system aimed at the systematic collection and molecular characterization of isolates using 18S rDNA PCR-RFLP and *gp60* sequencing (https://www.cdc.gov/parasites/crypto/cryptonet.html). In Germany, Norway, Spain, Ireland, Northern Ireland, Australia, and New Zealand, *Cryptosporidium* genotyping has been used in epidemiological research projects and/or for supporting outbreak investigations (28, 33, 34), while the focus in Asia, Africa, and South American countries has been on molecular epidemiological research (29, 30, 35).

Molecular surveillance data in the United Kingdom (UK) for example has shown that >95% of cases are caused by *C. hominis* or *C. parvum*. Two seasonal peaks in cases occur, with *C. parvum* consistently causing the majority of cases in spring and *C. hominis* predominating in the autumn peak, with much higher rates of foreign travel also reported during this second period (32, 36–38). A similar temporal pattern has been reported in New Zealand (39), but contrasts with the epidemiology in Ireland, where there is no autumn peak and *C. parvum* predominates all year (33, 40). This is likely due to the highly rural socio-geography of Ireland and the greater potential of zoonotic transmission, a feature also seen in rural regions of Great Britain (36, 38). In the UK, the highest incidence of cryptosporidiosis is in children under 5 years, with a second smaller peak in adults in their 20s and 30s; in England and Wales in the period 2000 to 2003, *C. hominis* predominated in infants and the 30–39 year age group (32), and in children <10 years and adults in the period 2004 to 2006 (37), suggesting transmission between children and caregivers. In Ireland, where *C. parvum* predominates, the adult peak does not appear but this may be a testing bias (33, 40).

Although the sentinel surveillance in France is not wholly representative of the French population due to the structure of the network resulting in the inclusion of a higher proportion of hospitalized cases (70%), particularly over-representing the proportion of HIV-infected patients, certain trends are noticeable (ANOFEL Cryptosporidium National Network, 2010). There appears to only be a late summer/autumn peak each year, but the case numbers per month were too low to determine any species-related seasonality. However, *C. parvum* was more prevalent each year compared to *C. hominis* (54.2 vs. 36.5%) and with the remaining 9.4% representing other species (particularly *C. felis*). The seemingly high number of unusual species were mainly found in the over-represented immunocompromised patients (82.8%), which may explain their higher prevalence than in the UK for example.

In the Netherlands, only an autumn peak in case numbers is present in surveillance data, and the predominant species infecting people does not seem to be stable between years. One study undertaken between 2003 and 2005 reported a higher prevalence of *C. hominis* (70.3%) than *C. parvum* (18.7%), with 9.9% cases having both species, and a single case of *C. felis* (41). The infecting species was significantly associated with patient age, with children (aged 0–9 years) more frequently infected with *C. hominis* and adults (over 25 years old) more frequently with *C. parvum* (41). However, over a 3-year study from April 2013, *C. parvum* was most prevalent in years one and two, but in year three (April 2015 to March 2016) *C. hominis* predominated and cases did not decline toward the winter as they had done in previous years (42). Whether these apparent shifts were a function of fluctuating participation in the sentinel scheme or another reason is not known. In England and Wales apparent shifts have also been seen; from 2000 to 2003 the ratio of *C. parvum*:*C. hominis* nationally was close to 1, but in the period 2004–2006 it was 1:1.5, most noticeable in 2005 when it was 1:2.3 and major *C. hominis* outbreaks may have influenced the distribution (37). The UK and the Netherlands both reported an excess in cases of *C. hominis* with similar epidemiology in the latter part of 2015, and despite *gp60* sequencing identifying subtype IbA10G2 and enhanced surveillance, no explanation was found. This was the second time an international *C. hominis* excess had been reported; in the late summer of 2012 the Netherlands, UK, and Germany reported similarly unexplained increases (43).

In the United States (US) national cryptosporidiosis surveillance through CryptoNet is in its infancy, but there seems to be a high diversity of *Cryptosporidium* species and subtypes causing human cryptosporidiosis compared to other industrialized nations (19). While *C. hominis* and *C. parvum* cause the majority of cases, unusual species such as *C. ubiquitum* and the chipmunk genotype are also seen, particularly in rural areas and may suggest an important role of wildlife in transmission, either directly or through drinking untreated water (19). While general surveillance of *Cryptosporidium* species and genotypes in the US is still fairly new, outbreak surveillance has been carried out for many years through the National Outbreak Reporting System (NORS). Analysis of 444 outbreaks of cryptosporidiosis between 2009 and 2017 demonstrated most were in the autumn and caused mainly by waterborne and person-person transmission (44). Molecular data are available for some of the outbreaks on the NORS website https://wwwn.cdc.gov/norsdashboard/. Genotyping data for 131/178 (74%) outbreaks in the same time period in England and Wales showed 69 were caused by *C. parvum* (which caused all animal and environmental contact and food-borne outbreaks, and a minority of recreational water outbreaks), 60 were caused by *C. hominis* (most of the recreational water and all person-to-person spread outbreaks) and in two outbreaks both species were identified (5). Both *C. parvum* and *C. hominis* caused drinking waterborne outbreaks. *Gp60* sequencing established linkage between cases and suspected sources in nine animal contact, three swimming pool, and one drinking water outbreaks (5). Thus, the public health benefits of identifying infecting species and subtypes lie in the ability to identify and strengthen epidemiologic links between cases, and in indicating possible exposures and sources to inform outbreak management (5). However, the ability to differentiate zoonotic and anthroponotic *C. parvum* routinely in all cases would be useful.

Identification by sequencing has established that unusual species of *Cryptosporidium*, previously considered without zoonotic potential, can infect people. Enhanced surveillance has provided some understanding of the transmission of these infections. In the UK, cases with unusual species often reported zoonotic exposures; contact with unwell pets was a significant association, and in particular, contact with cats was reported by significantly more cases with *C. felis* (45). Genotyping *C. ubiquitum* from patients in the US revealed mainly the rodent-adapted subtype families (XIIb-XIId) in contrast to the UK where infections were mainly the ruminant-adapted XIIa subtype family (19, 46).

The potential for outbreaks is not limited to *C. parvum* and *C. hominis*. In 2007 *Cryptosporidium cuniculus* (previously rabbit genotype) was first identified in a patient during routine molecular surveillance in the UK (47). The following year an investigation into a drinking water quality incident in England established that oocysts detected in treated water were *C. cuniculus*. Soon afterwards, primary and secondary *C. cuniculus* cases appeared in the supplied local population, with the same *gp60* subtype, VaA18 (48). Importantly, matching the *Cryptosporidium* isolated from the drinking water, the remains of a rabbit discovered in a chlorine contact tank, and the case samples provided strong evidence for waterborne transmission. This was the first outbreak reported to have caused cryptosporidiosis where the etiological agent was a species other than *C. parvum* or *C. hominis*, and established *C. cuniculus* as a human pathogen. It re-enforced the importance of protecting water supplies not only from livestock and sewage contamination, but also from wildlife.

Sequencing of the *gp60* gene has identified changes in the circulation of predominant subtypes, and the emergence of virulent subtypes. *C. hominis* IbA10G2 continues to predominate in northern Europe, but in the US in 2007, 40 of 57 sporadic cases from four states were a rare subtype, IaA28R4, with IbA10G2 accounting for just eight cases (49). Since 2013, IaA28R4 has been displaced by IfA12G1R5 as the predominant *C. hominis*

genotype in the US associated with both sporadic and outbreak cases (19). In Africa and Asia there is greater variation in *C. hominis* subtypes. For example, in Bangladesh where *C. hominis* is the most common species (>95% of cases) and the seasonality demonstrates a summer peak corresponding to the monsoon, *gp60* analysis revealed 13 different subtypes over a 2 year period (31). Some, for example IaA18R3 and IbA9G3 were present year on year, but other subtypes predominated in some years and disappeared in subsequent years (e.g., IdA15G1 was very common in 2015, but not in 2016 when IaA19R3 and IeA11G3T3 were dominant), indicating a dynamic and frequent transmission (31).

In Europe there is more variation among *C. parvum* than *C. hominis*, although IIaA15G2R1 and IIaA17G1R1 are often (but not always) the most common (5, 19, 50). Genotyping has increased our capacity to detect, investigate and interpret outbreaks. For example, in 2012, *C. parvum* IIaA15G2R1 was used as part of the case definition in an analytical study to investigate a large outbreak (>300 cases) across England and Scotland. A statistically significant association was identified with consumption of pre-cut, bagged mixed salad leaves from a specific national retailer (51). Also in 2012, an outbreak in schoolchildren was associated with a visit to a holiday farm in Norway (52). Genotyping of isolates from cases and potential animal sources on the farm revealed the same rare subtype of *C. parvum*, IIaA19G1R1, in the cases, lambs and goat kids (52). The same holiday farm was also involved in a previous outbreak in 2009 and the same subtype was identified retrospectively, suggesting that in the absence of newly introduced subtypes, existing subtypes can be stable and circulate on the farms for many years (52).

Although *gp60* sequencing has played an important role in refining epidemiological investigations, it is somewhat surprising that there is no standardized multilocus genotyping scheme for *Cryptosporidium* surveillance and outbreaks. Additionally, the lack of suitable markers has hampered our understanding of the main transmission pathway (zoonotic or anthroponotic) of *Cryptosporidium* species and subtypes. As discussed in this paper, genomics has an important role to play in the identification of new markers and the development of a MLG scheme, and the aspiration is that application would eventually become nationally systematic.

Multilocus Genotyping

Currently multilocus genotyping (MLG) is mainly applied to study the population structure of *Cryptosporidium* spp. with few reports describing its utility in surveillance or outbreaks. One example is an investigation into a Swedish swimming pool outbreak in 2002, where multilocus genotyping revealed two concurrent *C. parvum* outbreaks, with different subtypes linked to the use of either the indoor or outdoor pool, indicating multiple contamination events (53). In England, the epidemiological association of *C. parvum* cases with a drinking water supply was strengthened by MLG (54). However, more often investigations have explored the population structure and biology of *Cryptosporidium*.

In 2015, Widmer and Caccio investigated the relationship between sequence and length polymorphism within a set of biomarkers in the *Cryptosporidium* genome. They compared genetic distances of sequence and length polymorphism, finding that there was a weak correlation between the two distance measures. Their results also indicated that the resolution of *Cryptosporidium* population structure was dependent on the genotyping method used (55). Differences in varying extents of host-associated (56, 57) and geographical segregation (24, 58–60), and the extent of panmixia vs. clonality, depending on the population studied (21), have been reported. For example, in Spain, *C. parvum* in cattle herds was reported to show a panmitic population structure contrasting with sheep where *C. parvum* populations appeared more clonal (19, 61, 62). This may have been a function of the predominance of *C. parvum gp60* subtype family (IId) in sheep in the study region of Northeastern Spain (63) as IId has been reported to be clonal in other regions/countries (64).

Pamixia in *Cryptosporidium* spp. may reflect the increased potential for genetic recombination between more diverse isolates than is available in these supposed clonal populations of parasites. The presence of mixed populations with multiple alleles is the driving force of increased variation and the emergence of new subtypes due to recombination events (65–67). In some studies, for example in Scotland *C. hominis* populations have shown clonality (58), but in a cohort of children in Peru, genetic recombination was detected in some *C. hominis* IbA10G2 samples using MLST of 32 polymorphic loci, despite the overall clonality of the *C. hominis* population (65).

However, with the vast majority of *C. hominis* isolates in many areas, including northern Europe and Australia, demonstrating the dominant IbA10G2 (21) the potential for recombination with other more diverse subtypes may be reduced through lack of exposure in those regions. In contrast, the wide variety of different *C. parvum* subtypes usually present in local geographic areas make mixed populations more likely. This has been suggested in a study of the global population structures of both *Cryptosporidium* species, where samples from Uganda showed similar panmitic population structures, contrasting with *C. hominis* samples from the United Kingdom and *C. parvum* from New Zealand which showed much more clonal population structures (68). The authors suggest that both *C. parvum* and *C. hominis* population structures appear to be shaped by local or host-related factors rather than being species-specific (68). This was borne out by a study in Sweden that applied a nine-locus SNP-based method to differentiate *C. hominis* IbA10G2 and grouped 44 isolates, from 12 countries (including 7 non-European), into 10 MLSTs with known epidemiologically-linked samples clustering together; geographical clustering was not obvious, however the numbers of isolates from each country were small (69). In the USA, the emergence and spread of *C. hominis* IaA28R4 was investigated by sequencing eight loci (67). Of 95 *C. hominis* samples (62 IaA28R4 samples) from four states, the sequence diversity identified two clear sub-populations separated geographically between Ohio and three southwestern states, and suggested that the Ohio subpopulation was a descendant of the subpopulation in the southwestern states. Furthermore,

genetic recombination was seen to occur in IaA28R4 isolates and was likely an important factor in its emergence (67), a finding supported by a comparative study of the genome along with the previously dominant IbA10G2 subtype (70).

For disease surveillance and outbreak investigations, there is a need to establish a common multilocus genotyping scheme to track the sources and spread of infection. In a review published in 2012, Robinson and Chalmers reported that different combinations of loci and methods of analysis had been used, with very few groups using comparable loci (27). For public health purposes it is desirable to have consensus to enable cross-boundary comparisons and investigations and track international spread. An initiative funded by EU COST Action FA1408 "A European Network for Foodborne Parasites: Euro-FBP" (http://www.euro-fbp.org) enabled a workshop to be held between 23 scientists and experts in public and animal health from 12 European countries and the USA on *Cryptosporidium* genotyping (71). The participants discussed the need for, and potential directions of, a standardized typing scheme specifically for surveillance and outbreak investigations. There was general agreement that a robust multilocus genotyping scheme should be developed through collaborative laboratory studies, to standardize a method for meaningful interpretation of genotype occurrence and distribution trends, and where possible incorporate into national surveillance programs (71). To achieve this multiple markers spread, sufficiently across the genome, are required. The recent generation of genome data facilitates the identification of markers that show potential to be combined for MLG investigations specifically for surveillance and outbreak investigations (15).

WHOLE GENOME SEQUENCING

While we aspire to using WGS routinely in public health investigations of *Cryptosporidium* cases in the way it is applied to some bacterial pathogens (72–74), the reality is that this is still a way off. Direct sequencing would provide timely investigation of public health incidents, but it poses a challenge for this parasite; it is difficult to culture and bioinformatics pipelines have not been validated for public health purposes as *Cryptosporidium* has suffered from lack of prioritization in genomics programs.

The first technical problem is the amount of DNA that is required. Although this varies depending on the technology used, for example, the Nextera XT DNA kits that have been used in several publications require 1 ng of DNA, and as each oocyst contains 40 fg of DNA it means that 2.5×10^4 oocysts are required without losses and in a practical volume (75). To generate sufficient DNA, oocysts may be propagated through animals, but *Cryptosporidium* populations have been shown to change through natural host-based preferential selection of individual subtypes or further recombination into new subtypes. For example, the "isolate" that provided the first reference *C. hominis* genome in 2004 (TU502) was subsequently serially propagated in gnotobiotic pigs over many years resulting in a different subtype in 2012, which was likely due to the original population being overgrown by another contaminating isolate (76). Additionally, the availability of host animals appropriate to the *Cryptosporidium* species in question (**Table 1**), and the ethics, time and cost resources that are associated with propagation are prohibitive. As propagating oocysts is not a practical solution, obtaining enough clinical sample is the next hurdle, as the volume of stools often submitted is very small. Purity is also a challenge because feces is the starting point, so *Cryptosporidium* DNA is overshadowed by non-target DNA from the biome and host. Lack of purity has been overcome by the combination of several techniques including harvesting by flotation, further purifying by immunomagnetic separation and using the natural chlorine resistance of *Cryptosporidium* oocysts to surface-sterilize them with bleach (75, 77).

The sufficiency of available *Cryptosporidium* DNA has also been addressed through the use of whole genome amplification (WGA) techniques, which now mean that very small amounts of DNA, even from single oocysts, can be used for genome sequencing (77, 78). Guo et al. used WGA to enrich *Cryptosporidium* DNA from six discrete species/genotypes extracted from 24 human and animal fecal samples (77). The results were encouraging, showing that *Cryptosporidium* DNA was significantly enriched, allowing for coverage of $>$ 94% of the genome (77). This ability to whole genome sequence from very low numbers of oocysts is a development that may help when investigating environmental samples and other transmission pathways. Additionally, it may also alleviate problems encountered when whole genome sequencing a mixed population of oocysts. The concern that WGA could result in higher numbers of errors introduced into the genome sequence due to the fidelity of the enzymes used is also unfounded. The presence of four sporozoite genomes in a single oocyst helps, as any errors introduced in the first cycle are unlikely to occur at exactly the same place in more than one genome, so subsequent copies from the other genomes (containing the correct sequence) should overshadow any errors. Although WGS technology has developed and some of the technical hurdles have been overcome to enable direct sequencing (75, 77, 78), we are still not at a point where it can be used to inform in real-time for meaningful surveillance or during outbreak investigations. Aside from technical and resource issues, the lack of user-friendly, validated pipelines specifically designed to generate data in a form that is useful to public health practitioners during the management of incidents, make direct whole genome sequencing currently impractical. Nevertheless, genomic data are being used for biomarker discovery and to understand genetic diversity in parasite populations in different settings. These developments are described below, and arise from the progression of *Cryptosporidium* whole genome sequencing and assembly over the last two decades.

Progression of Whole Genome Sequencing and Assembly

Attempts to sequence the genome of *Cryptosporidium* began in the early 2000s. Initial attempts involved cloning sheared fragments into plasmid vectors and Sanger sequencing. This approach resulted in $>$ 9x coverage of the genome and yielded a fragmented assembly of 221 contigs of length $>$ 5 kbp (79). A more advanced sequencing project was undertaken to resolve gaps, using large *C. parvum* fragments contained within lambda DASH II libraries, and sequence missing DNA using a primer

walk strategy (79). The completed genome of *C. parvum* (Iowa II) along with a preliminary annotation was first published in 2004 by Abrahamsen et al. (80) who passaged oocysts through an animal donor to produce enough parasitic material for the extraction and purification of sufficient amounts of DNA. A random shotgun sequencing approach was used, which yielded a complete genome with coverage of 13x over 18 large contigs (80) and was shortly followed by the publication of the first draft genome of *C. hominis* (TU502) in late 2004. However, this *C. hominis* genome proved to be much more fragmented than that of *C. parvum*, resulting in a sequence consisting of 1,422 contigs (81).

In 2015, the *C. parvum* (Iowa II) reference genome was reassembled and reannotated, and a new *C. hominis* reference genome (UdeA01) published (82). The updated assembly resolved all eight chromosomes from the 18 scaffolds in the previous genome, representing the first chromosome level assembly of *C. parvum*. The reannotation effort increased the number of putative genes from 3807 to 3865 for *C. parvum* Iowa II, and predicted the presence of 3819 genes in *C. hominis* UdeA01 (82). In 2016, Ifeonu et al. reassembled and reannotated the *C. hominis* TU502 genome, along with producing new draft genomes of human isolated *C. hominis* (UKH1) and *C. meleagridis* (UKMEL1) along with the avian species *Cryptosporidium baileyi* (TAMU-09Q1) (83). The *C. hominis* TU502 genome proved to be a considerable improvement on the previous 2004 version, being much more complete, and reducing the number of contigs down to 119. Annotation was facilitated by the RNAseq data generated from the oocyst stage of both *C. hominis* and *C. baileyi*, predicting the presence of 3745 protein coding genes in *C. hominis* TU502 and 3765 in *C. hominis* UKH1 (83).

As can be seen in **Table 2**, there is little difference between the genomes of *C. parvum* and *C. hominis*. They exhibit 95–97% DNA sequence identity; with 11 protein-coding sequences identified only in *C. hominis* and 5 in *C. parvum*, and no large indels or rearrangements apparent (84). The high conservation in the *C. hominis* genomes generated from European samples compared to the much more polymorphic *C. parvum* does not appear to be expressed in general observations on structure and base representation as illustrated in **Table 2**, suggesting that phenotypic differences are potentially due to more subtle sequence divergence (SNPs and Indels) and gene expression. This further illustrates the importance of large-scale sequence comparison of *Cryptosporidium* species to elucidate potentially exploitable variation. Widmer et al. identified a number of highly divergent genes by comparison of the genomes of *C. parvum gp60* subtype IIc and the Iowa II reference (85). Further investigation reveals that genomic evolution was largely reductive, resulting in *Cryptosporidium* depending mainly on host cells for basic nutrients (86).

As more genomes are becoming available at an ever-increasing rate, researchers are able to explore further the biology and evolution of *Cryptosporidium*. Recently, Nader et al. (87) used 21 whole genome sequences to show the existence of two subspecies lineages of *C. parvum* (*C. parvum parvum* and *C. parvum anthroponosum*) with different host-adapted infectivity. Additionally, they identified some of the historic genetic exchanges that have occurred between these lineages and *C. hominis* during the evolution of these different species and subspecies, even suggesting rough time-lines for when these events occurred (87, 88).

In an important epidemiological development, Gilchrist et al. (31) used the methods described by Hadfield et al. (75), to study the genetic diversity of *C. hominis* in slum dwelling infants in Dhaka, Bangladesh, over a 2-year period. As mentioned above, they found that *C. hominis* was more abundant during the monsoon periods and showed high levels of diversity at *gp60* locus. Furthermore, WGS revealed extensive SNP diversity, and very high levels of variation at seven distinct loci. They also detected high levels of recombination within the *C. hominis* populations, evidenced by linkage disequilibrium decay. The genetic diversity of *C. hominis* encountered in the Bangladesh study was found to be far greater than that seen in northern Europe, where the predominant *C. hominis* IbA10G2 subtype is highly conserved at the genome level (50, 71). This study reveals the importance of high-throughput, wide scale genomic sequencing and analysis in elucidating the complex population structure of the parasite worldwide (31).

In another study, WGS was also used for a comparative genomic analysis between two subtypes of *C. hominis* that have been dominant in the US at various times, IbA10G2 and IaA28R4, and *C. parvum* (70). Their genome comparison revealed evidence of genetic recombination in the two *C. hominis* subtypes, and also some unique genetic differences between *C. hominis* and *C. parvum*, and multigene families that may contribute to the host variation between these two species (70).

Genome Availability

The advent of the new techniques to facilitate the DNA extraction, enrichment, sequencing, and assembly of high quality *Cryptosporidium* genomes from clinical samples, provides

TABLE 2 | The progression of *C. hominis* and *C. parvum* whole genome assembly from initial attempts in 2004 to the completed genomes in 2015 and 2016 (80–83).

Feature	*C. parvum* Iowa II (2004)	*C. hominis* TU502 (2004)	*C. hominis* UdeA01 (2015)	*C. parvum* Iowa II (2015)	*C. hominis* TU502 (2016)
Genome length	9.10 Mbp	9.16 Mbp	9.05 Mbp	9.10 Mbp	9.10 Mbp
Coding genes (% genome)	3807 (75.3%)	3994 (69%)	3819 (75.4%)	3865 (75.7%)	3745 (77.8%)
GC content	0.3	0.32	0.32	0.32	0.3
Introns	0.05	0.05-0.20	0.109	0.108	not reported
Fragments	18	1422	8	8	119

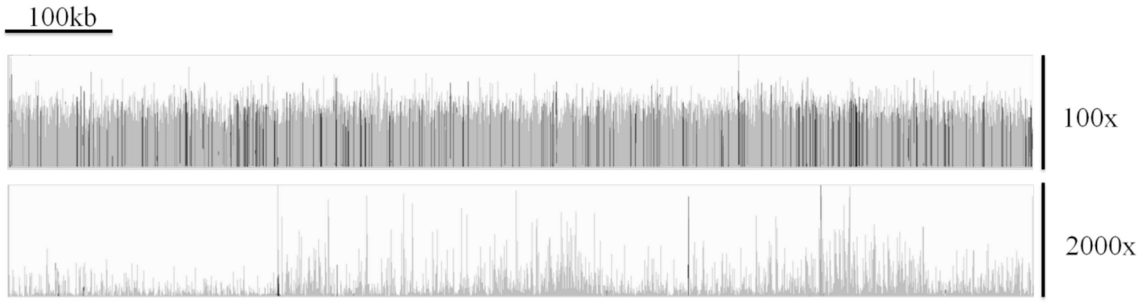


FIGURE 3 | A comparison of the coverage over chromosome 1 *of C. parvum* Iowa II (top track) and the clinical isolate UKP3 (bottom track), showing the highly uneven coverage typically exhibited from many clinical isolates. Reads were mapped using Bowtie v2.3.3.1 (93) and visualized using Integrative Genomics Viewer v2.4.16 (94).

an opportunity to greatly expand the number of genomes available. An EU funded collaboration (Aquavalens project, www.aquavalens.org) between several institutions generated 27 assemblies of *C. parvum*, *C. hominis*, *Cryptosporidium viatorum*, *C. ubiquitum*, *C. cuniculus*, and *C. meleagridis* directly from clinical isolates using the DNA extraction and purification protocol described by Hadfield et al. (75) and Nader et al. (87). Under another EU funded project, COMPARE (https://www.compare-europe.eu/), 31 new *C. parvum* and 19 new *C. hominis* genome assemblies were generated from clinical isolates, using the DNA extraction and purification protocol described by Hadfield et al. (75), and the DNA enrichment protocol described by Guo et al. (77). A further 14 *C. hominis* genomes, representing 9 different *gp60* subtypes, have also been published (89) and are available as a Bioproject (PRJNA307563) on the National Center for Biotechnology Information (NCBI) online databases. Currently, whole genome assemblies of isolates from human and animal derived *Cryptosporidium* spanning 9 species, are available as Bioprojects on NCBI databases (see **Table 1**), but this number is rapidly increasing as methods and technology become more available. The *Cryptosporidium* genomics resource CryptoDB (http://cryptodb.org/), provides access to species including *C. hominis*, *C. parvum*, other zoonotic species including *C. meleagridis*, and host-adapted species rarely found in humans (*Cryptosporidium muris*, *Cryptosporidium andersoni*, *C. baileyi*, and *Cryptosporidium tyzzeri*) and provides analytical tools to mine and compare the genomes sequences and their functionality (90, 91). A number of unassembled, unprocessed raw read sequences are also publically available via online repositories such as GenBank and the Welcome Trust Sanger Institute FTP servers.

Sequencing Using Long-Read Technology

Recently, there have been attempts to generate *Cryptosporidium* sequences using long-read technology, such as MinION by Oxford Nanopore, and Pacific Biosciences. There exist a few draft genomes from long reads generated by PacBio, but most are yet unpublished. However, a *C. parvum* PacBio sequence is available on the Welcome Trust Sanger Institute FTP servers (ftp://ftp.sanger.ac.uk/project/pathogens/Cryptosporidium) that was generated to map shorter Illumina reads to during the study in Dhaka that explored the genetic diversity of *C. hominis* (31). Currently, there have been no successful attempts at sequencing the genome using the MinION platform published. This is likely due to the large amount of DNA required to generate such reads using this particular technology, which is a known difficulty associated with *Cryptosporidium* genomic sequencing.

Pitfalls in Genome Assembly

Morris et al. have outlined difficulties associated with generating reliable and accurate genome assemblies from clinical isolates of *Cryptosporidium* (92). They demonstrated that the issues surrounding extracting sufficient DNA from clinical isolates resulted in highly uneven depth of coverage across the genome (for an example, see **Figure 3**) which can be seen in sequences generated from clinical isolates by a number of research teams. This, in tandem with the large number of low complexity regions within the *Cryptosporidium* genome, results in widespread genome misassembly when using the Spades assembler (95). Peng et al. further proposed an approach to generating reliable draft assemblies from clinical samples, and demonstrated how accurate resolution of low complexity regions are essential for biomarker discovery using the Iterative De-Bruijn Assembler (IDBA) (96).

Assembly of *C. parvum* and *C. hominis* is facilitated by high quality reference sequences (*C. parvum* IowaII and *C. hominis* UdeA01) which allow for reference-guided assembly. This, however, is not the case for other species of *Cryptosporidium*. It is therefore important to consider whether a reference guided assembly should be attempted, and what reference genome to use. The application of an inappropriate reference sequence may result genome assembly errors.

APPLICATIONS, FUTURE ISSUES, AND RESEARCH DIRECTIONS

With the recent expansion in the number of available raw read archives and genome assemblies generated from clinical samples, further *in silico* investigation can be carried out in an attempt to resolve a number of biological questions, such as:

- Can biomarkers differentiate genetic lineages of *Cryptosporidium* spp. virulence or pathogenicity, and therefore act as targets for diagnostic interrogation or novel therapeutics?

- How much variability exists within intergenic regions in species of *Cryptosporidium*?
- To what extent do multiple sub-populations of *Cryptosporidium* spp. exist within an infected host and in single clinical samples and impact of these during onward transmission and even the evolution of the parasite?

Biomarker Discovery and Analysis

The state of *Cryptosporidium* genotyping is far from resolved, and there is still a large amount of work to be done regarding the discovery, assessment, and selection of suitable biomarkers and genotyping conventions. Subsequent to the increasing availability of genomes is a bottle-neck in the analysis of these data, and there is a need to develop time-efficient, computationally inexpensive and high-throughput (automated) methods of genome analysis. "In house" pipelines have been used for biomarker detection and analysis. A typical example was reported by Perez-Cordon et al. (15), who used Tandem Repeats Finder (TRF) (97) to detect Variable Number Tandem Repeat (VNTR) regions within the genome of *Cryptosporidium parvum* Iowa II isolate and aligned them to homologues within a dataset of genomes generated by Hadfield et al. (75). This pipeline consisted of three primary steps:

1. Tandem Repeat (TR) identification in a reference genome.
2. Discovery of the TR's around the genome of a dataset of assembled genomes.
3. Assessment of these TR's for variation and subsequent viability as Biomarkers.

Using this pipeline, bioinformatic analysis of the Hadfield dataset alone has yielded a large number of novel VNTR regions (15), some of which compare favorably to the commonly used *gp60* marker in their ability to resolve discrete subtypes of *C. parvum*. Automating pipelines, can utilize the increasing amounts of whole genome sequence data available for *Cryptosporidium* allowing for the discovery of novel VNTRs in a high-throughput manner.

In addition to novel VNTR markers, genome analysis of other *Cryptosporidium* species and genotypes can allow for the redescription of known markers in these for the development of new subtyping tools. One example, is with the zoonotic species *Cryptosporidium ubiquitum*, where the homolog of *gp60* was diverse from those of *C. hominis* and *C. parvum* so could not be used to differentiate isolates (46). Li et al. used whole genome sequence data to identify and develop a *gp60* subtyping tool that allowed the differentiation and showed apparent host-adaptation (46). Another example, described the development from whole genome sequencing data of a two marker subtyping tool (*gp60* and a mucin protein gene) for the zoonotic chipmunk genotype I (98).

When developing genotyping assays, it is important that biomarkers are selected so as not to influence the outcome of the analysis. For example, markers must be distant enough from each other on the same chromosome or spread over the eight chromosomes to ensure genetic linkage does not occur, and markers must give high enough discrimination when combined to be appropriate for the application in question, such as demonstrating epidemiological relationships (27, 84).

Multiplicity of Infection in *Cryptosporidium*

It is both biologically plausible (due to unrestricted sexual recombination between sub-populations), and there is strong evidence (described below) that infections can arise from, and give rise to, multiple sub-populations of *Cryptosporidium* spp. which will be present in individual hosts (termed here multiplicity of infection—MOI) and thus clinical samples. This is driven by meiotic division in the zygote resulting in potential re-assortment of chromosomes (**Figure 2b**). As a result, the genomes of the haploid sporozoites within an oocyst may differ from each other and the parent sporozoites. Grinberg and Widmer demonstrated the common occurrence of MOI and provided evidence that the degree of MOI may depend on prevailing transmission patterns within geographical regions (25). The current approaches of Sanger sequencing results in the resolution of a single allele at each locus for the population, which, if MOI is present, would in effect simply represent the most populous sequence variant at each locus within the assembly. Grinberg and Widmer illustrated this from three hypothetical infections (25), but the potential extent for MOI is theoretically even greater (**Figure 2b**). This may confound epidemiological analysis, which generally relies on the assumption that large-scale genetic recombination does not occur within a host, and that a single host exhibits a single, clonal population. Furthermore, it has been suggested that MOI is a driving force behind the evolution of virulence, and has a complex relationship with both the virulence experienced by the host, and transmission (99, 100). It is therefore essential that MOI is well-understood and accounted for in order to develop novel prevention strategies in the fight against cryptosporidiosis and other parasitic diseases. The investigation into the impact of MOI relies on the accurate and reliable detection and discrimination of discrete populations of parasites, not readily achieved by current genotyping approaches. There are a few major alternatives to achieve this:

- Cloning and sequencing key loci to detect variation.
- Isolating and sequencing single oocysts from clinical samples.
- Comparing length polymorphism at multiple loci.
- Investigating sequence variation among reads within short read archives generated by Next Generation Sequencing (NGS).

These approaches investigate MOI from very different angles: variable locus cloning and single cell sequences from an experimental angle, and length polymorphism and sequence variation within reads from an *in silico* angle. This lends them unique challenges to overcome. By cloning PCR amplicons of selected loci (*gp60* and hsp70) and utilizing Next Generation Sequencing (NGS), Grinberg et al. reported the presence of numerous sub-populations within single isolates of *C. parvum.* They demonstrated the presence of two *hsp70* and 10 *gp60* alleles within their two isolate dataset. Furthermore, they reported that in both isolates there was a dominant allele, which represented the majority of the amplicons sequenced (101). Single oocysts were isolated and sequenced by Troell et al. (78) with a

view to elucidate these putative intra-isolate sub-populations. Sequencing 10 oocysts individually resulted in assemblies of 49.4–91.8% of the size of the *C. parvum* Iowa II reference genome. By pooling the reads from all 10 oocysts, they generated a 94.4% complete genome. Variation at multiple loci was detected between the assembled genomes, verifying the presence of discrete populations within the "isolate" (78). Analysis of fragment length polymorphism can highlight MOI, however, due to PCR-based amplification of the fragments, minority variants are largely undetectable (25). To compare the results obtained from Sanger sequencing and NGS, Zahedi et al. investigated *gp60* amplicons from 11 *C. hominis*, 22 *C. parvum*, and 8 *C. cuniculus* animal samples from Australia and China (102). They demonstrated that NGS is more effective at resolving the presence of multiple populations of *Cryptosporidium* within a sample, and the extent of MOI. There was concordance between the subtypes identified by both platforms, but additional subtypes were identified using NGS on *C. parvum* and *C. cuniculus gp60* amplicons, but not *C. hominis*.

The major issue with the experimental approaches detailed above is that they are expensive, extremely labor intensive and time consuming, leading to poor scalability. This leads to a major problem in generating sufficient data with which to begin to unravel the role of these parasite sub-populations, and to understand their overall impact on global public health. It is expected that they will have roles in affecting transmission by reducing host-fitness (virulence), and in generating novel subtypes via sexual recombination. There is therefore a great need to develop strategies which allow us to carry out investigations in a high-throughput manner, utilizing the wealth of raw genomic data is available for *Cryptosporidium* and other related parasites. Using biomarkers discovered from the analysis of the increasing number of high quality genomes, the opportunity arises to start to investigate MOI using *in silico* techniques, by mining raw read sets sequenced from clinical samples for information, which may have been previously unattainable. This approach involves three stages:

1. Identification of target regions for read interrogation. It is essential to select target regions, which are likely to show variation in-host, and it is therefore wise to select loci which show large amounts of variation between hosts.
2. Identification of reads within a single-host read set which have captured the target region.
3. Assessment of variation of the target sequence amongst reads which were identified in step 2.

A high level of variation within a single-host read set indicates the presence of multiple populations. Preliminary analysis of the Hadfield et al. dataset (75) indicated extensive variation at multiple tandem repeat loci around the *Cryptosporidium* genome, indicating highly complex in-host population structure. Results for variance mining at the *gp60* locus can be seen in **Figure 4**, which shows high levels of fragment length variation. However, there is invariably a single allele which appears to be

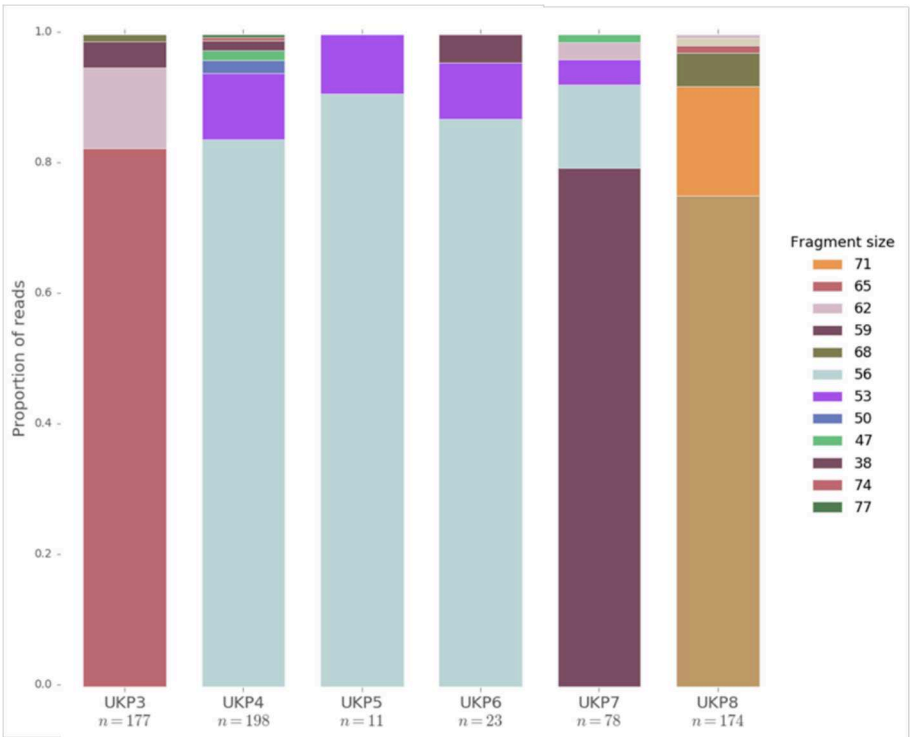


FIGURE 4 | The distribution of fragment lengths at the *gp60* locus mined from raw read sets generated from human clinical samples of UK isolated *C. parvum* by Hadfield et al. (75). Fragment lengths are given in the legend. *n* refers to the number of reads which fully captured the *gp60* region, and are therefore presented in the data.

most frequently exhibited within reads, and therefore considered dominant. This is in agreement with the findings reported by others, which show similar population structure (78, 101). There is, however, a disparity in the extent of MOI in *Cryptosporidium* between laboratory evidence by fragment sizing of key loci, and by mining NGS data. This is potentially due to the limited sensitivity of such approaches to identify multiple alleles of similar fragment size. Furthermore, PCR may preferentially amplify more abundant alleles, resulting in the less abundant alleles being obscured, as shown by Grinberg et al. who initially only identified the predominant alleles in their samples by PCR and Sanger sequencing (101). It may also be the case that such studies were not designed to detect multiple alleles within a single sample, and therefore underestimate the incidence of MOI. Consequently, care should be taken when interpreting entirely *in silico* results in the absence of experimental data. Due to MOI being a new area of investigation in *Cryptosporidium* research, the reliability of *in silico* approaches to elucidate in-host population diversity is still unclear, particularly in the light of studies indicating extensive contamination of samples (77). Preliminary results, however, appear to make predictions which are in accordance with experimental and epidemiological evidence, giving confidence in such data.

Natural transmission studies from analyzing secondary infections and those in farm settings has shown that dominant subtypes can be stable for many years or they can vary from year to year. For example, the outbreaks among visiting children on a holiday farm in Norway showed the same *gp60* subtype, IIaA19G1R1, was still circulating over several years and an investigation into secondary transmission within households after the children returned home also found the same subtype (103). While there was no evidence at the *gp60* gene of mixed populations in this example, in farm settings it is common for multiple subtypes to be present (104, 105). During a study of household transmission in a rural and urban setting in Bangladesh, a wide variety of *gp60* subtypes were found, particularly in the urban setting, but often there were concurrent infections with the same subtype within households and therefore it was mostly impossible to know the directionality of transmission (106). Where there were different subtypes within households it is unclear whether these stemmed from external sources rather than secondary transmission within the household (106). However, despite these studies there is a lack of data from mixed natural infections and the changes or dominance of subtypes that may occur during onward transmission, something that warrants further investigation using multilocus tools or whole genome data. Cama et al. used MLST to characterize differences in Iowa reference *C. parvum* isolates that had been maintained in different laboratories and described differences that were likely the results of passages through calves infected with exogenous *C. parvum* (107). This genetic drift in reference isolates was also seen with the TU502 reference *C. hominis* isolate between 2005 and 2012 following multiple animal passages (76). Therefore, the implications of MOI for surveillance and outbreak investigations are uncertain. As drift may happen in the longer term but not necessarily in the short term, detecting an outbreak "type" is reasonable, but equally it could be that two cases with apparently different subtypes are still actually linked if there is bias in the detection of dominant alleles.

CONCLUSIONS

WGS holds tantalizing promise for better understanding the transmission of cryptosporidosis, but there are still good reasons as to why it is not used routinely for diagnostics in a clinical setting. These include issues with extracting high quality pure DNA from clinical samples and issues with uneven depth of read coverage that leads to gaps in the assembled genome sequence. This later issue has important implications for cost: reducing costs by sequencing at a low depth of coverage is problematic, because it will increase the size and frequency of gaps in the assembled genome sequences. Nonetheless, while WGS is not yet on the horizon as method for routine clinical genotyping, it is indirectly having an important influence on clinical diagnostics. For instance, WGS is being used to guide and inform the development of MLST schemes, such as those based on VNTRs and fragment sizing. It is providing key insights into the evolutionary development of *Cryptosporidium*, including the discovery of new subspecies. Perhaps most important in terms of understanding the transmission of the disease, WGS is providing key insights into MOI. While evidence for MOI is occasionally found using fragment sizing, preliminary WGS analysis shows that is it is much more prevalent than the evidence from fragment sizing might suggest. WGS shows that although clinical samples do indeed contain multiple alleles, a single highly abundant allele usually dominates the data sets. It is highly likely that only the dominant allele that is detected via fragment sizing, with the other alleles remaining undetected. Resolution of these multiple populations is a stepping-stone to understanding the driving factors behind the evolution of virulence, and how new subtypes and genotypes arise in *Cryptosporidium*.

AUTHOR CONTRIBUTIONS

RC devised and revised the manuscript. GR, AM, and MS drafted the manuscript. All authors approved the final manuscript.

FUNDING

This work was funded by the Knowledge Economy Skills Scholarships (KESS 2), a pan-Wales higher level skills initiative led by Bangor University on behalf of the HE sector in Wales. It is part funded by the Welsh Government's European Social Fund (ESF) convergence programme for West Wales and the Valleys.

ACKNOWLEDGMENTS

We would like to thank Gregorio Perez Cordon for preparing **Figure 1** and for helpful comments on the manuscript, and Kevin Tyler for helpful discussions and assistance with **Figure 2b**.

REFERENCES

1. Current WL, Garcia LS. Cryptosporidiosis. *Clin Microbiol Rev*. (1991) 4:325–58. doi: 10.1128/cmr.4.3.325
2. Okhuysen PC, Chappell CL, Crabb JH, Sterling CR, DuPont HL. Virulence of three distinct *Cryptosporidium parvum* isolates for healthy adults. *J Infect Dis*. (1999) 180:1275–81. doi: 10.1086/315033
3. King BJ, Monis PT. Critical processes affecting *Cryptosporidium* oocyst survival in the environment. *Parasitology*. (2007) 134:309–23. doi: 10.1017/S0031182006001491
4. Jenkins MB, Eaglesham BS, Anthony LC, Kachlany SC, Bowman DD, Ghiorse WC. Significance of wall structure, macromolecular composition, and surface polymers to the survival and transport of *Cryptosporidium parvum* oocysts. *Appl Environ Microbiol*. (2010) 76:1926–34. doi: 10.1128/AEM.02295-09
5. Chalmers RM, Robinson G, Elwin K, Elson R. Analysis of the *Cryptosporidium* spp. and *gp60* subtypes linked to human outbreaks of cryptosporidiosis in England and Wales, 2009 to 2017. *Parasit Vectors*. (2019) 12:95. doi: 10.1186/s13071-019-3354-6
6. Sponseller JK, Griffiths JK, Tzipori S. The evolution of respiratory cryptosporidiosis: evidence for transmission by inhalation. *Clin Microbiol Rev*. (2014) 27: 575–86. doi: 10.1128/CMR.00115-13
7. Flores J, Okhuysen PC. Genetics of susceptibility to infection with enteric pathogens. *Curr Opin Infect Dis*. (2009) 22:471–6. doi: 10.1097/QCO.0b013e3283304eb6
8. Borad A, Ward H. Human immune responses in cryptosporidiosis. *Future Microbiol*. (2010) 5:507–19. doi: 10.2217/fmb.09.128
9. Chalmers RM, Davies AP. Minireview: clinical cryptosporidiosis. *Exp Parasitol*. (2010) 124:138–46. doi: 10.1016/j.exppara.2009.02.003
10. Checkley W, White AC Jr, Jaganath D, Arrowood MJ, Chalmers RM, Chen XM, et al. A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for *Cryptosporidium*. *Lancet Infect Dis*. (2015) 15:85–94. doi: 10.1016/S1473-3099(14)70772-8
11. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet*. (2013) 382:209–22. doi: 10.1016/S0140-6736(13)60844-2
12. GBD Diarrhoeal Diseases Collaborators. Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015. *Lancet Infect Dis*. (2017) 17:909–48. doi: 10.1016/S1473-3099(17)30276-1
13. Bartelt LA, Lima AA, Kosek M, Peñataro Yori P, Lee G, Guerrant RL. "Barriers" to child development and human potential: the case for including the "neglected enteric protozoa" (NEP) and other enteropathy-associated pathogens in the NTDs. *PLoS Negl Trop Dis*. (2013) 7:e2125. doi: 10.1371/journal.pntd.0002125
14. Savioli L, Smith H, Thompson A. *Giardia* and *Cryptosporidium* join the 'Neglected Diseases Initiative'. *Trends Parasitol*. (2006) 22:203–8. doi: 10.1016/j.pt.2006.02.015
15. Pérez-Cordón G, Robinson G, Nader J, Chalmers RM. Discovery of new variable number tandem repeat loci in multiple *Cryptosporidium parvum* genomes for the surveillance and investigation of outbreaks of cryptosporidiosis. *Exp Parasitol*. (2016) 169:119–28. doi: 10.1016/j.exppara.2016.08.003
16. Bouzid M, Hunter PR, Chalmers RM, Tyler KM. *Cryptosporidium* pathogenicity and virulence. *Clin Microbiol Rev*. (2013) 26:115–34. doi: 10.1128/CMR.00076-12
17. Khan A, Shaik JS, Grigg ME. Genomics and molecular epidemiology of *Cryptosporidium* species. *Acta Trop*. (2018) 184:1–14. doi: 10.1016/j.actatropica.2017.10.023
18. Baragaña B, Forte B, Choi R, Nakazawa Hewitt S, Bueren-Calabuig JA, Pisco JP, et al. Lysyl-tRNA synthetase as a drug target in malaria and cryptosporidiosis. *Proc Natl Acad Sci USA*. (2019) 116:7015–20. doi: 10.1073/pnas.1814685116
19. Xiao L, Feng Y. Molecular epidemiologic tools for waterborne pathogens *Cryptosporidium* spp. and *Giardia duodenalis*. *Food Waterborne Parasitol*. (2017) 8–9:14–32. doi: 10.1016/j.fawpar.2017.09.002
20. Xiao L. Molecular epidemiology of cryptosporidiosis: an update. *Exp Parasitol*. (2010) 124:80–9. doi: 10.1016/j.exppara.2009.03.018
21. Feng Y, Ryan UM, Xiao L. Genetic diversity and population structure of *Cryptosporidium*. *Trends Parasitol*. (2018) 34:997–1011. doi: 10.1016/j.pt.2018.07.009
22. Feng X, Rich SM, Tzipori S, Widmer G. Experimental evidence for genetic recombination in the opportunistic pathogen *Cryptosporidium parvum*. *Mol Biochem Parasitol*. (2002) 119:55–62. doi: 10.1016/S0166-6851(01)00393-0
23. Tanriverdi S, Blain JC, Deng B, Ferdig MT, Widmer G. Genetic crosses in the apicomplexan parasite *Cryptosporidium parvum* define recombination parameters. *Mol Microbiol*. (2007) 63:1432–9. doi: 10.1111/j.1365-2958.2007.05594.x
24. Mallon M, MacLeod A, Wastling J, Smith H, Reilly B, Tait A. Population structures and the role of genetic exchange in the zoonotic pathogen *Cryptosporidium parvum*. *J Mol Evol*. (2003) 56:407–17. doi: 10.1007/s00239-002-2412-3
25. Grinberg A, Widmer G. *Cryptosporidium* within-host genetic diversity: systematic bibliographical search and narrative overview. *Int J Parasitol*. (2016) 46:465–71. doi: 10.1016/j.ijpara.2016.03.002
26. Widmer G, Lee Y. Comparison of single- and multilocus genetic diversity in the protozoan parasites *Cryptosporidium parvum* and *C. hominis*. *Appl Environ Microbiol*. (2010) 76:6639–44. doi: 10.1128/AEM.01268-10
27. Robinson G, Chalmers RM. Assessment of polymorphic genetic markers for multi-locus typing of *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Exp Parasitol*. (2012) 132:200–15. doi: 10.1016/j.exppara.2012.06.016
28. Chalmers RM, Pérez-Cordón G, Cacció SM, Klotz C, Robertson LJ, on behalf of the participants of the *Cryptosporidium* genotyping workshop (EURO-FBP). *Cryptosporidium* genotyping in Europe: the current status and processes for a harmonised multi-locus genotyping scheme. *Exp Parasitol*. (2018) 191:25–30. doi: 10.1016/j.exppara.2018.06.004
29. Aldeyarbi HM, Abu El-Ezz NM, Karanis P. *Cryptosporidium* and cryptosporidiosis: the African perspective. *Environ Sci Pollut Res Int*. (2016) 23:13811–21. doi: 10.1007/s11356-016-6746-6
30. Mahmoudi MR, Ongerth JE, Karanis P. *Cryptosporidium* and cryptosporidiosis: the Asian perspective. *Int J Hyg Environ Health*. (2017) 220:1098–109. doi: 10.1016/j.ijheh.2017.07.005
31. Gilchrist CA, Cotton JA, Burkey C, Arju T, Gilmartin A, Lin Y, et al. Genetic diversity of *Cryptosporidium hominis* in a Bangladeshi community as revealed by whole-genome sequencing. *J Infect Dis*. (2018) 218:259–64. doi: 10.1093/infdis/jiy121
32. Chalmers RM, Elwin K, Thomas AL, Guy EC, Mason B. Long-term *Cryptosporidium* typing reveals the aetiology and species-specific epidemiology of human cryptosporidiosis in England and Wales, 2000 to 2003. *Euro Surveill*. (2009) 14:19086. doi: 10.2807/ese.14.02.19086-en
33. Zintl A, Proctor AF, Read C, Dewaal T, Shanaghy N, Fanning S, et al. The prevalence of *Cryptosporidium* species and subtypes in human faecal samples in Ireland. *Epidemiol Infect*. (2009) 137:270–7. doi: 10.1017/S0950268808000769
34. Waldron LS, Ferrari BC, Cheung-Kwok-Sang C, Beggs PJ, Stephens N, Power ML. Molecular epidemiology and spatial distribution of a waterborne cryptosporidiosis outbreak in Australia. *Appl Environ Microbiol*. (2011) 77:7766–71. doi: 10.1128/AEM.00616-11
35. Garcia-R JC, French N, Pita A, Velathanthiri N, Shrestha R, Hayman D. Local and global genetic diversity of protozoan parasites: spatial distribution of *Cryptosporidium* and *Giardia* genotypes. *PLoS Negl Trop Dis*. (2017) 11:e0005736. doi: 10.1371/journal.pntd.0005736
36. McLauchlin J, Amar C, Pedraza-Díaz S, Nichols GL. Molecular epidemiological analysis of *Cryptosporidium* spp. in the United Kingdom: results of genotyping *Cryptosporidium* spp. in 1,705 fecal samples from humans and 105 fecal samples from livestock animals. *J Clin Microbiol*. (2000) 38:3984–90.
37. Chalmers RM, Smith R, Elwin K, Clifton-Hadley FA, Giles M. Epidemiology of anthroponotic and zoonotic human cryptosporidiosis in England and Wales, 2004-2006. *Epidemiol Infect*. (2011) 139:700–12. doi: 10.1017/S0950268810001688
38. Pollock KG, Ternent HE, Mellor DJ, Chalmers RM, Smith HV, Ramsay CN, et al. Spatial and temporal epidemiology of sporadic human

cryptosporidiosis in Scotland. *Zoonoses Public Health*. (2010) 57:487–92. doi: 10.1111/j.1863-2378.2009.01247.x
39. Learmonth J, Ionas G, Pita A, Cowie R. Seasonal shift in *Cryptosporidium parvum* transmission cycles in New Zealand. *J Eukaryot Microbiol*. (2001) 48:34S—5. doi: 10.1111/j.1550-7408.2001.tb00444.x
40. Garvey P, McKeown P. Epidemiology of human cryptosporidiosis in Ireland, 2004–2006: analysis of national notification data. *Euro Surveill*. (2009) 14:19128. doi: 10.2807/ese.14.08.19128-en
41. Wielinga PR, de Vries A, van der Goot TH, Mank T, Mars MH, Kortbeek LM, et al. Molecular epidemiology of Cryptosporidium in humans and cattle in The Netherlands. *Int J Parasitol*. (2008) 38:809–17. doi: 10.1016/j.ijpara.2007.10.014
42. Nic Lochlainn LM, Sane J, Schimmer B, Mooij S, Roelfsema J, van Pelt W, et al. Risk factors for sporadic cryptosporidiosis in the netherlands: analysis of a 3-year population based case-control study coupled with genotyping, 2013-2016. *J Infect Dis*. (2019) 219:1121–9. doi: 10.1093/infdis/jiy634
43. Roelfsema JH, Sprong H, Cacciò SM, Takumi K, Kroes M, van Pelt W, et al. Molecular characterization of human *Cryptosporidium* spp. isolates after an unusual increase in late summer 2012. *Parasit Vectors*. (2016) 9:138. doi: 10.1186/s13071-016-1397-5
44. Gharpure R, Perez A, Miller AD, Wikswo ME, Silver R, Hlavsa MC. Cryptosporidiosis Outbreaks - United States, 2009-2017. *MMWR Morb Mortal Wkly Rep*. (2019) 68:568–72. doi: 10.15585/mmwr.mm6825a3
45. Elwin K, Hadfield SJ, Robinson G, Chalmers RM. The epidemiology of sporadic human infections with unusual cryptosporidia detected during routine typing in England and Wales, 2000-2008. *Epidemiol Infect*. (2012) 140:673–83. doi: 10.1017/S0950268811000860
46. Li N, Xiao L, Alderisio K, Elwin K, Cebelinski E, Chalmers R, et al. Subtyping *Cryptosporidium ubiquitum*, a zoonotic pathogen emerging in humans. *Emerg Infect Dis*. (2014) 20:217–24. doi: 10.3201/eid2002.121797
47. Robinson G, Elwin K, Chalmers RM. Unusual *Cryptosporidium* genotypes in human cases of diarrhea. *Emerg Infect Dis*. (2008) 14:1800–2. doi: 10.3201/eid1411.080239
48. Chalmers RM, Robinson G, Elwin K, Hadfield SJ, Xiao L, Ryan U, et al. *Cryptosporidium* sp. rabbit genotype, a newly identified human pathogen. *Emerg Infect Dis*. (2009) 15:829–30. doi: 10.3201/eid1505.081419
49. Xiao L, Hlavsa MC, Yoder J, Ewers C, Dearen T, Yang W, et al. Subtype analysis of *Cryptosporidium* specimens from sporadic cases in Colorado, Idaho, New Mexico, and Iowa in 2007: widespread occurrence of one *Cryptosporidium hominis* subtype and case history of an infection with the *Cryptosporidium* horse genotype. *J Clin Microbiol*. (2009) 47:3017–20. doi: 10.1128/JCM.00226-09
50. Cacciò SM, Chalmers RM. Human cryptosporidiosis in Europe. *Clin Microbiol Infect*. (2016) 22:471–80. doi: 10.1016/j.cmi.2016.04.021
51. McKerr C, Adak GK, Nichols G, Gorton R, Chalmers RM, Kafatos G, et al. An outbreak of *Cryptosporidium parvum* across England & Scotland associated with consumption of fresh pre-cut salad leaves, May 2012. *PLoS ONE*. (2015) 10:e0125955. doi: 10.1371/journal.pone.0125955
52. Lange H, Johansen OH, Vold L, Robertson LJ, Anthonisen IL, Nygard K. Second outbreak of infection with a rare *Cryptosporidium parvum* genotype in schoolchildren associated with contact with lambs/goat kids at a holiday farm in Norway. *Epidemiol Infect*. (2014) 142:2105–13. doi: 10.1017/S0950268813003002
53. Mattsson JG, Insulander M, Lebbad M, Björkman C, Svenungsson B. Molecular typing of *Cryptosporidium parvum* associated with a diarrhoea outbreak identifies two sources of exposure. *Epidemiol Infect*. (2008) 136:1147–52. doi: 10.1017/S0950268807009673
54. Hunter PR, Wilkinson DC, Lake IR, Harrison FC, Syed Q, Hadfield SJ, et al. Microsatellite typing of *Cryptosporidium parvum* in isolates from a waterborne outbreak. *J Clin Microbiol*. (2008) 46:3866–7. doi: 10.1128/JCM.01636-08
55. Widmer G, Cacciò SM. A comparison of sequence and length polymorphism for genotyping *Cryptosporidium* isolates. *Parasitology*. (2015) 142:1080–5. doi: 10.1017/S0031182015000396
56. Drumo R, Widmer G, Morrison LJ, Tait A, Grelloni V, D'Avino N, et al. Evidence of host-associated populations of *Cryptosporidium parvum* in Italy. *Appl Environ Microbiol*. (2012) 78:3523–9. doi: 10.1128/AEM.07686-11
57. Quílez J, Vergara-Castiblanco C, Monteagudo L, del Cacho E, Sánchez-Acedo C. Host association of *Cryptosporidium parvum* populations infecting domestic ruminants in Spain. *Appl Environ Microbiol*. (2013) 79:5363–71. doi: 10.1128/AEM.01168-13
58. Mallon ME, MacLeod A, Wastling JM, Smith H, Tait A. Multilocus genotyping of *Cryptosporidium parvum* Type 2: population genetics and sub-structuring. *Infect Genet Evol*. (2003) 3:207–18. doi: 10.1016/S1567-1348(03)00089-3
59. Gatei W, Hart CA, Gilman RH, Das P, Cama V, Xiao L. Development of a multilocus sequence typing tool for *Cryptosporidium hominis*. *J Eukaryot Microbiol*. (2006) 53:S43–8. doi: 10.1111/j.1550-7408.2006.00169.x
60. Cacciò SM, de Waele V, Widmer G. Geographical segregation of *Cryptosporidium parvum* multilocus genotypes in Europe. *Infect Genet Evol*. (2015) 31:245–9. doi: 10.1016/j.meegid.2015.02.008
61. Ramo A, Quílez J, Monteagudo L, Del Cacho E, Sánchez-Acedo C. Intra-species diversity and panmictic structure of *Cryptosporidium parvum* populations in cattle farms in Northern Spain. *PLoS ONE*. (2016) 11:e0148811. doi: 10.1371/journal.pone.0148811
62. Ramo A, Monteagudo LV, Del Cacho E, Sánchez-Acedo C, Quílez J. Intra-species genetic diversity and clonal structure of *Cryptosporidium parvum* in sheep farms in a confined geographical area in Northeastern Spain. *PLoS ONE*. (2016) 11:e0155336. doi: 10.1371/journal.pone.0155336
63. Quílez J, Torres E, Chalmers RM, Hadfield SJ, Del Cacho E, Sánchez-Acedo C. *Cryptosporidium* genotypes and subtypes in lambs and goat kids in Spain. *Appl Environ Microbiol*. (2008) 74:6026–31. doi: 10.1128/AEM.00606-08
64. Wang R, Zhang L, Axén C, Bjorkman C, Jian F, Amer S, et al. *Cryptosporidium parvum* IId family: clonal population and dispersal from Western Asia to other geographical regions. *Sci Rep*. (2014) 4:4208. doi: 10.1038/srep04208
65. Li N, Xiao L, Cama VA, Ortega Y, Gilman RH, Guo M, et al. Genetic recombination and *Cryptosporidium hominis* virulent subtype IbA10G2. *Emerg Infect Dis*. (2013) 19:1573–82. doi: 10.3201/eid1910.121361
66. Feng Y, Torres E, Li N, Wang L, Bowman D, Xiao L. Population genetic characterisation of dominant *Cryptosporidium parvum* subtype IIaA15G2R1. *Int J Parasitol*. (2013) 43:1141–7. doi: 10.1016/j.ijpara.2013.09.002
67. Feng Y, Tiao N, Li N, Hlavsa M, Xiao L. Multilocus sequence typing of an emerging *Cryptosporidium hominis* subtype in the United States. *J Clin Microbiol*. (2014) 52:524–30. doi: 10.1128/JCM.02973-13
68. Tanriverdi S, Grinberg A, Chalmers RM, Hunter PR, Petrovic Z, Akiyoshi DE, et al. Inferences about the global population structures of *Cryptosporidium parvum* and *Cryptosporidium hominis*. *Appl Environ Microbiol*. (2008) 74:7227–34. doi: 10.1128/AEM.01576-08
69. Beser J, Hallström BM, Advani A, Andersson S, Östlund G, Winiecka-Krusnell J, et al. Improving the genotyping resolution of *Cryptosporidium hominis* subtype IbA10G2 using one step PCR-based amplicon sequencing. *Infect Genet Evol*. (2017) 55:297–304. doi: 10.1016/j.meegid.2017.08.035
70. Guo Y, Tang K, Rowe LA, Li N, Roellig DM, Knipe K, et al. Comparative genomic analysis reveals occurrence of genetic recombination in virulent *Cryptosporidium hominis* subtypes and telomeric gene duplications in *Cryptosporidium parvum*. *BMC Genomics*. (2015) 16:320. doi: 10.1186/s12864-015-1517-1
71. Chalmers RM, Caccio S. Towards a consensus on genotyping schemes for surveillance and outbreak investigations of *Cryptosporidium*, Berlin, June 2016. *Euro Surveill*. (2016) 21:30338. doi: 10.2807/1560-7917.ES.2016.21.37.30338
72. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, et al. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol*. (2012) 13:R118. doi: 10.1186/gb-2012-13-12-r118
73. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *Lancet Infect Dis*. (2013) 13:137–46. doi: 10.1016/S1473-3099(12)70277-3
74. Dallman TJ, Byrne L, Ashton PM, Cowley LA, Perry NT, Adak G, et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing *Escherichia coli* O157. *Clin Infect Dis*. (2015) 61:305–12. doi: 10.1093/cid/civ318

75. Hadfield SJ, Pachebat JA, Swain MT, Robinson G, Cameron SJ, Alexander J, et al. Generation of whole genome sequences of new *Cryptosporidium hominis* and *Cryptosporidium parvum* isolates directly from stool samples. *BMC Genomics*. (2015) 16:650. doi: 10.1186/s12864-015-1805-9
76. Widmer G, Ras R, Chalmers RM, Elwin K, Desoky E, Badawy A. Population structure of natural and propagated isolates of *Cryptosporidium parvum*, *C. hominis* and *C. meleagridis*. *Environ Microbiol*. (2015) 17:984–93. doi: 10.1111/1462-2920.12447
77. Guo Y, Li N, Lysén C, Frace M, Tang K, Sammons S, et al. Isolation and enrichment of *Cryptosporidium* DNA and verification of DNA purity for whole-genome sequencing. *J Clin Microbiol*. (2015) 53:641–7. doi: 10.1128/JCM.02962-14
78. Troell K, Hallström B, Divne AM, Alsmark C, Arrighi R, Huss M, et al. *Cryptosporidium* as a testbed for single cell genome characterization of unicellular eukaryotes. *BMC Genomics*. (2016) 17:471. doi: 10.1186/s12864-016-2815-y
79. Widmer G, Lin L, Kapur V, Feng X, Abrahamsen MS. Genomics and genetics of *Cryptosporidium parvum*: the key to understanding cryptosporidiosis. *Microbes Infect*. (2002) 4:1081–90. doi: 10.1016/S1286-4579(02)01632-5
80. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, Lancto CA, et al. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*. (2004) 304:441–5. doi: 10.1126/science.1094786
81. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, et al. The genome of *Cryptosporidium hominis*. *Nature*. (2004) 431:1107–12. doi: 10.1038/nature02977
82. Isaza JP, Galván AL, Polanco V, Huang B, Matveyev AV, Serrano MG, et al. Revisiting the reference genomes of human pathogenic *Cryptosporidium* species: reannotation of *C. parvum* Iowa and a new *C. hominis* reference. *Sci Rep*. (2015) 5:16324. doi: 10.1038/srep16324
83. Ifeonu OO, Chibucos MC, Orvis J, Su Q, Elwin K, Guo F, et al. Annotated draft genome sequences of three species of *Cryptosporidium*: *Cryptosporidium meleagridis* isolate UKMEL1, *C. baileyi* isolate TAMU-09Q1 and *C. hominis* isolates TU502 2012 and UKH1. *Pathog Dis*. (2016) 74:ftw080. doi: 10.1093/femspd/ftw080
84. Widmer G, Sullivan S. Genomics and population biology of *Cryptosporidium* species. *Parasite Immunol*. (2012) 34:61–71. doi: 10.1111/j.1365-3024.2011.01301.x
85. Widmer G, Lee Y, Hunt P, Martinelli A, Tolkoff M, Bodi K. Comparative genome analysis of two *Cryptosporidium parvum* isolates with different host range. *Infect Genet Evol*. (2012) 12:1213–21. doi: 10.1016/j.meegid.2012.03.027
86. Keeling PJ. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. *Dev Cell*. (2004) 6:614–6. doi: 10.1016/S1534-5807(04)00135-2
87. Nader JL, Mathers TC, Ward BJ, Pachebat JA, Swain MT, Robinson G, et al. Evolutionary genomics of anthroponosis in *Cryptosporidium*. *Nat Microbiol*. (2019) 4:826–36. doi: 10.1038/s41564-019-0377-x
88. Kissinger JC. Evolution of *Cryptosporidium*. *Nat Microbiol*. (2019) 4:730–1. doi: 10.1038/s41564-019-0438-1
89. Sikora P, Andersson S, Winiecka-Krusnell J, Hallström B, Alsmark C, Troell K, et al. Genomic variation in IbA10G2 and other patient-derived *Cryptosporidium hominis* subtypes. *J Clin Microbiol*. (2017) 55:844–58. doi: 10.1128/JCM.01798-16
90. Puiu D, Enomoto S, Buck GA, Abrahamsen MS, Kissinger JC. CryptoDB: the *Cryptosporidium* genome resource. *Nucleic Acids Res*. (2004) 32:D329–31. doi: 10.1093/nar/gkh050
91. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, Kaluskar N, et al. CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Res*. (2006) 34:D419–22. doi: 10.1093/nar/gkj078
92. Morris A, Pachebat J, Robinson G, Chalmers R, Swain M. Identifying and resolving genome misassembly issues important for biomarker discovery in the protozoan parasite, *Cryptosporidium*. In Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies. *Bioinformatics*. (2019) 3:90–100. doi: 10.5220/0007397200900100
93. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol*. (2009) 10:R25. doi: 10.1186/gb-2009-10-3-r25
94. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. (2013) 14:178–92. doi: 10.1093/bib/bbs017
95. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. (2012) 19:455–77. doi: 10.1089/cmb.2012.0021
96. Peng Y, Leung HC, Yiu SM, Chin FY. IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. (2012) 28:1420–8. doi: 10.1093/bioinformatics/bts174
97. Benson G. Tandem Repeats Finder: a program to analyse DNA sequences. *Nucleic Acids Res*. (1999) 27:573–8. doi: 10.1093/nar/27.2.573
98. Guo Y, Cebelinski E, Matusevich C, Alderisio KA, Lebbad M, McEvoy J, et al. Subtyping novel zoonotic pathogen *Cryptosporidium* chipmunk genotype I. *J Clin Microbiol*. (2015) 53:1648–54. doi: 10.1128/JCM.03436-14
99. Alizon S, de Roode JC, Michalakis Y. Multiple infections and the evolution of virulence. *Ecol Lett*. (2013) 16:556–67. doi: 10.1111/ele.12076
100. Sondo P, Derra K, Lefevre T, Diallo-Nakanabo S, Tarnagda Z, Zampa O, et al. Genetically diverse *Plasmodium falciparum* infections, within-host competition and symptomatic malaria in humans. *Sci Rep*. (2019) 9:1–9. doi: 10.1038/s41598-018-36493-y
101. Grinberg A, Biggs PJ, Dukkipati VS, George TT. Extensive intra-host genetic diversity uncovered in *Cryptosporidium parvum* using Next Generation Sequencing. *Infect Genet Evol*. (2013) 15:18–24. doi: 10.1016/j.meegid.2012.08.017
102. Zahedi A, Gofton AW, Jian F, Paparini A, Oskam C, Ball A, et al. Next Generation Sequencing uncovers within-host differences in the genetic diversity of *Cryptosporidium gp60* subtypes. *Int J Parasitol*. (2017) 47:601–7. doi: 10.1016/j.ijpara.2017.03.003
103. Johansen ØH, Hanevik K, Thrana F, Carlson A, Stachurska-Hagen T, Skaare D, et al. Symptomatic and asymptomatic secondary transmission of *Cryptosporidium parvum* following two related outbreaks in schoolchildren. *Epidemiol Infect*. (2015) 143:1702–9. doi: 10.1017/S095026881400243X
104. Wells B, Shaw H, Hotchkiss E, Gilray J, Ayton R, Green J, et al. Prevalence, species identification and genotyping Cryptosporidium from livestock and deer in a catchment in the Cairngorms with a history of a contaminated public water supply. *Parasit Vectors*. (2015) 8:66. doi: 10.1186/s13071-015-0684-x
105. Thomson S, Innes EA, Jonsson NN, Katzer F. Shedding of *Cryptosporidium* in calves and dams: evidence of re-infection and shedding of different gp60 subtypes. *Parasitology*. (2019) 146:1404–13. doi: 10.1017/S0031182019000829
106. Korpe PS, Gilchrist C, Burkey C, Taniuchi M, Ahmed E, Madan V, et al. Case-control study of *Cryptosporidium* transmission in bangladeshi households. *Clin Infect Dis*. (2019) 68:1073–9. doi: 10.1093/cid/ciy593
107. Cama VA, Arrowood MJ, Ortega YR, Xiao L. Molecular characterization of the *Cryptosporidium parvum* IOWA isolate kept in different laboratories. *J Eukaryot Microbiol*. (2006) 53:S40–2. doi: 10.1111/j.1550-7408.2006.00168.x

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

CURRICULUM, INSTRUCTION, AND PEDAGOGY
published: 25 February 2020
doi: 10.3389/fpubh.2020.00038

Addressing Learning Needs on the Use of Metagenomics in Antimicrobial Resistance Surveillance

Ana Sofia Ribeiro Duarte[1], Katharina D. C. Stärk[2], Patrick Munk[1], Pimlapas Leekitcharoenphon[1], Alex Bossers[3,4], Roosmarijn Luiken[4], Steven Sarrazin[5], Oksana Lukjancenko[1], Sünje Johanna Pamp[1], Valeria Bortolaia[1], Jakob Nybo Nissen[6], Philipp Kirstahler[1], Liese Van Gompel[4], Casper Sahl Poulsen[1], Rolf Sommer Kaas[1], Maria Hellmér[7], Rasmus Borup Hansen[8], Violeta Munoz Gomez[2] and Tine Hald[1]*

[1] Research Group for Genomic Epidemiology, National Food Institute, Technical University of Denmark, Lyngby, Denmark, [2] SAFOSO AG, Bern-Lieberfeld, Switzerland, [3] Department of Infection Biology, Wageningen Bioveterinary Research, Lelystad, Netherlands, [4] Faculty of Veterinary Medicine, Institute for Risk Assessment Sciences, Utrecht University, Utrecht, Netherlands, [5] Veterinary Epidemiology Unit, Faculty of Veterinary Medicine, Ghent University, Merelbeke, Belgium, [6] Department of Health Technology, Technical University of Denmark, Lyngby, Denmark, [7] Research Group for Microbiology and Hygiene, National Food Institute, Technical University of Denmark, Lyngby, Denmark, [8] Intomics A/S, Lyngby, Denmark

OPEN ACCESS

Edited by:
Marc Jean Struelens, European Centre for Disease Prevention and Control (ECDC), Sweden

Reviewed by:
Teresa M. Coque, Ramón y Cajal Institute for Health Research, Spain
Olov Johan Svartström, Public Health Agency of Sweden, Sweden

***Correspondence:**
Ana Sofia Ribeiro Duarte
asrd@food.dtu.dk

Specialty section:
This article was submitted to Infectious Diseases Surveillance, Prevention and Treatment, a section of the journal Frontiers in Public Health

Received: 23 August 2019
Accepted: 05 February 2020
Published: 25 February 2020

Citation:
Duarte ASR, Stärk KDC, Munk P, Leekitcharoenphon P, Bossers A, Luiken R, Sarrazin S, Lukjancenko O, Pamp SJ, Bortolaia V, Nissen JN, Kirstahler P, Van Gompel L, Poulsen CS, Kaas RS, Hellmér M, Hansen RB, Gomez VM and Hald T (2020) Addressing Learning Needs on the Use of Metagenomics in Antimicrobial Resistance Surveillance. Front. Public Health 8:38. doi: 10.3389/fpubh.2020.00038

One Health surveillance of antimicrobial resistance (AMR) depends on a harmonized method for detection of AMR. Metagenomics-based surveillance offers the possibility to compare resistomes within and between different target populations. Its potential to be embedded into policy in the future calls for a timely and integrated knowledge dissemination strategy. We developed a blended training (e-learning and a workshop) on the use of metagenomics in surveillance of pathogens and AMR. The objectives were to highlight the potential of metagenomics in the context of integrated surveillance, to demonstrate its applicability through hands-on training and to raise awareness to bias factors[1]. The target participants included staff of competent authorities responsible for AMR monitoring and academic staff. The training was organized in modules covering the workflow, requirements, benefits and challenges of surveillance by metagenomics. The training had 41 participants. The face-to-face workshop was essential to understand the expectations of the participants about the transition to metagenomics-based surveillance. After revision of the e-learning, we released it as a Massive Open Online Course (MOOC), now available at https://www.coursera.org/learn/metagenomics. This course has run in more than 20 sessions, with more than 3,000 learners enrolled, from more than 120 countries. Blended learning and MOOCs are useful tools to deliver knowledge globally and across disciplines. The released MOOC can be a reference knowledge source for international players in the application of metagenomics in surveillance.

Keywords: surveilance, metagenomics, MOOC, antimicrobial resistance, one health

[1]Metagenomics Training Report. Available Online at: http://www.effort-against-amr.eu/page/metagenomics-training-report.php.

INTRODUCTION

The dissemination of knowledge on antimicrobial resistance (AMR) is, like AMR itself, a global, transversal challenge, and needs to be tackled in collaboration between the public health, veterinary and food systems, i.e., in a One Health or integrated approach. A One Health AMR surveillance is challenged by the need to coordinate between surveillance programmes, distinct for each sector. It is therefore important to develop harmonized methods for detection of AMR determinants across sectors (1). In Europe, several initiatives are contributing to the development of integrated AMR surveillance, including the European Epidemiologic Network (Epi-NET),[2] the European Union Joint Programming Initiative on Antimicrobial Resistance (JPIAMR)[3], the Joint Interagency Antimicrobial Consumption and Resistance Analysis (JIACRA)[4] and the EU One Health Action Plan against AMR[5].

The development of integrated surveillance depends on the definition of AMR itself and the choice of a quantitative measure that can be used for comparisons within and between different target populations. AMR can be defined based on established phenotypic thresholds (i.e., interpretation of minimum inhibitory concentration (MIC) or inhibition zone according to specific guidelines [e.g., CLSI and EUCAST]) and based on gene-centric definitions (2). Traditional AMR surveillance relies on the monitoring of phenotypic AMR in indicator organisms (e.g., *Escherichia coli*) and selected pathogens (e.g., serotypes of *Salmonella enterica* subsp. *enterica*), while in metagenomic studies the definition of AMR is gene-centric.

Recent studies have shown that gene-centric AMR monitoring using whole genome sequencing (WGS) of isolates can be highly concordant with observed phenotypic resistance (3–6), although at different levels of accuracy between antibiotic classes. Gene-centric approaches allow to differentiate whether AMR is due to the presence of acquired resistance genes or due to mutations in chromosomal genes, and to identify genes embedded into mobile genetic elements, which are transferable among bacteria.

Although such findings encourage the implementation of WGS in AMR monitoring (7), WGS remains a culture-based method, which challenges the production of real-time actionable information.

Shotgun metagenomics is the culture-independent, untargeted sequencing of all DNA present in a sample, and it therefore offers the possibility to investigate taxonomic composition (including viable and non-viable, culturable, and non-culturable organisms), to predict microbial functions (including AMR) and to recover whole genome sequences (8) (which may reveal yet undiscovered reservoirs of ARGs). A gene-centric, culture-independent method, such as metagenomics allows monitoring AMR with a common measure across surveillance programs, which is independent of the choice of sector-specific indicator- and pathogenic-organisms. Indicator organisms, such as *E. coli*, have often been selected due to convenience and scalability, and not necessarily for being the most appropriate organism to monitor overall AMR trends in a microbial community. Furthermore, it is possible with metagenomics to investigate interactions between species in a microbial community (9) which may determine the occurrence of resistant organisms. Finally, it also has the potential to overcome infrastructure limitations hampering reliable AMR surveillance in low- and middle-income countries, since it requires less tightly controlled environmental conditions and less diversified laboratory supplies compared to traditional microbiology methods (10). Finally, metagenomics surveillance yields data in a standardized format that can be stored and shared electronically with overall modest investments.

There are however shortcomings and biasfactors that need to be taken into account when applying metagenomics (11). The results may be biased due to sampling (including the sample matrix) (9, 12), and the community composition can be affected by sample handling (12, 13). Furthermore, DNA extraction (12, 14, 15), sequencing library preparation (16), the sequencing technology (17, 18), the bioinformatics approach (19), the reference databases used (2), and downstream statistical analyses (20) may also bias results. Finally, there are concerns related to the low sensitivity of metagenomics, which probably constitutes the main obstacle to its application in AMR surveillance. There is an obvious need for benchmarking studies targeting different steps of the process and it is essential to be aware of the importance of method validation and protocol harmonization.

AMR surveillance is a complex topic under rapid scientific development, and the potential to embed new methods into policy in the future calls for an appropriate knowledge dissemination strategy. Open online education (e-learning) is an effective, flexible, and cost-efficient way to disseminate knowledge to a large and diverse range of target learners, at a global level. The delivery of online courses has been greatly facilitated by web-based platforms that host massive open online courses (MOOCs), generally offered free of charge (21). Blended learning, i.e., a mix of training delivery formats, allows for the combination of traditional conceptual lectures delivered through e-learning with face-to-face sessions of hands-on work with tutor support[1]. This facilitates learning in topics where practical data analysis and data interpretation are relevant, and

[2]EPI-Net: Epidemiologic network. Available online at: https://www.combacte.com/about/epi-net.

[3]Joint Programming Initiative on Antimicrobial Resistance (JPIAMR). Available online at: https://www.jpiamr.eu.

[4]Analysis of antimicrobial consumption and resistance ('JIACRA' reports). Available online at: https://www.ema.europa.eu/en/veterinary-regulatory/overview/antimicrobial-resistance/analysis-antimicrobial-consumption-resistance-jiacra-reports.

[5]Action at EU level: The new EU One Health Action Plan against Antimicrobial Resistance. Available online at: https://ec.europa.eu/health/amr/action_eu_en.

additionally facilitates discussions and networking between course participants.

There are several internationally available MOOCs covering the topics of antimicrobial resistance (21),[6,7,8,9], genomics[10,11,12,13,14] or One Health[15]. However, to the best of our knowledge, there are no current initiatives to provide information and training on the use of metagenomics in the context of AMR surveillance, particularly in a transdisciplinary way (i.e., covering topics from sampling strategy to data analysis).

The goal of the European project Ecology from Farm to Fork Of Microbial drug Resistance and Transmission (EFFORT) is to provide scientific evidence on the epidemiology and consequences of AMR in the food chain, while implementing metagenomics[16] (22, 23). Within the scope of EFFORT, we developed a blended training programme on the use of metagenomics in surveillance of pathogens and AMR to (1) Highlight the potential of metagenomics in a global, integrated surveillance context, (2) Demonstrate its applicability by providing hands-on training on a surveillance case-study, and (3) Raise awareness for the factors that may bias metagenomics results[1]. The training consisted of an e-learning component delivered 1 month ahead of a one-and-a-half-day hands-on workshop. After the workshop, we re-evaluated and revised the e-learning, before its stand-alone launch as a MOOC[1].

PEDAGOGICAL FRAMEWORK

The blended training programme consisted of an e-learning component and a one-time face-to-face workshop. The resources used for development of lectures and practical exercises included peer-reviewed scientific publications and the instructors' own expertise. The instructors' background included a variety of disciplines, such as bioinformatics, microbiology, epidemiology, and veterinary medicine[1]. The target group of learners included staff of competent authorities responsible for AMR monitoring (i.e., veterinary services, food safety authorities and reference laboratories), as well as academic staff[1].

The development of the training was led by the Research Group for Genomic Epidemiology at the National Food Institute, Technical University of Denmark (DTU FOOD), which is the EU reference laboratory for antimicrobial resistance (EURL-AR) and comprises multidisciplinary expertise relevant to metagenomics-based AMR surveillance. The objective was to cover the different stages of the workflow in metagenomics-based surveillance, providing the learners with a practical overview of how to conduct each step[1]. Individual lectures from all instructors were subject to peer-review, to avoid overlaps and ensure message consistency.

PROGRAMME DEVELOPMENT AND DELIVERY

Pedagogical Format

E-Learning

The online course was originally organized in "four modules intended to be delivered over 4 weeks, with a separate graded assessment after each module. The modules were: (1) *Introduction*, (2) *From sampling to sequencing*, (3) *From reads to results*, and (4) *Potential of metagenomics for surveillance*. On average, the expected learning time per week was 2 h" minimum[1].

The course was implemented and delivered in the platform Coursera[17], which gathers e-learning courses from the world's top universities and education providers[1]. Before its delivery to the workshop participants, it was offered to a private group of volunteers, in order to gather feedback. The e-learning was released 1 month before the workshop. The e-learning component was subsequently revised and adapted to a MOOC, with the title "Metagenomics applied to surveillance of pathogens and antimicrobial resistance," and it is freely available at https://www.coursera.org/learn/metagenomics. On Coursera, public courses run in 4-weeks sessions, and learners in the same session work through the course together. Sessions start automatically on a regular schedule, and enrolment for each session opens and closes automatically[1].

Table 1 summarizes the course structure and content, as it is presently available online. E-learning elements include video lectures, in-video quizzes, complementary reading, case-study reports and module assessment quizzes. "Lectures are delivered in English, with English subtitles, and pdfs from every lecture are available from the course page. In most videos, non-graded quizzes are included to ensure the engagement of the learners in the lecture and consolidate the learning of key concepts. Reading elements are provided as a complement to most lectures to reinforce the knowledge transmitted, and eventually provide

[6]Antimicrobial resistance–theory and methods. Available online at: https://www.coursera.org/learn/antimicrobial-resistance.

[7]Antimicrobial & Antibiotic Resistance Courses. Available online at: https://www.futurelearn.com/courses/categories/health-and-psychology-courses/antimicrobial-and-antibiotic-resistance.

[8]Antimicrobial Resistance in the Food Chain. Available online at: https://www.futurelearn.com/courses/antimicrobial-resistance-food-chain.

[9]Bacterial Genomes: Disease Outbreaks and Antimicrobial Resistance. Available online at: https://www.futurelearn.com/courses/introduction-to-bacterial-genomics.

[10]Genomic Data Science Specialization. Available online at: https://www.coursera.org/specializations/genomic-data-science.

[11]Whole genome sequencing of bacterial genomes–tools and applications. Available online at: https://www.coursera.org/learn/wgs-bacteria.

[12]Bacterial Genomes: Accessing and Analysing Microbial Genome Data. Available online at: https://www.futurelearn.com/courses/bacterial-genomes-access-and-analysis.

[13]Whole Genome Sequencing: Decoding the Language of Life and Health. Available online at: https://www.futurelearn.com/courses/whole-genome-sequencing.

[14]Bacterial Genomes: From DNA to Protein Function Using Bioinformatics. Available online at: https://www.futurelearn.com/courses/bacterial-genomes-bioinformatics.

[15]One Health: Connecting Humans, Animals and the Environment. Available online at: https://www.futurelearn.com/courses/one-health.

[16]Ecology from Farm to Fork Of Microbial drug Resistance and Transmission (EFFORT). Available online at: http://www.effort-against-amr.eu.

[17]Coursera. Available online at: https://about.coursera.org.

TABLE 1 | MOOC structure and content and corresponding learners' feedback (accessed 31/01/2020).

Module	Elements	Topic	Lecture	Likes	Dislikes
	1 lecture 2 readings		Welcome lecture	97	2
From sampling to sequencing	9 lectures 9 readings	Introduction to metagenomics and antimicrobial resistance	Introduction to Metagenomics	72	
			Considerations and controls for metagenomic/microbiome projects	52	
			Introduction to antimicrobial resistance	49	
		Sampling and sample handling	Sampling for surveillance	38	
			Sampling at farms and slaughterhouses	30	2
			Sample storage	19	1
		DNA and RNA extraction methods	Isolation of DNA from complex samples	27	
			Sample processing for viral metagenomics	11	1
		Sequencing	Notes on library preparation	11	
			Sequencing platforms	29	
Module 1 assessment	2 quizzes 2 readings			59	3
From reads to results	6 lectures 5 readings	Bioinformatics concepts and tools for metagenomics analysis	General intro to bioinformatics analysis of metagenomics data	24	3
			Overview of available metagenomics analysis tools	23	5
			MG mapper	35	1
			ResFinder database	20	
			Demo of metagenomic classification using KRAKEN	12	1
			Real example of metagenomic analysis–lessons learned	13	1
Module 2 assessment	1 quiz 6 readings			25	1
Interpretation of results and potential of metagenomics for surveillance	5 lectures 6 readings	Interpretation of results and application of metagenomics in surveillance	Virtual machine setup	5	
			Analysis and visualization of read count data	12	
			Metagenomic assembly and binning–reconstructing genomes from reads	23	1
			Application of metagenomics in surveillance–methods	20	
			Application of metagenomics in surveillance–opportunities and challenges	15	
Module 3 assessment	1 quiz 3 readings			13	1
Final assessment	5 quizzes 7 readings			23	8
	1 lecture		Farewell lecture	9	

Likes/dislikes for each topic include lecture videos and corresponding reading(s), or all elements of a module assessment.

additional information on the topic. Also, a glossary of the terms used during the course is provided in the first module"[1].

The course assessment is divided in three module-specific graded multiple option quizzes and a final quiz. Each module quiz includes "questions to assess the theoretical knowledge obtained in the corresponding module, and questions based on a surveillance case-study, transversal to the overall course"[1]. The case-study material includes an outline of the exercise step at each module, and module-specific reports for interpretative analysis. "Quiz questions are presented in a multiple-choice format, some with a single correct answer, and others with multiple correct options. In order to complete a module successfully, the learners are required to answer 80% of the quiz correctly"[1].

The final assessment quiz includes questions which require hands-on work by the learners, similarly to what was required to the workshop participants. This is expected to improve the active learning potential of the MOOC. Tutorials for the different steps of the final quiz (*virtual machine setup, introduction, sampling, quality control, bioinformatics analysis of metagenomics results* and *analysis of metagenomics results in a surveillance context*) are provided as additional course elements.

Workshop

Part of the workshop program was based on a recapitulation of the e-learning and the remaining consisted on new content, particularly hands-on training, with exercise sessions following a case-study. Workshop lectures were complemented with discussion sessions, which were distributed throughout the programme in order to foster the exchange of impressions among participants. Two quizzes, at the beginning and at the end of the workshop, were used in order to collect the background information

of the participants, their feedback on the training and their opinion on the use of metagenomics for AMR surveillance. A report on the blended training is available at the EFFORT website[1].

The participants worked in groups during the exercises. "A virtual machine (including user guide) was built for the purpose of the workshop to make use of specific software"[1], including FastQC (24) for quality control, MGmapper (25) for read classification and R (26) for read count analysis and epidemiological analysis. The participants were also introduced to and had the opportunity to apply Linux command-line. They were provided with fictional metagenomics and epidemiological data of a hypothetical case-study in order to perform the analyses. Teaching materials are publicly available at Metagenomics Training Report[1].

Learning Environment

The e-learning was first delivered in a pilot session to a group of 14 volunteers from the EFFORT consortium to gather feedback before launching. After launching, it was delivered to a group of 155 registered learners, including all workshop participants[1].

"A total number of 41 participants and 7 speakers from 14 countries attended the workshop"[1]. Most participants had a research and microbiology background, and were employed at University (52%) or at a Government research institute (32%). Competent authorities (5%) and the Industry (5%) were also represented among participants. The two top reasons for registering on the workshop were "a general interest in the topic" and "a continuing education for the current job." These were followed by "informing current research" and "continuing education for a future job."

By January 2020, 52.0% of the MOOC enrolled learners were students, and the percentage holding a post-graduate degree, Master's (33.0%) or Doctorate (29.9%), was above Coursera averages, 25.7 and 4.09%, respectively. The learners originated relatively more from Europe (32.3%), Africa (9.6%) and Oceania (3.1%), and less from Asia (24.9%), North America (22.7%) and South America (7.3%) compared to Coursera corresponding averages.

Learning Objectives

The learning objectives cover the basics of metagenomics and the background knowledge necessary to consider the implementation of metagenomics in surveillance. They are enumerated for each MOOC module below, as published in the course platform[17].

Module 1:

- "Distinguish between the concepts of metagenomics and other microbial genomics
- Give examples of the application of metagenomics
- Critique the need to use controls in different steps of a metagenomics study
- List types of controls that can be used in a metagenomics study
- Conclude on the advantages of metagenomics for the surveillance of antimicrobial resistance
- Evaluate how sampling design, sample size, sample material and sample handling influence the outcome of a metagenomics study
- Describe current sample processing for bacterial and viral metagenomics
- Explain different sequencing platforms and their possibilities regarding metagenomics
- Summarize the impact that library preparation may have on metagenomics results."

Module 2:

- "Demonstrate the steps involved in a general bioinformatics analysis, including quality control and mapping to different databases
- Outline the principle behind various tools available for analysis of metagenomics data and explain the situations where each tool is appropriate to use
- Interpret the outputs of bioinformatics pipelines (read classification for antimicrobial resistance genes and bacterial species)
- Interpret the possibilities to use a database of antimicrobial resistance genes."

Module 3:

- "Justify the need for epidemiology in surveillance
- Discriminate challenges for the use of metagenomics in surveillance
- Examine the potential of metagenomics for surveillance of pathogens and antimicrobial resistance
- Explain the concept of global and integrated surveillance
- Conclude on metagenomics findings together with explanatory data
- Employ methods for analysis and visualization of read counts."

Assessment

E-learning lecture- and quiz-specific feedback was retrieved from the trial run with volunteers. "The main outcome in terms of course improvement was the development of complementary reading material summarizing the content of the lectures, and the compilation of a glossary"[1]. Both were added to the revised e-learning version, before release as a MOOC. The Coursera platform offers several possibilities for learners' feedback. Module-specific feedback obtained from MOOC learners is presented in **Table 1** including "likes" and "dislikes" given for each course element[1].

"Additionally, an interactive voting tool[18] was used during the workshop, at the end of each day, in order to collect feedback on both components of the training"[1]. 58% of all workshop participants had completed the e-learning and 7% planned to complete it after the workshop. 77% considered the blended learning more useful than a stand-alone e-learning or workshop. An online questionnaire was also used for the evaluation of the workshop and for collecting the participants' opinions on the workshop topic. Response rate was 80.5% (33/41 participants). Respondents assessed again positively the combination of the

[18]Mentimeter. Available Online at: https://www.mentimeter.com.

e-learning and the workshop, considering the workshop as an essential component of the training package. However, many would have liked to have longer practical sessions[1].

At the time of writing (January 2020), the MOOC has run in 22 consecutive 4-weeks sessions, with a total of 3,346 learners enrolled, including 2,180 active learners (enrolled learners who have started a course item), of which 186 passed all assessments and were issued a course certificate. It has been rated as 4.7/5, with 95% of likes and 5% of dislikes. The highest drop rate among all eligible learners (81.9%) is in module 1. This is not surprising, as we expected most learners to explore the course content before deciding to complete it. Furthermore, it is in accordance with the 90-9-1 rule that describes most participation in online communities (90% consume content, 9% engage with content sporadically, and 1% regularly) (27).

DISCUSSION

Lessons Learned

At the end of the workshop, the majority of the participants (90.2%) responded that they expected the use of metagenomics in AMR surveillance to increase, slowly (63.4%) or rapidly (26.8%), in the near future (**Figure 1**). The participants were asked to assess the main challenges and gaps for the implementation of metagenomics in surveillance (**Figure 2**), and the results showed that harmonization of protocols and interpretation of results (including uncertainty and association of metagenomics data with risk factors) are considered main hurdles. The lack of standards and legislation, and the implementation costs were also mentioned. Infrastructure challenges, such as data sharing and storage were considered less relevant. Improvement of metagenomics analysis was also considered by the participants the priority in order to increase the understanding of AMR. However, the improvement of surveillance programmes and international guidelines, and an increase in harmonized reporting were considered similarly important (results not shown). Food safety risk assessment was clearly the area where participants considered metagenomics will have the largest impact (**Figure 3**).

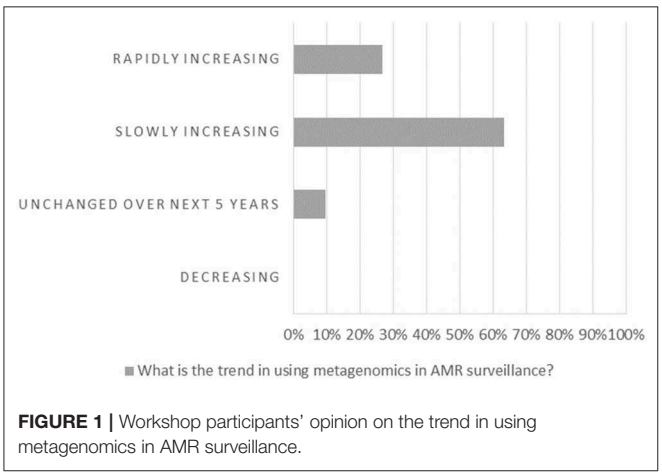


FIGURE 1 | Workshop participants' opinion on the trend in using metagenomics in AMR surveillance.

Practical Implications

The future of antimicrobial resistance surveillance needs to be tackled with a multinational, multidisciplinary One Health approach (1, 21). While many countries are already engaging in the use of whole genome sequencing for surveillance (9), outbreak investigation, source-attribution and microbial risk assessment (11), the implementation of metagenomics in those areas still resides in the future due to its novelty, among other reasons.

One of the main concerns about the routine use of metagenomics is that it may lead to a decrease in pathogen isolation from humans and along transmission pathways (9, 11, 28). However, the potential of metagenomics is significant. It allows the detection of pathogens in mixed cultures, the identification of (new) non-culturable pathogens, the characterization of bacterial diversity and its effect on pathogen presence and diversity, and the characterization of resistomes and mobilomes (sequences attributed to mobile genetic elements, involved in horizontal gene transfer). To engage in these diverse aspects of AMR surveillance and future methodological options, professionals from a variety of disciplines should co-develop a joint understanding of the strengths and weaknesses of this approach. Blended learning courses and MOOCs can be successfully applied in this context to deliver knowledge, to provide a platform to engage across disciplines, and to facilitate peer-learning.

The interaction with the course participants provided general information on the readiness of the community for using metagenomics in AMR surveillance. Harmonization of protocols was highlighted as an important challenge by the workshop participants. There is a general concern about the numerous sources of bias in metagenomics studies, and the need for validation and benchmarking exercises is recognized (11). Recently, there is a growing number of studies addressing this concern (29), which represent valuable input for a conscious application of metagenomics in surveillance. The lack of standards and legislation, lack of harmonized reporting and lack of international guidelines were also among the participants' apprehensions. Undeniably, metagenomics conveys sequence data that may contain indication of hazards which would otherwise not be investigated and/or detected with isolate-based monitoring methods. Additionally, host sequence data can potentially allow the identification of human subjects. These issues must be addressed in international guidelines developed for the ethical use of metagenomics (28). Improvement of metagenomics analysis was considered by the training participants the first priority in order to increase the understanding of AMR-related outputs (results not shown). "Improvement" may in this context relate to different factors that are considered potential limitations of metagenomics studies. One of the main challenges is that the detected DNA can originate from both dead and alive cells, which may be perceived as a shortcoming in the context of policy-based monitoring and risk assessment studies (11). Potential solutions could be to complement metagenomics with metatranscriptomics (28) or to use algorithms that infer microbial population replication rates from metagenomics data (30). However, the detection of

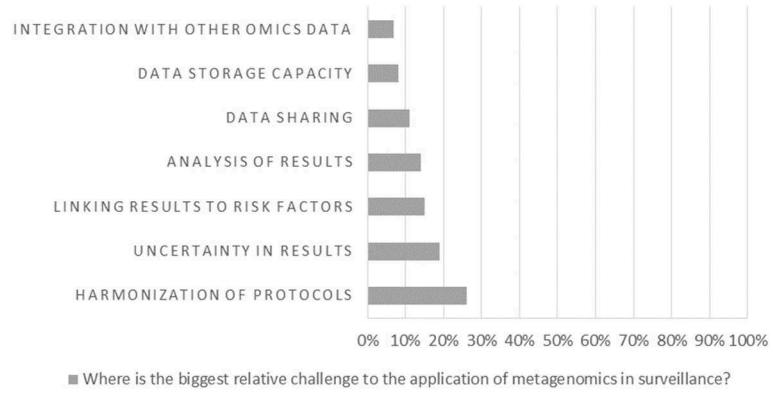


FIGURE 2 | Workshop participants' opinion on the biggest relative challenge to the application of metagenomics in surveillance.

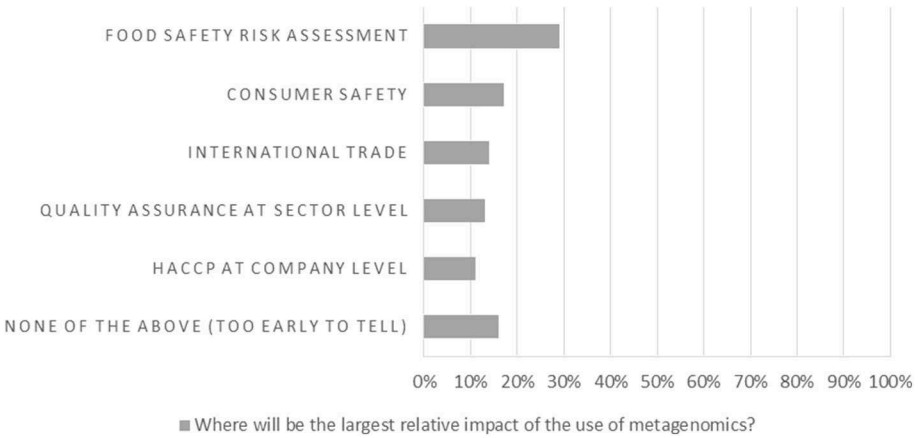


FIGURE 3 | Workshop participants' opinion on where to expect the largest relative impact of the use of metagenomics.

non-viable microorganisms, particularly pathogens, may also be seen as an opportunity. Although dead bacteria may not constitute an immediate risk for the exposed population, their detection is an opportunity to prompt investigation of the source of contamination and to apply corrective preventive measures before transmission occurs. In a surveillance program, detecting the presence of pathogens (eventually carrying high-risk ARGs), viable or not, should therefore be desired–if the microorganisms are viable, their spread can be contained; if they are non-viable, source tracking can be performed and preventive measures applied to avoid infections. Also, attributing detected ARGs to their bacterial host, and classifying their transferability between hosts may be necessary in many circumstances. Metagenomic assembly and binning (31) help overcoming the first issue, and many recent developments have contributed to increase the number of genomes assembled from metagenomics datasets, including the methods of Hi-C Chromatin conformation capture (31), DNA methylation profiling (32) and co-assembly and co-binning (33). A greater challenge remains with the second issue-disclosing the link between ARGs and mobile genetic elements. The joint analysis of resistome and microbiome has been used to investigate the occurrence of horizontal gene transfer, with recent studies suggesting an infrequent exchange of ARGs between human gut flora and pathogenic organisms (34, 35). Another route to address this issue is the use of single cell sequencing (36, 37). A further concern is that the resolution in the profiling of resistomes, i.e., the accuracy of ARG typing, may be insufficient due to a high similarity shared between ARG reference sequences. This may produce ambiguous alignment, false negatives due to non-alignment, or false positives due to misannotation. Recent bioinformatics developments have also addressed this concern (35, 38). Similarly, the low sensitivity of metagenomics to capture low abundant ARGs, has also been recently addressed by combining targeted metagenomics with novel bioinformatics tools for the analysis of resistomes (39), however further developments and validation studies are still needed in order to confidently approach the sensitivity levels presently achieved with phenotypic methods.

Food safety risk assessment and consumer safety might benefit from metagenomics, in the participants' opinion. However, ARGs detected in metagenomics studies should undergo an assessment regarding their public health risk potential, since they

do not all represent an actual hazard (2). The application of metagenomics in risk assessment is therefore dependent on a new hazard definition concept, and the nature of the hazard will determine the nature of the estimated risk. With metagenomics, "hazard" covers the microbial community, the resistome, and the potential for horizontal transmission of ARGs. As a result, risk may refer to the development of disease due to infection with a resistant pathogen, and/or the spread of ARGs between pathogens and commensal bacteria in the human host (40). Traditional microbial risk assessment methods need to undergo an adaptation in order to accommodate these new considerations of hazard and risk (40).

We developed and delivered a blended-training on "Metagenomics applied to surveillance of pathogens and AMR." After the training, the e-learning component was revised and an updated version is now publicly available as a MOOC at https://www.coursera.org/learn/metagenomics[1], on which more than 3,000 learners have already enrolled. The MOOC conveys the idea of the workflow, the requirements, the benefits and the challenges of AMR surveillance by metagenomics, which could help inform the design of future AMR surveillance programs.

Constraints and Future Perspectives

Throughout the training, the main challenge has been to adjust to the variable level of background and skills of the participants. In general, the hands-on training was well-received, both during the workshop, and by the MOOC learners. However, when technical difficulties arise in operating the software programs for data analysis, it is difficult to provide adequate support to those in need. Furthermore, in the context of education at the global level, the uneven access of learners to infrastructures (internet bandwidth, computer processor, operating system and memory) will impact on the learning outcome and the likelihood of course completion. This mirrors one of the expected challenges in the implementation of a metagenomics-based global surveillance–the uneven and variable levels of capacity among countries.

A future perspective for improvement of the MOOC is to provide less technically demanding and infrastructure-dependent practical exercises. Furthermore, we intend to periodically review the course content and update it following the latest research developments. For example, many studies have recently investigated the impact of different normalization approaches for metagenomics data (41–43), a topic that has not been addressed in the current MOOC version. With future content updates, the course will maintain a high educative value and can be established as a reference international source of information for the implementation of metagenomics in surveillance.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

AUTHOR CONTRIBUTIONS

TH and AD conceived the e-learning curriculum with input from PM, OL, SP, VB, JN, PK, LV, CP, RK, MH, and RH. AD coordinated the development of the MOOC. AD, TH, PM, SP, VB, OL, JN, PK, LV, CP, RK, MH, and RH developed and recorded the lectures for the MOOC. KS conceived the workshop curriculum with input from AD, PM, AB, RL, SS, LV, and PL. KS and VG coordinated the development of the workshop in collaboration with AD, and were responsible for advertisement, registration of participants and development of participant surveys. OL, PM, AB, RL, SS, LV, PL, and AD developed the lectures and practical exercises for the workshop. AD was in charge of overall direction and planning of the training and wrote the manuscript with critical feedback from all authors.

FUNDING

This work was part of the EC FP7 project Ecology from Farm to Fork of Microbial Drug Resistance and Transmission (grant agreement No. 613754). The MOOC content counted with a large contribution of collaborators from the EC Horizon 2020 project COllaborative Management Platform for detection and Analyses of (Re-) emerging and foodborne outbreaks in Europe (grant agreement No. 643476). Both projects focus on the use of next generation sequencing in surveillance of pathogens and AMR.

ACKNOWLEDGMENTS

We thank all media partners who advertised the training in their channels and all the participants who provided feedback on the training.

REFERENCES

1. Aarestrup FM. The livestock reservoir for antimicrobial resistance: a personal view on changing patterns of risks, effects of interventions and the way forward. *Philos Trans Royal Soc Lond B Biol Sci*. (2015) 370:20140085. doi: 10.1098/rstb.2014.0085
2. Martínez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. *Nat Rev Microbiol*. (2014) 13:116–23. doi: 10.1038/nrmicro3399
3. Do Nascimento V, Day MR, Doumith M, Hopkins KL, Woodford N, Godbole G, Jenkins C. Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of enteroaggregative *Escherichia coli* isolated from cases of diarrhoeal disease in England, 2015–16. *J Antimicrob Chemother*. (2017) 72:3288–97. doi: 10.1093/jac/dkx301
4. Moran RA, Anantham S, Holt KE, Hall RM. Prediction of antibiotic resistance from antibiotic resistance genes detected in antibiotic-resistant commensal *Escherichia coli* using PCR or WGS. *J Antimicrob Chemother*. (2017) 72:700–4. doi: 10.1093/jac/dkw511.
5. Stubberfield E, AbuOun M, Sayers E, O'Connor, HM, Card RM, Anjum MF. Use of whole genome sequencing of commensal *Escherichia coli* in pigs for antimicrobial resistance surveillance, United Kingdom, 2018. *Euro Surveill*. (2019) 24:1900136. doi: 10.2807/1560-7917.ES.2019.24.50.1900136

6. Guo S, Tay MYF, Aung KT, Seow KLG, Ng LC, Purbojati RW, et al. Phenotypic and genotypic characterization of antimicrobial resistant *Escherichia coli* isolated from ready-to-eat food in Singapore using disk diffusion, broth microdilution and whole genome sequencing methods. *Food Control*. (2019) 99:89–97. doi: 10.1016/j.foodcont.2018.12.043
7. EFSA (European Food Safety Authority), Aerts M, Battisti A, Hendriksen R, Kempf I, Teale C, et al. (2019). Scientific report on the technical specifications on harmonised monitoring of antimicrobial resistance in zoonotic and indicator bacteria from food-producing animals and food. *EFSA J*. (2019) 17:5709. doi: 10.2903/j.efsa.2019.5709
8. Quince C, Walker AW, Simpson JT, Loman NJ, Segata N. Shotgun metagenomics, from sampling to analysis. *Nat Biotechnol*. (2017) 35:833–44. doi: 10.1038/nbt.3935
9. Jagadeesan B, Gerner-Smidt P, Allard MW, Leuillet S, Winkler A, Xiao Y, et al. The use of next generation sequencing for improving food safety: translation into practice. *Food Microbiol*. (2019) 79:96–115. doi: 10.1016/j.fm.2018.11.005
10. Nkengasong JN, Yao K, Onyebujoh P. Laboratory medicine in low-income and middle-income countries: progress and challenges. *Lancet*. (2018) 391:1873–5. doi: 10.1016/S0140-6736(18)30308-8
11. EFSA BIOHAZ Panel (EFSA Panel on Biological Hazards), Koutsoumanis K, Allende A, Alvarez-Ordóñez A, Bolton D, Bover-Cid S, et al. Scientific Opinion on the whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms. *EFSA J*. (2019) 17:5898. doi: 10.2903/j.efsa.2019.5898
12. Knudsen BE, Bergmark L, Munk P, Lukjancenko O, Priemé A, Aarestrup FM, Pamp SJ. Impact of sample type and DNA isolation procedure on genomic inference of microbiome composition. *MSystems*. (2016) 1: e00095-16. doi: 10.1128/mSystems.00095-16
13. Walker AW, Martin JC, Scott P, Parkhill J, Flint HJ, Scott KP. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome*. (2015) 3:26. doi: 10.1186/s40168-015-0087-4
14. McOrist AL, Jackson M, Bird AR. A comparison of five methods for extraction of bacterial DNA from human faecal samples. *J Microbiol Methods*. (2002) 50:131–9. doi: 10.1016/S0167-7012(02)00018-0
15. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to basics–the influence of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities. *PLoS ONE*. (2015) 10:e0132783. doi: 10.1371/journal.pone.0132783
16. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques*. (2014) 56:61–4. doi: 10.2144/000114133
17. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. *J Biomed Biotechnol*. (2012) 2012:11. doi: 10.1155/2012/251364
18. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*. (2016) 17:333–51. doi: 10.1038/nrg.2016.49
19. De Filippis F, Parente E, Zotta T, Ercolini D. A comparison of bioinformatic approaches for 16S rRNA gene profiling of food bacterial microbiota. *Int J Food Microbiol*. (2018) 265:9–17. doi: 10.1016/j.ijfoodmicro.2017.10.028
20. Sudarikov K, Tyakht A, Alexeev D. Methods for the metagenomic data visualization and analysis. *Curr Issues Mol Biol*. (2017) 24:37–58. doi: 10.21775/cimb.024.037
21. Sneddon J, Barlow G, Bradley S, Brink A, Chandy SJ, Nathwani D. Development and impact of a massive open online course (MOOC) for antimicrobial stewardship. *J Antimicrob Chemother*. (2018) 73:1091–7. doi: 10.1093/jac/dkx493
22. Munk P, Knudsen BE, Lukjancenko O, Duarte ASR, Van Gompel L, Luiken REC, et al. Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. *Nat Microbiol*. (2018)3:898–908. doi: 10.1038/s41564-018-0192-9
23. Van Gompel L, Luiken REC, Sarrazin S, Munk P, Knudsen BE, Hansen RB, et al. The antimicrobial resistome in relation to antimicrobial use and biosecurity in pig farming, a metagenome-wide association study in nine European countries. *J Antimicrob Chemother*. (2019) 74:865–76. doi: 10.1093/jac/dky518
24. Andrews S. (2010). *FastQC: A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
25. Petersen TN, Lukjancenko O, Thomsen MCF, Maddalena Sperotto M, Lund O, Møller Aarestrup F, et al. MGmapper: reference based mapping and taxonomy annotation of metagenomics sequence reads. *PLoS ONE*. (2017) 12:e0176469. doi: 10.1371/journal.pone.0176469
26. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. (2018) Available Online at: https://www.R-project.org
27. Honeychurch S, Bozkurtaras A, Singh L, Koutropoulos A. Learners on the periphery: lurkers as invisible learners. *Eur J Open Dist e-Learning*. (2017) 20:192–212. doi: 10.1515/eurodl-2017-0012
28. Bergholz TM, Moreno Switt AI, Wiedmann M. Omics approaches in food safety: fulfilling the promise? *Trends Microbiol*. (2014) 22:275–81. doi: 10.1016/j.tim.2014.01.006
29. Sczyrba A, Hofmann P, Belmann P, Koslicki D, Janssen S, Dröge J, et al. Critical assessment of metagenome interpretation—a benchmark of metagenomics software. *Nat Methods*. (2017) 14:1063–71. doi: 10.1038/nmeth.4458
30. Brown CT, Olm MR, Thomas BC, Banfield JF. Measurement of bacterial replication rates in microbial communities. *Nat Biotechnol*. (2016) 34:1256. doi: 10.1038/nbt.3704
31. Stewart RD, Auffret MD, Warr A, Wiser AH, Press MO, Langford KW, et al. Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun*. (2018) 9:870. doi: 10.1038/s41467-018-03317-6
32. Beaulaurier J, Zhu S, Deikus G, Mogno I, Zhang XS, Davis-Richardson A, et al. Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation. *Nat Biotechnol*. (2018) 36:61–69. doi: 10.1038/nbt.4037
33. Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. (2019) 176:649–62.e20. doi: 10.1016/j.cell.2019.01.001
34. Sommer M, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science*. (2009) *325*:1128–31. doi: 10.1126/science.1176950
35. Ruppé E, Ghozlane A, Tap J, Pons N, Alvarez AS, Maziers N, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. *Nat Microbiol*. (2019) 4:112–123. doi: 10.1038/s41564-018-0292-6
36. Spencer SJ, Tamminen MV, Preheim SP, Guo MT, Briggs AW, Brito IL, et al. Massively parallel sequencing of single cells by epicPCR links functional genes with phylogenetic markers. *ISME J*. (2015) 10:427–436. doi: 10.1038/ismej.2015.124
37. Lan F, Demaree B, Ahmed N, Abate, AR. Single-cell genome sequencing at ultra-high-throughput with microfluidic droplet barcoding. *Nat Biotechnol*. (2017) 35:640–6. doi: 10.1038/nbt.3880
38. Rowe WPM, Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. *Bioinformatics*. (2018) 34:3601–8. doi: 10.1093/bioinformatics/bty387
39. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, et al. In-depth resistome analysis by targeted

metagenomics. *Microbiome*. (2018) 6:11. doi: 10.1186/s40168-017-0387-y
40. Pires SM, Duarte AS, Hald T. Source attribution and risk assessment of antimicrobial resistance. *Microbiol Spectr*. (2018) 6. doi: 10.1128/microbiolspec.ARBA-0027-2017
41. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*. (2017) 5:27. doi: 10.1186/s40168-017-0237-y
42. Du R, An L, Fang Z. Performance evaluation of normalization approaches for metagenomic compositional data on differential abundance analysis. In: Zhao Y, Chen DG, editors. *New Frontiers of Biostatistics and Bioinformatics. ICSA Book Series in Statistics*. Cham: Springer (2018) doi: 10.1007/978-3-319-99389-8_16
43. Pereira MB, Wallroth M, Jonsson V, Kristiansson E. Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics*. (2018) 19:274. doi: 10.1186/s12864-018-4637-6

Conflict of Interest: KS and VG were employed by the company SAFOSO AG. RH was employed by the company Intomics A/S.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: info@frontiersin.org | +41 21 510 17 00

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership