Chapter 5
The CGE Tool Box

Mette Voldby Larsen, Katrine G. Joensen, Ea Zankari, Johanne Ahrenfeldt, Oksana Lukjancenko, Rolf Sommer Kaas, Louise Roer, Pimlapas Leekitcharoenphon, Dhany Saputra, Salvatore Cosentino, Martin Christen Frølund Thomsen, Jose Luis Bellod Cisneros, Vanessa Jurtz, Simon Rasmussen, Thomas Nordahl Petersen, Henrik Hasman, Thomas Sicheritz-Ponten, Frank M. Aarestrup, and Ole Lund

Introduction

Human and animal health worldwide is increasingly threatened by new and re-emerging epidemics and foodborne pathogens, placing a burden on health and veterinary systems, reducing consumer confidence in food, and negatively affecting trade, food chain sustainability and food security. Rapid identification of emerging and foodborne pathogens and subsequent provision of timely insights into the modes of transmission, prevention, and control, pathogenesis, and clinical impact of such diseases is essential to reduce the impact, time, and costs of disease outbreaks.

M.V. Larsen (✉) • J. Ahrenfeldt • D. Saputra • M.C.F. Thomsen • J.L.B. Cisneros • V. Jurtz • S. Rasmussen • T.N. Petersen • T. Sicheritz-Ponten • O. Lund
Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
e-mail: metteb@cbs.dtu.dk; johah@cbs.dtu.dk; dhany@cbs.dtu.dk; mcft@cbs.dtu.dk; cisneros@cbs.dtu.dk; vanessa@cbs.dtu.dk; simon@cbs.dtu.dk; tnp@cbs.dtu.dk; thomas@cbs.dtu.dk; lund@cbs.dtu.dk

K.G. Joensen • H. Hasman • F.M. Aarestrup
Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark
e-mail: KNJ@ssi.dk; henh@ssi.dk; fmaa@food.dtu.dk

E. Zankari • O. Lukjancenko
Department of Systems Biology, Center for Biological Sequence Analysis, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

National Food Institute, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark
e-mail: east@food.dtu.dk; oklu@food.dtu.dk

X. Deng et al. (eds.), *Applied Genomics of Foodborne Pathogens*, Food Microbiology and Food Safety, DOI 10.1007/978-3-319-43751-4_5

A potential breakthrough is offered by the revolution in genome technology, leading to increasing speed and reducing costs of sequencing. As the common denominator to all pathogens and hosts, regardless of species and domain, is the presence of a genome, the ability to rapidly determine the sequence provides a common language by which data on pathogens can be compared. Such a single technology applicable to different disciplines (bacteriology, virology, parasitology) and domains (human, food, animal, environment) would facilitate global cross-cutting collaboration and information exchange (integrated surveillance), leading to rapid and coordinated responses to novel and known health threats as they emerge [1].

Conditional to this success is the capacity to generate and analyze the complex genome data in a manner that addresses clinical and public health questions reliably and timely. Thus, one of the biggest obstacles for the implementation of Whole Genome Sequence (WGS) data in clinical, animal and food microbiological laboratories is the absence of bioinformatics expertise to handle the vast amount of data. If we can provide reliable real-time bioinformatics services for frontline diagnostics, we might also be able to capture this information globally and thus create real-time global surveillance.

Center for Genomic Epidemiology (GGE) at the Technical University of Denmark was initiated in 2010 to provide a proof of concept for this. The center was funded by a grant from the Danish Council for Strategic Research. It had Prof. Frank M. Aarestrup as the coordinator and Prof. Ole Lund as the deputy coordinator.

Basically, the aim of CGE was to develop methods that use WGS data for discovering the content of a sample (typing), predict its pathogenic potential, and which antimicrobials it might be resistant towards (phenotyping). For epidemiological purposes, it was furthermore the aim to develop methods for examining the evolutionary relationship of the isolate vs. other samples. Table 5.1 provides an overview of the 16 methods that have so far been developed at CGE, and which are all made public available via easy-to-use web-services. Each method is described in more detail in the remainder of the chapter. Further, throughout the chapter the use of the web-services is exemplified with a case story employing WGS data from verotoxigenic *Escherichia coli* (VTEC) (see the boxes).

R.S. Kaas • L. Roer • P. Leekitcharoenphon
National Food Institute, Technical University of Denmark, 2800, Kgs. Lyngby, Denmark
e-mail: rkmo@food.dtu.dk; lroe@food.dtu.dk; pile@food.dtu.dk

S. Cosentino
Department of Infection Metagenomics, Genome Information Research Center, Research Institute for Microbial Diseases, Osaka University, 3-1 Yamadaoka, Suita, Osaka 565-0871, Japan
e-mail: salvocos@gen-info.osaka-u.ac.jp

Table 5.1 Overview of 16 public available web-services from CGE by Oct. 2014

Name of method	Description	URL	Publication
CSIPhylogeny	SNP-based creation of phylogenetic trees	https://cge.cbs.dtu.dk/services/CSIPhylogeny	Published Aug 2014 PMID: 25110940 [2]
KmerFinder	Species identification by co-occurring 16-mers	https://cge.cbs.dtu.dk/services/KmerFinder	Published Jan 2014 PMID: 24172157 [3]
MLST	Multilocus sequence typing	https://cge.cbs.dtu.dk/services/MLST	Published Apr 2012 PMID: 22238442 [4]
MyDbFinder	Identification of genes in user-made database	https://cge.cbs.dtu.dk/services/MyDbFinder	Published here
NDtree	Creation of phylogenetic trees	https://cge.cbs.dtu.dk/services/NDtree	Published Feb 2014 PMID: 24505344 [5]
PanFunPro	Groups homologous proteins based on functional domain content	https://cge.cbs.dtu.dk/services/PanFunPro	Published Dec 2013 [6]
PathogenFinder	Prediction of pathogenic potential	https://cge.cbs.dtu.dk/services/PathogenFinder	Published Oct 2013 PMID: 24204795 [7]
PlasmidFinder	Plasmid identification in *Enterobacteriaceae*	https://cge.cbs.dtu.dk/services/PlasmidFinder	Published Apr 2014 PMID: 24777092 [8]
pMLST	pMLST of plasmids in *Enterobacteriaceae*	https://cge.cbs.dtu.dk/services/pMLST	Published Apr 2014 PMID: 24777092 [8]
Reads2Type	Species identification on client computer	https://cge.cbs.dtu.dk/services/Reads2Type	Published Feb 2014 PMID: 24574292
ResFinder	Identification of acquired antimicrobial resistance genes	https://cge.cbs.dtu.dk/services/ResFinder	Published Nov 2012 PMID: 22782487 [9]
SerotypeFinder	WGS-based serotyping of *Escherichia coli*	https://cge.cbs.dtu.dk/services/serotypefinder	Published May 2015 PMID: 25972421 [10]
SnpTree	SNP-based creation of phylogenetic trees	https://cge.cbs.dtu.dk/services/snpTree	Published Dec 2012 PMID: 23281601 [11]
SpeciesFinder	16S rRNA-based species identification	https://cge.cbs.dtu.dk/services/SpeciesFinder	Published Feb 2014 PMID: 24574292 [12]
TaxonomyFinder	Taxonomy identification using functional protein domains	https://cge.cbs.dtu.dk/services/TaxonomyFinder	Published Feb 2014 PMID: 24574292 [12]
VirulenceFinder	Identification of virulence genes in *E. coli*	https://cge.cbs.dtu.dk/services/VirulenceFinder	Published Feb 2014 PMID: 24574290 [5]

VTEC Case Study
Verocytotoxin-producing *E. coli* (VTEC), also commonly referred to as Shiga toxin-producing *E. coli* (STEC) is a gastrointestinal pathogen, which causes disease due to production of verocytotoxins as well as several other virulence factors [13, 14]. Some VTEC cause severe infection with bloody diarrhea and at times life-threatening Hemolytic Uremic Syndrome (HUS) [15]. Around 5–10 % of VTEC infections lead to the development of HUS, and although most patients recover within a few weeks, it can be fatal or lead to permanent kidney damage. VTEC is usually contracted by ingestion of contaminated food or water, or through person-to-person contact, and it is estimated that 265,000 VTEC infections occur each year in the US [16]. In Denmark, routine typing of VTEC infections for surveillance is carried out at Statens Serum Institut (SSI). It includes serotyping (O:(K):H), which identifies the Lipopolysaccharide (O-antigen), capsular (K) antigen, and the flagellar (H) antigen. Isolates are further examined for β-glucuronidase activity, haemolysin production, and for the production of verocytotoxin, as well as for specific virulence factors, most importantly, verotoxin 1 (*vtx1*), verotoxin 2 (*vtx2*) and intimin (*eae*), which are detected by DNA hybridization, and further subtyping of the verocytotoxins is carried out by PCR. Isolates of the same serotype with similar toxin profiles and phenotypic features that are considered to be potential outbreak isolates are further subjected to PFGE typing for comparison. Due to the many different analysis that are necessary for accurate routine typing and surveillance, it is time-consuming, laborious and costly. Thus, as a proof-of-concept that WGS-based typing of VTEC could be an attractive alternative, real-time WGS-based typing of VTEC were performed during 7 weeks, in parallel to the routine typing carried out at SSI. The study included a set of 46 suspected VTEC isolates and has previously been described [5]. Throughout this chapter, the same set of suspected VTEC isolates are used to exemplify the use and output of selected CGE web-services.

Prokaryotic Taxonomy

From a pragmatic point of view "the ultimate goal of taxonomy is to construct a classification that is of operative and predictive use for any discipline in microbiology and that is also essentially stable" [17]. Most taxonomists agree that phylogeny—the study of the evolutionary history of organisms—should be the underlying basis of taxonomy. Historically, the first attempts on bacterial classification were based on morphology, later the phylogenetic reconstructions were based on physiological properties. A milestone in classification of prokaryotes was the introduction of 16S rRNA sequence data [18] and it has dominated molecular taxonomy since. Tremendous amounts of 16S rRNA gene sequence data are

available in public databases [19, 20]. Several concerns about its use have, however, been raised. These include low resolution [21, 22], the presence of several, and sometimes-different 16S rRNA genes in some genomes [23], and the fact that the 16S rRNA gene only represents a tiny fraction of microbial genomes [24]. The introduction of WGS data enables alternative approaches for prokaryotic classification that utilize a larger portion of the genome. At CGE, a number of methods using WGS data have been implemented. Some examine only the 16S rRNA gene (SpeciesFinder and Reads2Type), while others take a larger portion of the genomes into account (TaxonomyFinder and KmerFinder) [25].

SpeciesFinder

SpeciesFinder predicts prokaryotic species based on the 16S rRNA gene. For this purpose, it uses a database of 16S rRNA genes isolated from reference genomes with annotated species [25]. The predicted species of a query genome is selected as the annotated species of the reference genome with the most similar 16S rRNA gene. The SpeciesFinder web-service exemplifies the most basic version of the generally simple CGE user interface (Fig. 5.1). The user only has to select the input file (short sequence reads in FASTQ format or a draft genome in FASTA format) containing the DNA sequence of the query isolate and click the "Submit" button.

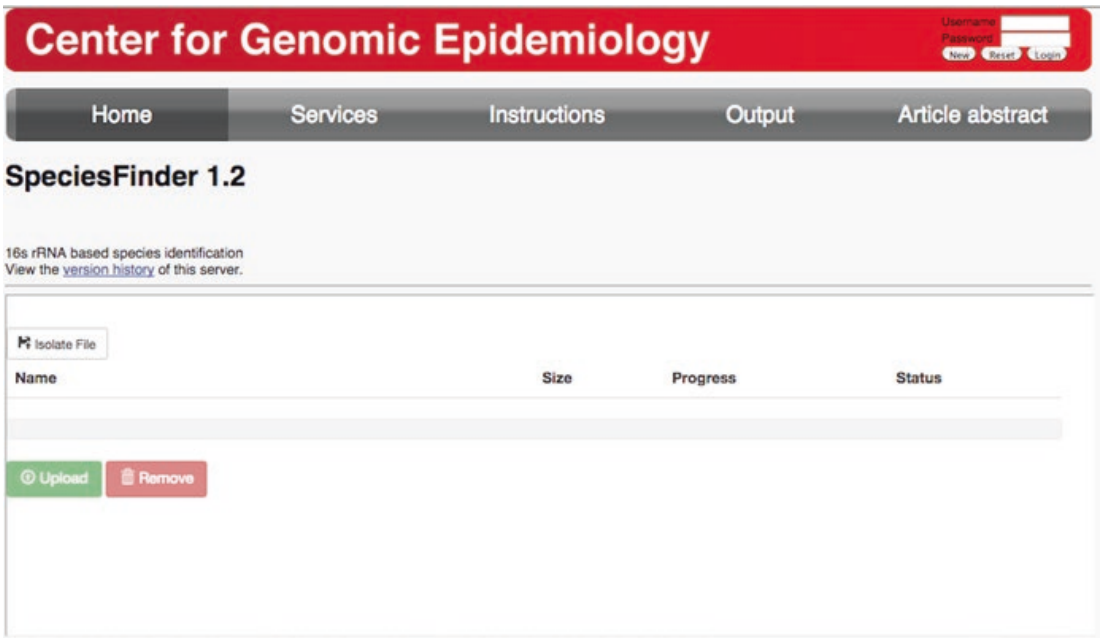


Fig. 5.1 User interface of the SpeciesFinder web-service. Using the "Browse" button the input file (short sequence reads in FASTQ format or a draft genome in FASTA format) containing the DNA sequence of the query isolate is selected. Then, the "Submit" button is clicked

VTEC Case Study: Identifying the Species Using SpeciesFinder
One of the suspected VTEC isolates (C757-12) [5] was run through the SpeciesFinder web-service to confirm the species as *E. coli*. Like the input page of the web-service, the output page is very simple (Fig. 5.2). Besides the predicted species and the GenBank accession number of the reference genome on which the prediction is based, the confidence level of the result is marked as PASS or FAIL; if the prediction is based on a similarity between the 16S rRNA gene of the SpeciesFinder database and the 16S rRNA gene of the query genome above 98 % identity on nucleotide level, the confidence of result is listed as "PASS". Otherwise it is listed as "FAIL".

SpeciesFinder is available at https://cge.cbs.dtu.dk/services/SpeciesFinder.

Reads2Type

Similar to SpeciesFinder, the 16S rRNA gene forms the basis of the Reads2Type method. However, instead of examining similarity across the entire gene, the method employs a small, pre-made database of species-specific 50-mers (stretches of DNA with the length of 50 nucleotides) from within the gene. Further, for the *Enterobacteriaceae* family, the *GyrB* gene is used as the species-specific marker gene, not the 16S rRNA gene. When using the Reads2Type web application, this small database of species-specific 50-mers is automatically transferred into memory and all computations are done on the computer of the user. This is an advantage, since it means that the much larger data amounts that the query genome sequence data represents do not need to be transferred from the computer of the user to the central server. Besides minimizing data transfer, bottleneck problems on the server are also avoided. The minimization of the data transfer may be particularly advantageous for users with limited Internet access [25].

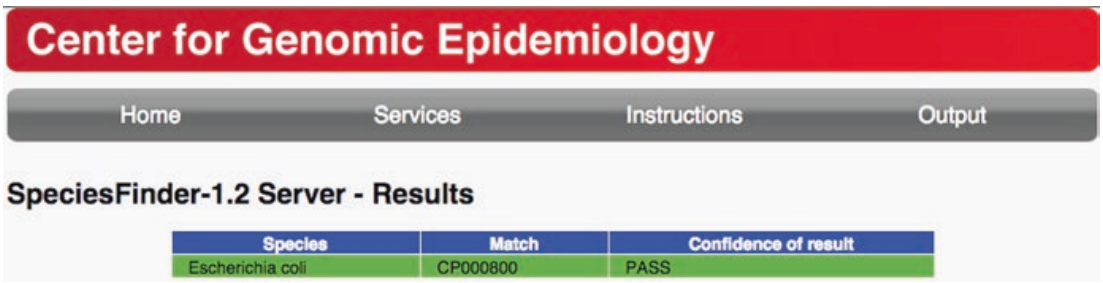


Fig. 5.2 The SpeciesFinder output when the suspected VTEC isolate C757-12 [5] is used as input. The output includes the predicted species (Species), the accession number of the reference genome on which the prediction is based (Match), and the level of confidence of the prediction (Confidence of result)

Reads2Type is available at https://cge.cbs.dtu.dk/services/Reads2Type.

TaxonomyFinder

The pan-genome of a given taxonomic group of genomes (phylum, genus, species) consists of a set of conserved genes, genes that are present in some, but not all genomes, and genes that are specific for particular strains. The typical approach for taxonomy prediction is using evolutionary conserved genes; such as 16S rRNA or a set of 'housekeeping' genes as in MLST (see below). However, taxonomic classification can also be performed using taxa-group specific proteins, which is the approach applied by TaxonomyFinder.

The TaxonomyFinder specific protein database was created using PanFunPro (Pan-genome analysis based on functional profiles) [6], homology detection, and a protein annotation tool. PanFunPro can be used for core-, pan- and accessory genome analysis, such as estimation of life's set of core genes, prediction of chromosome-specific families [26], analysis of differences between probiotic and pathogenic strains [6], as well as estimation of taxonomy-group specific proteins. Briefly, a set of proteins from a number of prokaryotic genomes are searched for functional domains using the InterProScan software [27] against the three Hidden Markov Model (HMM) collections: PfamA, SuperFamily and TIGRFAM. Subsequently, non-repeating and non-overlapping functional domains within each protein are combined into functional profiles, using the information of one database at a time, with respect to the order of scans. Homologous proteins are grouped into protein families, based on functional profiles. Next, taxa-specific profiles are predicted. A profile is considered to be specific, if it is 100 % conserved among the set of query genomes (genomes within a taxonomic group), and absent in the rest of the analyzed genomes. However, this approach may not be feasible if the number of members in the taxonomic group is very high, such as in the *Firmicutes* and *Proteobacteria* phyla, or *Escherichia* genus. Under these circumstances, the requirement is lowered, meaning that profiles remain specific to the taxonomic group, but may be missing in several genomes within the group.

TaxonomyFinder implements taxonomy prediction on species and phylum level. The database includes 33 phylum-specific and 1242 species-specific profile sets. Additionally, TaxonomyFinder provides protein annotation for the submitted isolate based on functional profiles.

PanFunPro is available at https://cge.cbs.dtu.dk/services/PanFunPro.

TaxonomyFinder is available at https://cge.cbs.dtu.dk/services/TaxonomyFinder.

KmerFinder

In their groundbreaking paper from 1977, Woese and Fox uncovered Archea as a separate branch in the tree of life [28]. As a measure of genetic relatedness, they used the number of co-occurring kmers in 16S (18S) rRNA genes. Kmers are

stretches of DNA with the length of k nucleotides (Woese and Fox used the term oligonucleotides). Taking advantage of the availability of complete prokaryotic genomes, KmerFinder uses a similar approach for identifying the species, but extends the analysis to kmers across the entire genome. More specifically, KmerFinder relies on a database of reference genomes with annotated species [25] that are each split into overlapping 16-mers with step-size one. This means that if the first 16-mer of a reference genome is initiated at position N and ends at position N + 15, then the next 16-mer is initiated at position N + 1 and ends at position N + 16 etc. To reduce the size of the final 16-mer database, only 16-mers with the prefix ATGAC are kept. For the prediction of the species of a query genome, the genome is likewise split into overlapping 16-mers and the species is predicted to be identical to the species of the reference genome with which it has the highest number of 16-mers in common regardless of position.

VTEC Case Study: Identifying the Species Using KmerFinder

The suspected VTEC isolate (C757-12) [5] was run through the KmerFinder web-service using the scoring method "winner takes it all". KmerFinder offers two different scoring schemes: "standard" and "the winner takes it all". In the standard scoring scheme, all identical Kmers between the query sequence and the reference genomes are reported and statistics are calculated based on this. When choosing "the winner takes it all" scoring scheme, the output for the top hit (the reference genome in which the highest number of identical 16-mers with the query sequence was found) is the same as for the standard scoring scheme. But for the following significant hits, only 16-mers that were not found before are counted. This scoring scheme leads to the indication of whether or not (and to which extent) the query sequence is chimeric—with two or more origins.

Figure 5.3 shows the KmerFinder output page. The "Hit" is the genome in the reference database with which the query genome (the suspected VTEC isolate, C757-12) has most co-occurring 16-mers. Hence, according to KmerFinder, the predicted species is *E. coli*.

For another suspected VTEC isolate, C484-12 [5], KmerFinder found that the isolate was actually a *Morganella Morganii*. This was later confirmed and the isolate was excluded from the remainder of the study.

KmerFinder is available at https://cge.cbs.dtu.dk/services/KmerFinder.

Performance of Methods for Species Identification

The performances of the above-mentioned four methods for species identification have been evaluated in terms of accuracy and speed [25]. More than 11,000 isolates covering 159 genera and 243 species were used in the evaluation. Datasets of both

KmerFinder 2.0 results:

Hit	Score	z-score	Query Coverage [%]	Template Coverage [%]	Depth	Total Query Coverage [%]	Total Template Coverage [%]	Total Depth
Escherichia coli, Escherichia coli O103:H2, Escherichia coli O103:H2 str. 12009 get sequence	10207	410.3	18.08	93.19	18.62	18.08	93.19	18.62

Fig. 5.3 The KmerFinder output when the suspected VTEC isolate C757-12, with serotype O103:H2 [5], is used as input. The columns of the output table contain, among others, the name of the reference genome with which the query genome has the highest number co-occurring 16-mers and a link to the sequence of the genome (first column; Hit) The remaining columns contain statistics on the 16-mers used for the comparison. A full description of the content of all columns can be found here: http://cge.cbs.dtu.dk/services/KmerFinder/output.php

short sequence reads and assembled draft genomes were included. The results indicated that methods, which only sample chromosomal, core genes (e.g. SpeciesFinder and Reads2Type) had difficulties in distinguishing closely related species. As an example, SpeciesFinder had problems distinguishing *Yersinia pestis* from *Yersinia pseudotuberculosis* and *Mycobacterium tuberculosis* from *Mycobacterium bovis*. Overall, the performances of SpeciesFinder and Reads2Type were found to be similar, ranging from 76 to 87 % correct species identification, when tested on the three different evaluation sets. TaxonomyFinder, on the other hand, had a higher performance, ranging from 85 to 95 % correct species identification. KmerFinder had the highest performance with 93–97 % correct identifications. The species that all methods had problems distinguishing were typically within the *Bacillus* genus or *Escherichia coli*—*Shigella* spp. mix-ups. Rather than pointing to flaws in the methods, these misclassifications are likely to highlight sub-optimal conventional classification: Species belonging to the *Bacillus cereus* group are notoriously difficult to distinguish, as they are genetically very similar. It has hence been suggested that all members of the *B. cereus* group (including *B. cereus*, *Bacillus anthracis*, and *Bacillus thuringiensis*) should be considered to be *B. cereus* and only subsequently differentiated by their content of plasmids [29]. Likewise, although *Shigella* spp. has for many years been considered a sub-strain of *E. coli* and the separation is mainly historical [30, 31], the practical implications of renaming it are considered insurmountable.

The speed of the four methods for species identification was likewise tested on both assembled draft genomes and short sequence reads (see Table 5.2). Since the actual speed experienced by the user will depend on, for instance, the network bandwidth capacity of their computer and the number of jobs queued at the server, it is the relative speed of the different methods in comparison to each other that should be noted, not the absolute speed. KmerFinder was found to be the fastest method, while TaxonomyFinder was the slowest. In contrary to the other methods, TaxonomyFinder does not work on the nucleotide sequence directly, but rather on the proteome, utilizing functional protein domain profiles for the species prediction. Hence, in return for the extra time, the user is rewarded with an annotated genome.

Table 5.2 Speed of four methods for whole genome-based species identification

Method	Speed on draft genomes (mm:ss)	Speed on short reads (mm:ss)
SpeciesFinder	00:13	3:14
Reads2Type	NA[a]	1:20
TaxonomyFinder	11:33	NA[a]
KmerFinder	00:09	03:10

[a]Reads2Type only takes short sequence reads as input, while TaxonomyFinder only takes assembled draft genomes as input

Subtyping

Once the species of the organism of interest has been identified, the next step is typically to identify the strain. A number of methods have been proposed—and are in use—for the purpose of differentiating microorganisms beyond the level of species or subspecies. Some of these methods are phenotype-based, e.g., phage-typing and serotyping, while others are founded on the genomes of the organisms. In 1998 a scheme for subtyping on the basis of internal nucleotide sequences of a small number of housekeeping genes was proposed for *Neisseria meningitidis* [32]. Unique sequences (alleles) for each housekeeping gene (locus) are assigned a random integer number and a unique combination of alleles at each locus, an "allelic profile", defines the sequence type (ST). The procedure is called multilocus sequence typing (MLST) and has been adopted to close to 100 additional microorganisms besides *N. meningitidis*. These additional microorganisms are mainly bacteria, but MLST schemes for fungal species have also been developed. The MLST allele sequences and ST profile tables are stored in curated databases hosted at different sites around the world, and made collectively available via the pubMLST site (http://pubmlst.org). One of the great advantages of MLST is that it is standardized and the nucleotide sequence of a particular allele of a particular locus is unambiguous, thus requiring a minimum of subjective interpretation by the person carrying out the analysis. For a handful of species, e.g., *Escherichia coli* [33, 34] and *Clostridium difficile* [35, 36], different groups have developed different MLST schemes, each employing a slightly different set of loci. But besides these inexpedient causes of confusion, MLST and the sequence types provide clinical microbiologists, food safety authorities and everyone else working with microorganisms a standardized way of performing subtyping and naming the bacteria. The latter is also extremely useful for communicative purposes.

The MLST Web-Service

Due to the above-mentioned advantages of MLST, it was for several years considered the gold standard of typing, even though it was traditionally carried out in a time-consuming and expensive manner. With the advent of next generation sequencing technologies, other typing schemes that take a larger proportion of the genome

into account are expected to become predominant; e.g., SNP, wgMLST or, cgMSLT based. Nevertheless, the first web-service made available by CGE was one that could perform MLST on the basis of WGS data [4]. The purpose of developing this web-service was less to confirm the importance of MLST as a reference genome typing method, but rather to enable comparison of isolates based on WGS data with those analyzed earlier by more traditional methods. In other words, the purpose was to provide backwards compatibility, while in parallel using WGS-based MLST as a spearhead for implementation of routine-use of WGS.

For use by the CGE MLST web-service, all MLST databases are automatically downloaded from the pubMLST.org site once a week. As input, the MLST web-service can receive either short sequence reads or assembled draft genomes. In the case of short sequence reads, they are assembled to draft genomes before the analysis [4]. Applying the user-specified MLST scheme to the input data, the best-matching MLST alleles is then found using BLAST [37]. Finally, the sequence type is determined by the combination of identified alleles. Currently (Oct. 2014), 116 different schemes are available for bacterial and fungal species, and more are being added, as they are developed and incorporated in pubMLST.org.

VTEC Case Study: Identifying Sequence Type Using the MLST Web-Service

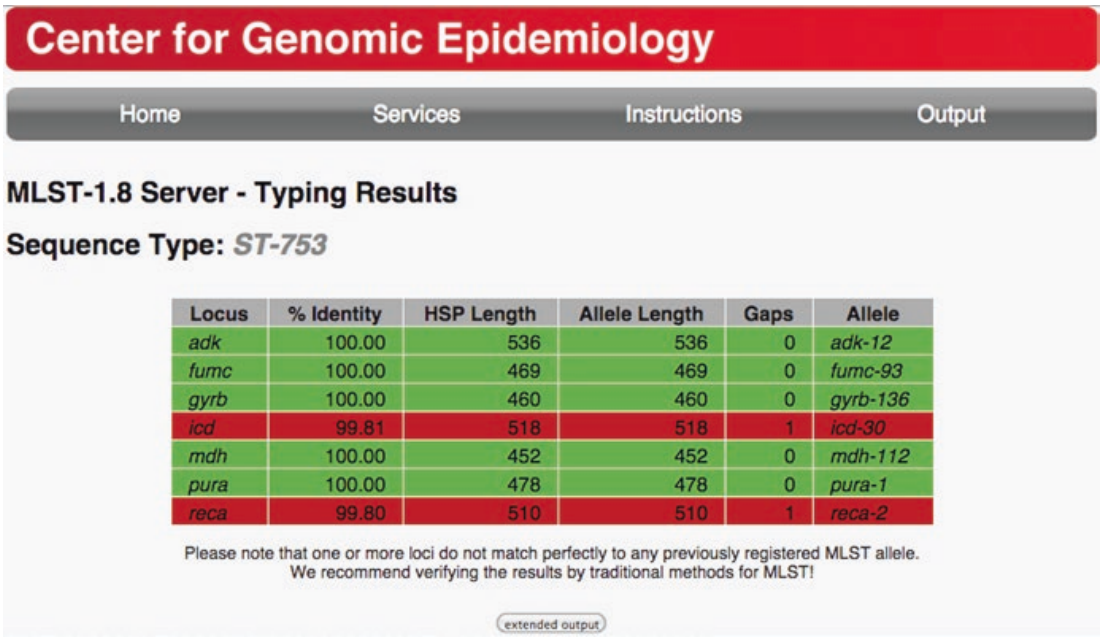
A suspected VTEC isolate (C770-12) [5] was run through the MLST web-service using the *E. coli* #1 MLST scheme. Figure 5.4 shows the output of the service. Below the listed predicted Sequence Type, a table shows the best-matching allele in the MLST database for each of the seven loci, along with information on the quality of the match.

By clicking the "extended output" button it is possible to examine the alignment between each of the MLST alleles and the corresponding sequences in the query sequence. Figure 5.5 shows the alignment of the *reca* locus, in which one gap occurs in the query sequence. All the suspected VTEC isolates were sequenced on an IonTorrent PGM sequencer, which is known to produce a large number of false-positive indels.

The remaining "Finder-tools" described below also all have the option to examine the actual alignments by selecting the extended output format.

*HSP: High-scoring Segment Pair. BLAST term that refers to the length of the alignment between the allele from the MLST database and the corresponding nucleotide sequence in the query genome.

The MLST web-service is currently (end-2015) the second-most used web-service provided by CGE (the most used web-service being ResFinder). From Sep. 2012 to Oct. 2015, the service was used more than 50,000 times in total. Examining the literature citing the CGE MLST web-service indicates that the service has most often been used for typing *E. coli*, *S. enterica*, and *S. aureus* (for examples see [3, 38–40]).



Locus	% Identity	HSP Length	Allele Length	Gaps	Allele
adk	100.00	536	536	0	*adk-12*
fumc	100.00	469	469	0	*fumc-93*
gyrb	100.00	460	460	0	*gyrb-136*
icd	99.81	518	518	1	*icd-30*
mdh	100.00	452	452	0	*mdh-112*
pura	100.00	478	478	0	*pura-1*
reca	99.80	510	510	1	*reca-2*

Fig. 5.4 Output from the MLST web-service when the suspected VTEC isolate, C770-12 [5], was run through the service using the *E. coli* #1 MLST scheme. Rows describing perfect matches between alleles in the database and the query sequence are colored *green*. In perfect matches the % identity is 100, the HSP* length equals the allele length, and there are no gaps. Rows describing imperfect matches are *red*

Fig. 5.5 Pairwise-alignment of the *reca-2* allele from the MLST database and the corresponding sequence in the C770-12 isolate

The MLST web-service is available at https://cge.cbs.dtu.dk/services/MLST.

Serotype

Serotyping has since its development in the 1940s become the gold standard for typing of several important pathogens, e.g. *E. coli* and *Salmonella*. Classical serotyping relies on serological detection of antigenic surface structures.

Serotyping of *E. coli*

For *E. coli* the most important antigens in typing are the somatic lipopolysaccharide O-antigen and the flagellar H-antigen. In order to transform this classical, phenotypic typing method into the WGS era, SerotypeFinder was constructed for WGS-based serotyping of *E. coli* [10]. The tool utilizes the O-antigen processing genes of *wzx*, *wxy*, *wzm*, and *wzt* to predict O-types and the flagellin genes *fliC*, *flkA*, *fllA*, *flmA*, and *flnA* for prediction of H-types. The SerotypeFinder database includes gene variants covering all 53 known H-types as well as all 188 known O-types, with the exception of O14 and O57.

The SerotypeFinder outputs O-types on basis of O-type specific gene variants, either by the combination of the variants detected by *wzx* and *wzy*, or by *wzm* and *wzt*. With a few exceptions, the two genes, e.g. *wzx/wzy*, output the same O-type, whereas in some cases the O-type will be predicted from just one of these gene variants. The H-type is predicted in SerotypeFinder by *fliC* gene variants alone, when this gene is the only flagellin gene present in the genome that is examined, whereas in cases of both a *fliC* and a non-*fliC* (*flkA*, *fllA*, *flmA*, *flnA*) gene being present, the non-*fliC* is set to predict the phenotype.

The SerotypeFinder is very robust and provides results directly comparable to the conventional serotyping of *E. coli*. In addition, it offers H-typing of all non-motile *E. coli* as well as O-typing of some rough strains that cannot be serotyped by conventional serotyping.

The SerotypeFinder is available at: https://cge.cbs.dtu.dk/services/SerotypeFinder.

Serotyping of *Salmonella*

Serotyping has likewise traditionally been a cornerstone in the surveillance of *Salmonella*. As for *E. coli*, the monitored antigens are the lipopolysaccharide O-antigen (encoded by the *rfb* gene cluster) and the flagellar H-antigen (encoded by the *fliC* and *fliB* genes). Researchers at the Center for Food Safety, University of

Georgia have developed a method, which is based on mapping of the raw reads from a sequencing run to curated databases of alleles of the *rfb* gene cluster and the *fliC* and *fliB* genes for WGS-based serotyping of *Salmonella* [41]. Although not having been involved in the development of the SeqSero web-service, CGE hosts the service, which is able to use raw sequence reads as well as assembled draft genomes as input.

The SeqSero web-service is available at: https://cge.cbs.dtu.dk/services/SeqSero and https://www.denglab.info/SeqSero.

Plasmids

Just as clones of bacteria might spread and are important to track, e.g., to find the source of an outbreak, so might plasmids spread horizontally among bacteria, conferring specific properties to their hosts. For the molecular epidemiological investigations of the major plasmid incompatibility groups among *Enterobacteriaceae*, plasmid-typing methods have been developed [42]. The initial method was developed to detect the replicons (part of the origin of replication and/or the replicase gene) of plasmids of the 18 major incompatibility (Inc) groups found in *Enterobacteriaceae*, but was extended and now contains 25 different replicons. Based on a database of 116 replicon sequences extracted from 559 plasmids, the PlasmidFinder method employs a BLAST-based search engine similar to the CGE MLST implementation for identification of plasmids [8]. For plasmid multilocus sequence typing (pMLST), a weekly updated database is furthermore generated from www.pubmlst.org/plasmid and integrated into a separate web-service named pMLST.

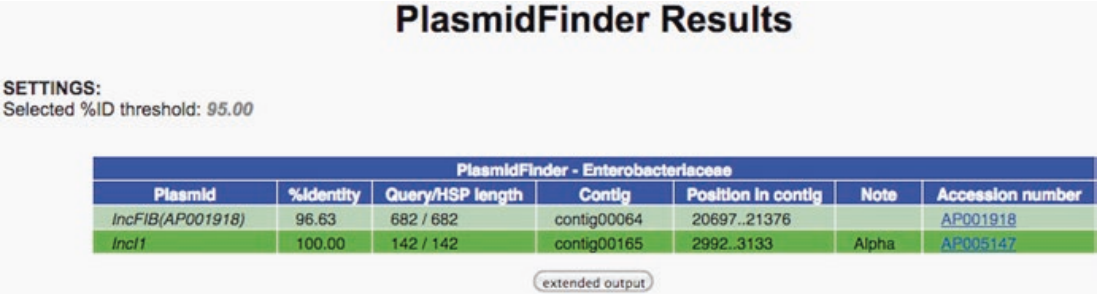
VTEC Case Study: Identifying Plasmids Using PlasmidFinder
Results of the PlasmidFinder web-service, when the isolate C757-12 [5] is used as input is shown in Fig. 5.6. Two plasmids (or actually replicons) were found: *I1* and *FIB(AP001918)*.

PlasmidFinder is available at https://cge.cbs.dtu.dk/services/PlasmidFinder.

The pMLST web-service is available at https://cge.cbs.dtu.dk/services/pMLST.

Phenotyping

Once the isolate is identified with adequate resolution, its potential phenotype can be investigated. For this purpose, we have developed a number of methods that typically search the input genome for the presence of particular genes that, if expressed at adequate levels, are likely to result in a particular phenotype of the isolate.


PlasmidFinder Results

SETTINGS:
Selected %ID threshold: 95.00

PlasmidFinder - Enterobacteriaceae

Plasmid	%Identity	Query/HSP length	Contig	Position in contig	Note	Accession number
IncFIB(AP001918)	96.63	682 / 682	contig00064	20697..21376		AP001918
IncI1	100.00	142 / 142	contig00165	2992..3133	Alpha	AP005147

extended output

Fig. 5.6 Results of the PlasmidFinder web-service. Note that it is the identified replicons (part of the origin of replication and/or the replicase gene) that are reported, not the entire plasmids

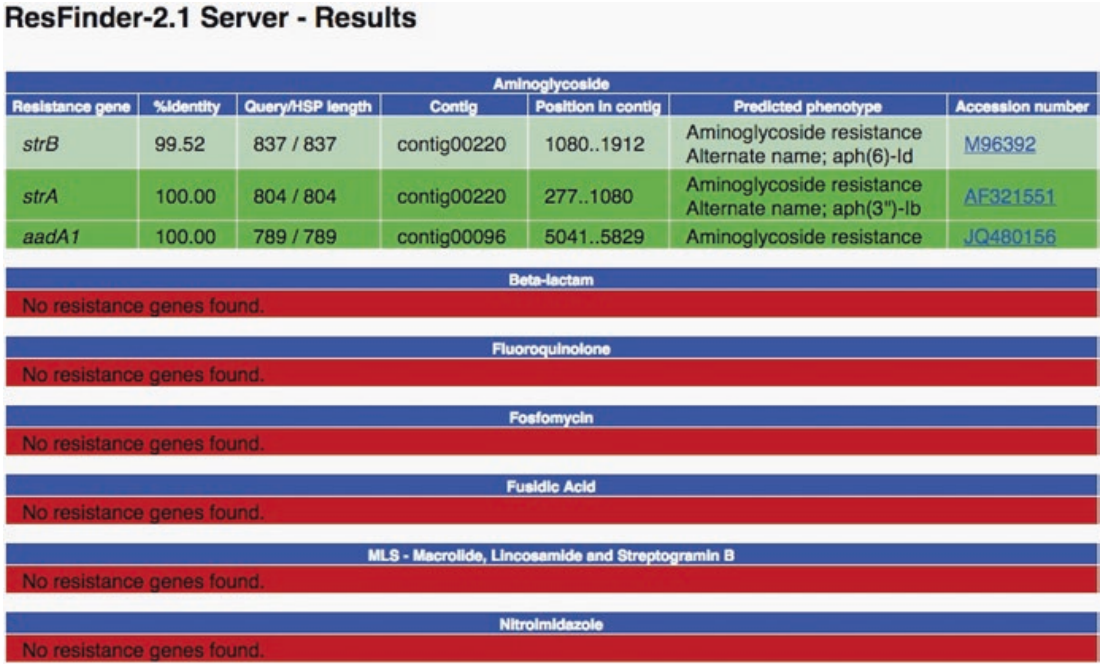
ResFinder

If the isolate is a pathogen, it is of importance to find out how it can be treated. To this end, a method for identification of acquired antimicrobial resistance genes was developed. The method is called ResFinder [9]. A major effort was initially put into compiling a curated database, based on public databases as well as on scientific papers. The database contains genes for the 13 major antimicrobial agent groups (Aminoglycosides, Beta-lactamases, Fluoroquinolone, Fosfomycin, Fusidic Acid, Glycopeptides, Macrolide-Lincosamide-StreptograminB, Nitroimidazole, Phenicol, Rifampicin, Sulfonamide, Tetracycline, and Trimethoprim) and is updated continuously. Query genomes are examined for the presence of any of these genes using the BLAST-based search engine.

VTEC Case Study: Identifying Acquired Antibiotic Resistance Genes with ResFinder

All the suspected VTEC isolates were examined for the presence of acquired antibiotic resistance genes. Overall, the isolates only contained very few of these genes. Figure 5.7 shows the results for C659-12 [5], which contained three genes known to confer resistance towards aminoglycosides.

Concerns have been raised that an assigned genotype may not always correspond to the actual phenotype, for instance due to mutations outside a particular gene, but affecting the expression of the gene product. A study was therefore conducted to compare antimicrobial resistance geno- and phenotypes. A surprisingly high concordance (99.74 %) was found between phenotypic and predicted antimicrobial susceptibility. Although the results were promising, it should be noted that the study was conducted in a population with relatively low levels of resistance, and lower levels of concordance may be found in other populations. It should likewise be noted that ResFinder is only able to discover antimicrobial resistance due to acquired antimicrobial resistance genes, not, e.g., point mutations in chromosomal genes. Nevertheless, it was concluded that genotyping using whole-genome sequencings is



ResFinder-2.1 Server - Results

Aminoglycoside

Resistance gene	%Identity	Query/HSP length	Contig	Position in contig	Predicted phenotype	Accession number
strB	99.52	837 / 837	contig00220	1080..1912	Aminoglycoside resistance Alternate name; aph(6)-Id	M96392
strA	100.00	804 / 804	contig00220	277..1080	Aminoglycoside resistance Alternate name; aph(3")-Ib	AF321551
aadA1	100.00	789 / 789	contig00096	5041..5829	Aminoglycoside resistance	JQ480156

Beta-lactam

No resistance genes found.

Fluoroquinolone

No resistance genes found.

Fosfomycin

No resistance genes found.

Fusidic Acid

No resistance genes found.

MLS - Macrolide, Lincosamide and Streptogramin B

No resistance genes found.

Nitroimidazole

No resistance genes found.

Fig. 5.7 An extract of the result from the ResFinder service when the C659-12 isolate is used as input. Besides the name of the resistance genes found (Resistance gene), the %identity between the gene in the ResFinder database and the corresponding sequence in the input isolate is shown (%identity). Also shown is the Query/HSP Length, where Query length is the length of the best matching resistance gene in the database, while HSP length is the length of the alignment between the best matching resistance gene and the corresponding sequence in the genome (also called the high-scoring segment pair (HSP)). The name of the contig and the position in the contig is also shown. Finally, the predicted phenotype based on the identified resistance gene is shown, as is the reference Genbank accession number according to NCBI of the resistance gene in the database

a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing [43].

ResFinder is currently the most frequently used web-services provided by CGE having been used almost 60,000 times by Oct. 2015, since its publication in end-2012. A review of the literature citing ResFinder shows that the method has so far mainly been used for identifying antimicrobial resistance genes in gram-negative bacteria, e.g., *K. pneumoniae* (for instance [44]), *E. coli* (see for instance [45]), and *S. enterica* (see for instance [46]).

ResFinder is available at https://cge.cbs.dtu.dk/services/ResFinder.

MyDbFinder

To accommodate that researchers may have interests in particular sets of genes, for which there is no in-house CGE databases, a special version of ResFinder, called MyDbFinder, was developed. Using this service, the user can generate their own database containing genes of interest, for which the program should search. The database must contain the DNA sequences of the genes that the user wishes to

```
>Seq1
ACTCGCGATCCGCATAGCGCATCGCATG
>Seq2 optional comment
ATGAAAACAATGATTTATCCTCACCAATATAATTATATCAGATCGGTTATT
TATGCGGCAATGATTTATCCTCACCAATGATGAGAGAGCAGATACTCTTTG
AACAAAGAAATTGAAGCAATACTTAATAAATTT
```

Fig. 5.8 Two DNA sequences in FASTA format

search for. As ResFinder, MyDbFinder uses BLAST to identify the genes in the query WGS data and outputs the best matching genes from the user's database. It is possible to select different settings depending on how strict an output is wanted.

How to Make a Database for MyDbFinder

The database should be made in a text editor (Notepad, TextEdit or equivalent) and must consist of DNA sequences in FASTA format. A sequence in FASTA format begins with a header, which is a single line description, followed by lines of sequence data in single-letter nucleotide code (A, T, C, and G). The header line always starts with a ">" (greater than) symbol, which distinguishes this line from the lines containing the sequence data. Note that empty lines are not accepted in FASTA files.

Figure 5.8 exemplifies two DNA sequences in FASTA format:

When making a database the user should be aware that MyDbFinder only shows the first word of the header (the characters until the first space) for outputted genes, thus different genes/sequence names without spaces are recommended.

MyDbFinder is available at https://cge.cbs.dtu.dk/services/MyDbFinder.

VirulenceFinder

The same BLAST-based methodology as above is also used for prediction of virulence factors in verotoxigenic *E. coli* on the basis of a database of known *E. coli* virulence genes [5]. In the study for which this VirulenceFinder tool was developed, it was used to examine 46 suspected VTEC isolates (the same isolates that are used throughout this chapter to exemplify the output of CGE web-services). VirulenceFinder quickly and accurately detected *eae*, *ehxA*, and *vtx* genes and was in addition more robust in assigning correct *vtx* subtypes than routine typing. Although poor sequencing quality and overall low coverage for some isolates caused VirulenceFinder to miss the detection of a few genes, it overall detected

the presence of many other important virulence genes, thus giving much more information on the virulence profiles of the isolates than was obtained by routine typing [5].

VTEC Case Study: Identifying Virulence Factors Using VirulenceFinder

The suspected VTEC isolates were examined for the presence of known virulence genes using VirulenceFinder. Figure 5.9 shows the results when using the isolate C892-12 [5] as input. Routine typing had assigned this isolate as *vtx2d*, while the subtype found by VirulenceFinder was *vtx2g* (reported in the column "protein function"). Retyping at SSI confirmed the subtype detected by VirulenceFinder. In a few other instances, where results obtained by routine typing and VirulenceFinder were not in agreement, poor WGS data quality was usually the cause.

VirulenceFinder is available at https://cge.cbs.dtu.dk/services/VirulenceFinder.

PathogenFinder

Whereas VirulenceFinder looks for the presence of known virulence factors previously described in the literature, Andreatta et al. took a radically different approach for determining the pathogenic potential of an organism [47]. In the study from 2010,

VirulenceFinder-1.5 Server - Results

SETTINGS:
Selected %ID threshold: *85.00*

Virulence - E. coli						
Virulence factor	%Identity	Query/HSP length	Contig	Position in contig	Protein function	Accession number
stx2A	100.00	960 / 960	contig00053	1039..1998	Shiga toxin 2, subunit A, variant g	GQ995452
lpfA	100.00	573 / 573	contig00003	165975..166547	Long polar fimbriae	AP010953
gad	99.83	1401 / 1151	contig00118	1..1150	Glutamate decarboxylase	AP009240
celb	100.00	144 / 144	contig00098	225..368	Endonuclease colicin E2	AF540491
katP	96.34	2211 / 2211	contig00035	33782..35992	Plasmid-encoded catalase peroxidase	AB011549
stx2B	100.00	270 / 270	contig00053	757..1026	Shiga toxin 2, subunit B, variant g	GQ995452

stx holotoxins						
Virulence factor	%Identity	Query/HSP length	Contig	Position in contig	Protein function	Accession number
stx2	99.84	1242 / 1242	contig00053	757..1998	Out S-8, variant g	AB048227

Fig. 5.9 Output of the VirulenceFinder service, when the C892-12 isolate was used as input. The columns correspond to those provided by the ResFinder tool, except for the "Protein function" column

genomes from *gamma-proteobacteria* were first grouped into those originating from pathogenic vs. non-pathogenic bacteria. Next, the genomes were examined for the presence of gene families that were statistically associated with being found in either the pathogenic or non-pathogenic groups. To the best of our knowledge, this is the first example of the use of machine learning techniques for determining the phenotype based on whole genome sequences. The method has later been extended to be applicable for all species of bacteria and made publicly available as the PathogenFinder method [7]. Since the method relies on groupings of proteins, without considering their annotated function (or even if they have any) or known involvement in pathogenicity, it can also aid the discovery of novel pathogenicity factors.

PathogenFinder is available at https://cge.cbs.dtu.dk/services/PathogenFinder.

Phylogeny

Nucleotide sequences have long been used to classify the species and taxonomy of bacteria and other organisms. But until recently it was only a few genes, including the 16s rRNA gene, that were used for making phylogenies. However, since the price of whole genome sequencing has gone down, whole genome based phylogeny has become increasingly used for both typing of bacteria and for disease outbreak detection. Previously phylogeny was mostly used to divide samples into families and species. But if the method is exact enough, it can even be used to follow and detect disease outbreaks. If, for example, WGS data from a number of samples from different patients are available, and the strains only differ by a few nucleotides, the strains are likely to have originated from the same strain, and hereby it can most likely be concluded that all the patients were contaminated from the same source. Another option that becomes possible when WGS data is available is to upload the data to a large database with good annotations and metadata, and make a phylogenetic tree of all similar strains. In this way, it might be possible to identify the possible contamination source.

SNPtree, CSIPhylogeny, and NDtree

At CGE, three tools based on Single Nucleotide Polymorphisms (SNPs) are available, which accept both short sequence reads as well as assembled genomes as input. The SNP tools are useful for identifying SNPs in closely related strains. The first developed tool was snpTree [11]. This tool first maps the query genomes to a reference genome selected by the user; either by MUMMER [48] (assembled sequences) or BWA [49] (short sequence reads). The reference genome can either be selected from the CGE database or uploaded by the user. After the mapping, the program localizes SNPs in the genomes by use of SAMtools [50] and nucmer [48]. Following the localization of all SNPs, they are filtered based on user-specified

settings. Default settings can also be used, which include a sequencing depth of ten and a minimum distance of ten bases between the SNPs. The SNPs that pass the filtering criteria for each genome are then concatenated to a continuous sequence, and a phylogenetic tree is made based on this multiple alignment. The output files include the alignment in different formats, the tree file, a file showing the individual SNPs in the genomes, VCF[1] files for each genome and a matrix with the difference between the genomes [11].

snpTree is widely used for genome analysis of different species. In 2014 Guio et al. [51] used snpTree to find 218 non-synonymous SNPs in the genome of *Mycobacterium tuberculosis* that could confer resistance towards antibiotics. The service has also been used by Teo et al. [52] to characterize an emerging new pathogen in the hospital environment, *Elizabethkingia anopheles*.

CSIPhylogeny is a further development of snpTree. CSIPhylogeny identifies the SNPs in the same way as snpTree, but is more strict when it comes to filtering (removing) the SNPs. SNPs are filtered if their mapping quality (which is calculated using BWA [49]) is below certain thresholds: If the SNP quality is below 30, or if the sequencing depth is below 10. SNPs are also removed in the filtering step if they are less than ten base pairs from the nearest SNP. Finally a Z-score is calculated for each SNP, which has to be above 1.96, for the SNP to be kept. The Z-score is calculated using the following equation:

$$Z = \left(X - Y\right) / \mathrm{sqrt}\left(X + Y\right) \tag{5.1}$$

Here X is the number of reads, having the most common nucleotide at that position, and Y the number of reads with any other nucleotide [2, 5]. CSIPhylogeny has the same output files as snpTree.

The main difference between snpTree and CSIPhylogeny lies in the site validation performed by CSIPhylogeny. The validation consists of checking all positions in the analysis and only using those that are considered valid in all the isolates analyzed. The snpTree method performs no validation and assumes all non-SNP positions to be valid, i.e., positions where no SNPs are found or where SNPs has been ignored are assumed to be identical to the base in the reference sequence. This assumption is inconsequently if the isolates compared (including the reference strain) are very closely related. However, the less related the compared isolates are the bigger the consequences will be, because more and more non-SNP positions will not be valid either due to low quality or simply because the DNA in which a SNP is found does not exist in all the genomes of the isolates.

The snpTree server is now deprecated and will no longer be updated. It will remain online, but the CSIPhylogeny server is the recommended tool.

Our third server for constructing phylogenetic trees is NDtree, which constructs the trees based on the number of nucleotide difference found between genomes [5]. It is intentionally made as simple as possible to study which features are important for making accurate phylogenies. NDtree can find a reference genome automati-

[1] VCF (Variant Call Format) files specify a type of text files used for storing sequence variation.

cally using KmerFinder, or may be given a reference sequence by the user. NDtree then maps the query reads to the reference sequence. This is done by splitting the reference sequence and the reads into 17-mers and storing them in a hash table. The 17-mers from the reads are then mapped to the reference to find a match, which is then extended to an optimally scoring ungapped alignment using a match score of 1 and a mismatch score of −3. If the match score is greater than 50 the alignment is used in the SNP calling. A position is significant if the Z-score, as described above in Eq. (5.1), is more than 1.96 and X is ten times larger than Y. If no base can be called based on the above criteria, an "N" is put in the sequence and the position is not used for phylogeny. These called sequences are then compared pairwise, counting the nucleotides that differ. An option can be chosen so all positions called in a given pair of sequences are used, even if that position is not called in one or more of the other sequences. In this case it is advised to increase the Z score cutoff to 3.29. After the pairwise comparisons, the tree is build from the distance matrix, using the UPGMA or neighbor joining (NJ) packages from Phylip. It is recommended to use the UPGMA if the samples are taken at the same time, and otherwise NJ. NDtree output files include the tree file in newick format and the distance matrix.

CSIPhylogeny and NDtree have recently been shown to be more accurate than the older SNPtree method and should be the preferred tools for phylogeny [2].

VTEC Case Study: Determining Phylogeny Using NDtree
The phylogeny of the suspected VTEC isolates was examined using NDtree (see Fig. 5.10). The isolates clustered completely according to serotype, and a clear concordance between serotype and MLST type could be observed. Further, there was a complete agreement with the epidemiological information and the observed clustering [5].

snpTree is available at https://cge.cbs.dtu.dk/services/snpTree.

CSIPhylogeny is available at https://cge.cbs.dtu.dk/services/CSIPhylogeny.

NDtree is available at https://cge.cbs.dtu.dk/services/NDtree.

Metagenomic Samples

Much attention has recently been given to the possibility of diagnosing diseases based on metagenomic samples, since this is faster and simpler than having to initially isolate the bacteria. Hasman et al. [3] were to the best of our knowledge the first to show that metagenomic samples (in this case urine) could be used to diagnose a pathogen without prior knowledge about which species it was. It was found that WGS improved the identification of the cultivated bacteria, and an

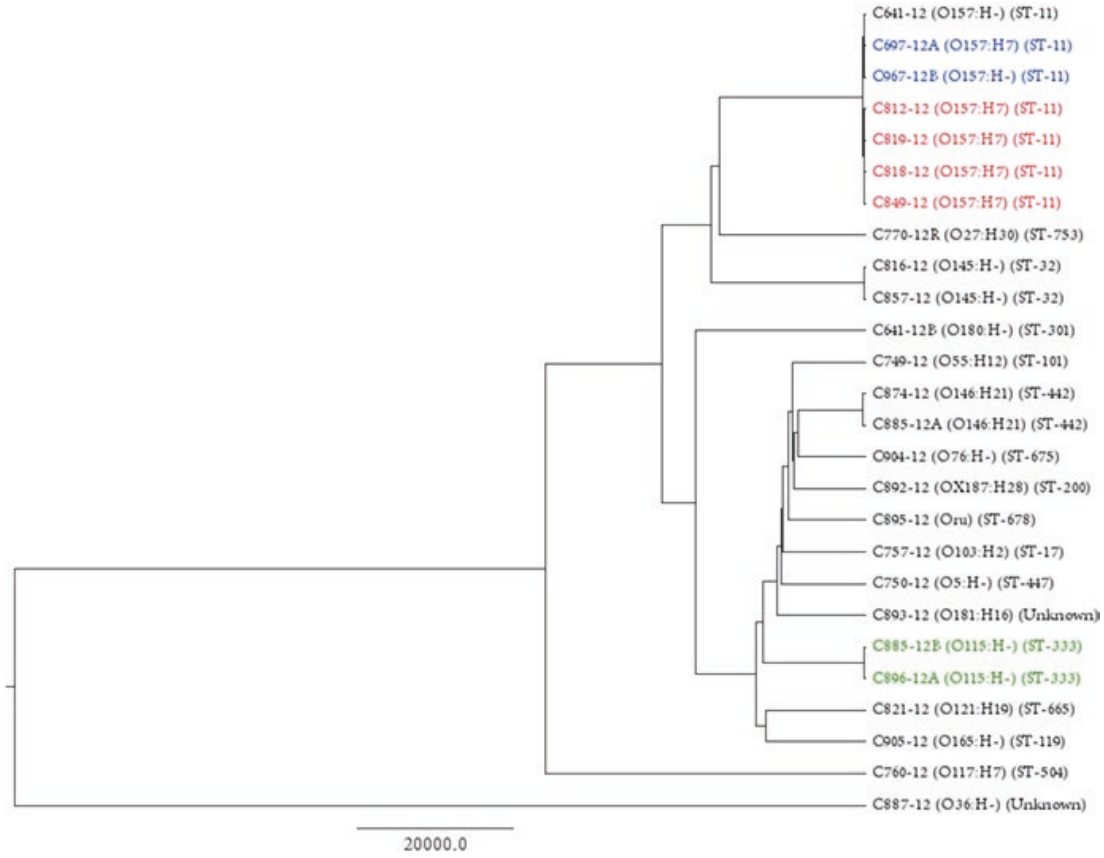


Fig. 5.10 Phylogeny of a subset of the suspected VTEC isolates according to the NDtree method. The tree has been constructed on basis of genome assemblies. Isolates known to be epidemiologically related are shown in the same color, with the red group constituting known outbreak isolates [53]. Serotypes and MLST types are shown for all isolates

almost complete agreement between phenotypic and predicted antimicrobial susceptibilities was observed [3]. Metagenomic analysis could also be used for monitoring purposes and was recently applied to the analysis of toilet waste from 18 international airplanes arriving in Copenhagen, Denmark. It was found that genes encoding antimicrobial resistance were more abundant and also of higher diversity in the planes from South Asia compared to North America. Likewise, the waste from the planes from South Asia indicated a higher presence of *S. enterica* and norovirus. Conversely, the waste from North America contained more *Clostridium difficile* [54].

Work in Progress

Besides finalizing a web-service, which will enable analysis of data from metagenomic samples, several other web-services are in development and expected to be published shortly. These include a method for identification of genes related

to restriction-modification systems, pathogenicity islands in S. enterica, and the prediction of the bacterial host of bacteriophages based on the genome sequence of the bacteriophage. To supplement the many stand-alone web-services, we are currently working on a number of improvements that will make the experience even more user-friendly. Specifically, we will include the possibility of batch-uploading several isolates in one go, followed by the automatic execution of several of the analytic tools for typing and phenotyping, and finally a graphical visualization of all isolates on a world map. Furthermore, we will add the option for registered users to manage their sequence data files and keep record of their results. These features are in dire need and their implementation will hopefully make the use of WGS for the analysis of pathogenic microorganisms even faster and more convenient.

Conclusion

Since the start of CGE, a large number of methods have been developed for the purpose of determining the species and subtypes of microorganisms, predict their phenotype, and investigate their phylogeny for epidemiological purposes. The methods have been made publicly available via web-services that are designed to be easy to use/"plug and play"—also for non-bioinformaticians. We will continue to update our existing methods as well as implement new ones, hopefully facilitating that the community can take full advantage of the genomics era.

Conclusion: VTEC Case Study
As a proof-of-concept that WGS-based typing of VTEC could be an attractive alternative in surveillance, real-time WGS-based typing of VTEC was performed during 7 weeks, in parallel to the routine typing carried out at SSI. The study included a set of 46 suspected VTEC isolates that were analyzed using the CGE tools. Overall, the results were concurrent with the routine typing, and the phylogenetic relationship determined was in agreement with epidemiological data. Furthermore, a small ongoing VTEC outbreak [53] was also easily distinguished by the WGS approach. We conclude that WGS-based typing of VTEC is an advantageous alternative to the current routine typing, producing comparable typing results faster and at a similar cost. For complete WGS-based VTEC surveillance WGS O:H serotyping is also needed, which we currently developing.

Acknowledgements The Center for Genomic Epidemiology at the Technical University of Denmark is funded by grant 09-067103/DSF from the Danish Council for Strategic Research.

References

1. Aarestrup FM, Brown EW, Detter C, Gerner-Smidt P, Gilmour D, Harmsen MW, Hendriksen RS, Hewson R, Heymann DL, Johansson K, Ijaz K, Keim PS, Koopmans M, Kroneman A, Lo Fo Wong D, Lund O, Palm D, Sawanpanyalert P, Sobel J, Schlundt J. Integrating genome-based informatics to modernize global disease monitoring, information sharing, and response. Emerg Infect Dis. 2012;18, e1.
2. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One. 2014;9, e104984.
3. Hasman H, Saputra D, Sicheritz-Ponten T, Lund O, Svendsen CA, Frimodt-Moller N, Aarestrup FM. Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol. 2014;52:139–46.
4. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 2012;50:1355–61.
5. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol. 2014;52:1501–10.
6. Lukjancenko O, Thomsen MCF, Larsen MV, Ussery DW. PanFunPro: PAN-genome analysis based on FUNctional PROfiles [v1; ref status: approved with reservations 3]. F1000Research. 2013;2:265. http://f1000r.es/2e1.
7. Cosentino S, Voldby Larsen M, Moller Aarestrup F, Lund O. PathogenFinder—distinguishing friend from foe using bacterial whole genome sequence data. PLoS One. 2013;8, e77302.
8. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, Moller Aarestrup F, Hasman H. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58:3895–903.
9. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67:2640–4.
10. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy in silico serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. J Clin Microbiol. 2015;53:2410–26.
11. Leekitcharoenphon P, Mortensen RK, Thomsen MCF, Friis C, Rasmussen S, Aarestrup FM. snpTree—a web-server to identify and construct SNP trees from whole genome sequence data. BMC Genomics. 2012;13 Suppl 7:S6.
12. Larsen J, Enright MC, Godoy D, Spratt BG, Larsen AR, Skov RL. Multilocus sequence typing scheme for *Staphylococcus aureus*: revision of the gmk locus. J Clin Microbiol. 2012;50:2538–9.
13. Boerlin P, McEwen SA, Boerlin-Petzold F, Wilson JB, Johnson RP, Gyles CL. Associations between virulence factors of Shiga toxin-producing *Escherichia coli* and disease in humans. J Clin Microbiol. 1999;37:497–503.
14. Karmali MA. Infection by verocytotoxin-producing *Escherichia coli*. Clin Microbiol Rev. 1989;2:15–38.
15. Karch H, Tarr PI, Bielaszewska M. Enterohaemorrhagic *Escherichia coli* in human medicine. Int J Med Microbiol. 2005;295:405–18.
16. CDC. Centers for Disease Control and Prevention—*E. coli*. General Information. 2012. http://www.cdc.gov/ecoli/general/#complications.
17. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 2009;106:19126–31.
18. Fox GE, Peckman KJ, Woese CE. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular approach to procaryotic systematics. Int J Syst Bacteriol. 1977;27:44–57.
19. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72:5069–72.

20. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 2007;35:7188–96.
21. Kampfer P. Systematics of prokaryotes: the state of the art. Antonie Van Leeuwenhoek. 2012;101:3–11.
22. Tindall BJ, Rossello-Mora R, Busse HJ, Ludwig W, Kampfer P. Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol. 2010;60:249–66.
23. Tindall BJ, Schneider S, Lapidus A, Copeland A, Glavina Del Rio T, Nolan M, Lucas S, Chen F, Tice H, Cheng JF, Saunders E, Bruce D, Goodwin L, Pitluck S, Mikhailova N, Pati A, Ivanova N, Mavrommatis K, Chen A, Palaniappan K, Chain P, Land M, Hauser L, Chang YJ, Jeffries CD, Brettin T, Han C, Rohde M, Goker M, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Klenk HP, Kyrpides NC, Detter JC. Complete genome sequence of *Halomicrobium mukohataei* type strain (arg-2). Stand Genomic Sci. 2009;1:270–7.
24. Klenk HP, Goker M. En route to a genome-based classification of Archaea and Bacteria? Syst Appl Microbiol. 2010;33:175–82.
25. Larsen MV, Cosentino S, Lukjancenko O, Saputra D, Rasmussen S, Hasman H, Sicheritz-Ponten T, Aarestrup FM, Ussery DW, Lund O. Benchmarking of methods for genomic taxonomy. J Clin Microbiol. 2014;52(5):1529–39.
26. Lukjancenko O, Ussery DW. Vibrio chromosome-specific families. Front Microbiol. 2014;5:73.
27. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R. A new bioinformatics analysis tools framework at EMBL-EBI. Nucleic Acids Res. 2010;38:W695–9.
28. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A. 1977;74:5088–90.
29. Helgason E, Okstad OA, Caugant DA, Johansen HA, Fouet A, Mock M, Hegna I, Kolsto AB. *Bacillus anthracis*, *Bacillus cereus*, and *Bacillus thuringiensis*—one species on the basis of genetic evidence. Appl Environ Microbiol. 2000;66:2627–30.
30. Karaolis DK, Lan R, Reeves PR. Sequence variation in *Shigella sonnei* (Sonnei), a pathogenic clone of *Escherichia coli*, over four continents and 41 years. J Clin Microbiol. 1994;32:796–802.
31. Lan R, Reeves PR. *Escherichia coli* in disguise: molecular origins of *Shigella*. Microbes Infect. 2002;4:1125–32.
32. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A. 1998;95:3140–5.
33. Jaureguy F, Landraud L, Passet V, Diancourt L, Frapy E, Guigon G, Carbonnelle E, Lortholary O, Clermont O, Denamur E, Picard B, Nassif X, Brisse S. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. BMC Genomics. 2008;9:560.
34. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 2006;60:1136–51.
35. Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, Fung R, Golubchik T, Harding RM, Jeffery KJ, Jolley KA, Kirton R, Peto TE, Rees G, Stoesser N, Vaughan A, Walker AS, Young BC, Wilcox M, Dingle KE. Multilocus sequence typing of *Clostridium difficile*. J Clin Microbiol. 2010;48:770–8.
36. Lemee L, Dhalluin A, Pestel-Caron M, Lemeland JF, Pons JL. Multilocus sequence typing analysis of human and animal *Clostridium difficile* isolates of various toxigenic types. J Clin Microbiol. 2004;42:2609–17.
37. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
38. Hendriksen RS, Joensen KG, Lukwesa-Musyani C, Kalondaa A, Leekitcharoenphon P, Nakazwe R, Aarestrup FM, Hasman H, Mwansa JC. Extremely drug-resistant *Salmonella enterica* serovar Senftenberg infections in patients in Zambia. J Clin Microbiol. 2013;51:284–6.

39. Rodriguez-Rivera LD, Moreno Switt AI, Degoricija L, Fang R, Cummings CA, Furtado MR, Wiedmann M, den Bakker HC. Genomic characterization of *Salmonella* Cerro ST367, an emerging *Salmonella* subtype in cattle in the United States. BMC Genomics. 2014;15:427.
40. Stegger M, Aziz M, Chroboczek T, Price LB, Ronco T, Kiil K, Skov RL, Laurent F, Andersen PS. Genome analysis of *Staphylococcus aureus* ST291, a double locus variant of ST398, reveals a distinct genetic lineage. PLoS One. 2013;8, e63008.
41. Zhang S, Yin Y, Jones MB, Zhang Z, Deatherage Kaiser BL, Dinsmore BA, Fitzgerald C, Fields PI, Deng X. Salmonella serotype determination utilizing high-throughput genome sequencing data. J Clin Microbiol. 2015;53:1685–92.
42. Carattoli A, Bertini A, Villa L, Falbo V, Hopkins KL, Threlfall EJ. Identification of plasmids by PCR-based replicon typing. J Microbiol Methods. 2005;63:219–28.
43. Zankari E, Hasman H, Kaas RS, Seyfarth AM, Agerso Y, Lund O, Larsen MV, Aarestrup FM. Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing. J Antimicrob Chemother. 2013;68:771–7.
44. Villa L, Feudi C, Fortini D, Garcia-Fernandez A, Carattoli A. Genomics of KPC-producing *Klebsiella pneumoniae* sequence type 512 clone highlights the role of RamR and ribosomal S10 protein mutations in conferring tigecycline resistance. Antimicrob Agents Chemother. 2014;58:1707–12.
45. Leonard SR, Lacher DW, Elkins CA, Jung CM. Draft genome sequence of the multidrug-resistant *Escherichia coli* strain LR09, isolated from a wastewater treatment plant. Genome Announc. 2014;2, e00272-14.
46. Kroft BS, Brown EW, Meng J, Gonzalez-Escalona N. Draft genome sequences of two Salmonella strains from the SARA collection, SARA64 (Muenchen) and SARA33 (Heidelberg). Provide insight into their antibiotic resistance. Genome Announc. 2013;1, e00806-13.
47. Andreatta M, Nielsen M, Moller Aarestrup F, Lund O. In silico prediction of human pathogenicity in the gamma-proteobacteria. PLoS One. 2010;5, e13680.
48. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30:2478–83.
49. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
50. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAM tools. Bioinformatics. 2009;25:2078–9.
51. Guio H, Tarazona D, Galarza M, Borda V, Curitomay R. Genome analysis of 17 extensively drug-resistant strains reveals new potential mutations for resistance. Genome Announc. 2014;2, e00759-14.
52. Teo J, Tan SY, Liu Y, Tay M, Ding Y, Li Y, Kjelleberg S, Givskov M, Lin RT, Yang L. Comparative genomic analysis of malaria mosquito vector-associated novel pathogen *Elizabethkingia anophelis*. Genome Biol Evol. 2014;6:1158–65.
53. Soborg B, Lassen SG, Muller L, Jensen T, Ethelberg S, Molbak K, Scheutz F. A verocytotoxin-producing *E. coli* outbreak with a surprisingly high risk of haemolytic uraemic syndrome, Denmark, September–October 2012. Euro Surveill. 2013;18:1–3.
54. Nordahl Petersen T, Rasmussen S, Hasman H, Caroe C, Baelum J, Schultz AC, Bergmark L, Svendsen CA, Lund O, Sicheritz-Ponten T, Aarestrup FM. 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.