frontiers in Cellular and Infection Microbiology

ORIGINAL RESEARCH
published: 10 January 2022
doi: 10.3389/fcimb.2021.772574

Improved Genomic Identification, Clustering, and Serotyping of Shiga Toxin-Producing *Escherichia coli* Using Cluster/Serotype-Specific Gene Markers

Xiaomei Zhang, Michael Payne, Sandeep Kaur and Ruiting Lan *

School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW, Australia

OPEN ACCESS

Edited by:
Floriana Campanile, University of Catania, Italy

Reviewed by:
Roberto Mauricio Vidal, University of Chile, Chile
Karen Keddy, South African Medical Research Council, South Africa
Joseph M. Bosilevac, U.S. Meat Animal Research Center, Agricultural Research Service (USDA), United States

***Correspondence:**
Ruiting Lan
r.lan@unsw.edu.au

Specialty section:
This article was submitted to Clinical Microbiology, a section of the journal Frontiers in Cellular and Infection Microbiology

Received: *08 September 2021*
Accepted: *03 December 2021*
Published: *10 January 2022*

Citation:
Zhang X, Payne M, Kaur S and Lan R (2022) Improved Genomic Identification, Clustering, and Serotyping of Shiga Toxin-Producing Escherichia coli Using Cluster/Serotype-Specific Gene Markers. Front. Cell. Infect. Microbiol. 11:772574. doi: 10.3389/fcimb.2021.772574

Shiga toxin-producing *Escherichia coli* (STEC) have more than 470 serotypes. The well-known STEC O157:H7 serotype is a leading cause of STEC infections in humans. However, the incidence of non-O157:H7 STEC serotypes associated with foodborne outbreaks and human infections has increased in recent years. Current detection and serotyping assays are focusing on O157 and top six ("Big six") non-O157 STEC serogroups. In this study, we performed phylogenetic analysis of nearly 41,000 publicly available STEC genomes representing 460 different STEC serotypes and identified 19 major and 229 minor STEC clusters. STEC cluster-specific gene markers were then identified through comparative genomic analysis. We further identified serotype-specific gene markers for the top 10 most frequent non-O157:H7 STEC serotypes. The cluster or serotype specific gene markers had 99.54% accuracy and more than 97.25% specificity when tested using 38,534 STEC and 14,216 non-STEC *E. coli* genomes, respectively. In addition, we developed a freely available *in silico* serotyping pipeline named STECFinder that combined these robust gene markers with established *E. coli* serotype specific O and H antigen genes and *stx* genes for accurate identification, cluster determination and serotyping of STEC. STECFinder can assign 99.85% and 99.83% of 38,534 STEC isolates to STEC clusters using assembled genomes and Illumina reads respectively and can simultaneously predict *stx* subtypes and STEC serotypes. Using shotgun metagenomic sequencing reads of STEC spiked food samples from a published study, we demonstrated that STECFinder can detect the spiked STEC serotypes, accurately. The cluster/serotype-specific gene markers could also be adapted for culture independent typing, facilitating rapid STEC typing. STECFinder is available as an installable package (https://github.com/LanLab/STECFinder) and will be useful for *in silico* STEC cluster identification and serotyping using genome data.

Keywords: STEC O157:H7, non-O157:H7 STEC serotypes, STEC phylogenetic clusters, cluster/serotype-specific gene markers, STEC serotyping, *in silico* STEC tying pipeline STECFinder, metagenomics

INTRODUCTION

Shiga toxin-producing *Escherichia coli* (STEC) are an important cause of foodborne disease worldwide (Tuttle et al., 1999; Teunis et al., 2008; World Health Organization, 2019). STEC causes human infections ranging from mild non-bloody diarrhea to haemorrhagic colitis (HC), haemolytic uraemic syndrome (HUS), end-stage renal disease (ESRD) and death (Paton and Paton, 1998; Tarr et al., 2005; Gould et al., 2009). Globally, an estimated 2.8 million STEC infections resulted in 3,890 cases of HUS, 270 cases of ESRD and 230 deaths in 2010 (Majowicz et al., 2014). Importantly, STEC infections were more frequent and severe in children younger than five years old (Gould et al., 2009; Buvens et al., 2012; Lozer et al., 2013).

Currently, there are over 470 STEC serotypes recognized based on *E. coli* O antigen (determination of O serogroup) and H (flagellar) antigen typing (Gyles, 2007; Mora et al., 2011; Ludwig et al., 2020). More than 130 STEC serotypes are associated with human STEC infections (Johnson et al., 1996; Bettelheim, 2000; Johnson et al., 2006; Valilis et al., 2018). STEC O157:H7 is the most frequent STEC serotype associated with foodborne outbreaks and human infections (Bettelheim, 2000; Qin et al., 2015; Li et al., 2017). However, other non-O157:H7 STEC serotypes have also been a major cause of foodborne outbreaks and sporadic cases, and are responsible for up to 50% STEC infections in recent years (Paton et al., 1999; McCarthy et al., 2001; Paciorek, 2002; Liptáková et al., 2005; Johnson et al., 2006; Zhang et al., 2007; European Food Safety Authority, 2011; Frank et al., 2011a; Käppeli et al., 2011; Verstraete et al., 2013; Zweifel et al., 2013; Morton et al., 2017). Among STEC non-O157:H7 serotypes, six serogroups O26, O45, O103, O111, O121 and O45, also known as "The Big six" (comprising nine serotypes: O26:H11/H-; O45:H2; O103:H2, H11, H25; O111:H8/H-; O121:H19 or H7; and O145:H28/H-) account for over 70% of non-O157:H7 STEC infections (Brooks et al., 2005; Hedican et al., 2009; Bosilevac and Koohmaraie, 2011).

Shiga toxin (Stx) is the main characteristic that defines STEC (Nataro and Kaper, 1998; Tarr et al., 2005), which is encoded by *stx* genes located within lambdoid prophages (Stx-converting phages or Stx-phages) (O'Brien et al., 1989; Mizutani et al., 1999; Bryan et al., 2015; Lacher et al., 2016). Shiga toxins are classified into two types, Stx1 and Stx2. Each of Stx type comprises several subtypes with three subtypes for Stx1 (Stx1a, Stx1c and Stx1d) and 10 subtypes for Stx2 (Stx2a, Stx2b, Stx2c, Stx2d, Stx2e, Stx2f, Stx2g, Stx2h, Stx2i and Stx2k) (Scheutz et al., 2012; Lacher et al., 2016; Bai et al., 2018; Yang et al., 2020). Stx1 and/or Stx2 carrying STEC can cause human disease, however, Stx2 is more often associated with HC and HUS (Lentz et al., 2011; Krüger and Lucchesi, 2015). Among Stx2 subtypes, Stx2a is the most prevalent subtype association with severe disease, followed by Stx2c and Stx2d (Feng and Reddy, 2013; Melton-Celsa, 2014; Krüger and Lucchesi, 2015). *Shigella dystenteriae* and some strains of *Shigella sonnei*, *Shigella flexneri* and *E. albertii* also produce Stx (Beutin et al., 1999; Gupta et al., 2007; Ooka et al., 2012; Gray et al., 2014; Murakami et al., 2014; Brandal et al., 2015). In addition to Shiga toxin, some STEC serotypes also carry the locus of enterocyte effacement (LEE) pathogenicity island (McDaniel and Kaper, 1997; Kaper et al., 2004) responsible for adherence during STEC infections.

STEC serotype detection and identification rely on the detection of Stx proteins by enzyme immune assays or detection of the presence of *stx* genes by molecular methods such as PCR (Brian et al., 1992; Milley and Sekla, 1993; Bélanger et al., 2002; Hara-Kudo et al., 2007; Teel et al., 2007; Zhang et al., 2012). Conventional phenotypic serotyping through antigenic agglutination can further classify STEC to the serotype level (Gyles, 2007). However, cross-reactivity, lack of expression of O antigens, a focus on STEC O157:H7 and novel serotypes may all prevent accurate serotyping and lead to under-detection of non-O157:H7 STEC (Liu et al., 2008; Stigi et al., 2012). Molecular methods, including microarrays, utilising the sequence variations in the O antigen gene clusters, have been developed to serotype STEC O157:H7, "Big six" STEC non-O157:H7 and other STEC serotypes (DebRoy et al., 2004; Gonzales et al., 2011; Lin et al., 2011; Norman et al., 2012; Iguchi et al., 2015; Ludwig et al., 2020). More recently, WGS based methods have been developed for *in silico* serotyping STEC, which allow phenotypically untypeable isolates be serotyped *in silico* using O antigen and flagellin H antigen genes (Inouye et al., 2014; Joensen et al., 2015).

Alongside STEC serotyping which is useful in outbreak investigation and for prevalence surveillance (FAO/WHO STEC EXPERT GROUP, 2019), other subtyping methods such as pulsed-field gel electrophoresis (PFGE), multiple locus variable-number tandem repeat analysis (MLVA) and multilocus sequence typing (MLST) were also used for STEC outbreak investigations (Gerner-Smidt et al., 2006; Gyles, 2007; Frank et al., 2011b). Recently, WGS based typing and metagenomic sequencing have shown great potential for STEC surveillance and outbreak investigation with high resolution and specificity (Leonard et al., 2015; Parsons et al., 2016).

STEC serotypes with the same O and H antigens were generally clustered together and share a common ancestor (Ju et al., 2012). A recent phylogenetic analysis on 276 STEC isolates belonging to 81 serotypes revealed that some STECs formed discrete clades with clustering associated with sequence types and serotypes (González-Escalona and Kase, 2019). Our present study aimed to i), identify phylogenetic clusters of STEC through large scale examination of publicly available genomes; ii), identify cluster/serotype-specific genes for detection of STEC isolates and for detection and serotyping of most frequent STEC serotypes through comparative genomic analysis of accessory genomes; iii), develop an automated pipeline for STEC *in silico* cluster typing and serotyping from WGS data based on cluster/serotype-specific gene markers combined with *E. coli* O and H antigen genes.

Abbreviations: STEC, Shiga toxin-producing *Escherichia coli*; HC, haemorrhagic colitis; HUS, haemolytic uraemic syndrome; ESRD, end-stage renal disease; Stx, Shiga toxin; LEE, locus of enterocyte effacement; EIEC, enteroinvasive *E. coli*; MLST, multi-locus sequence typing; rSTs, ribosomal MLST STs; TP, true positives; TPR, true positive rate; TN, true negatives; TNR, true negative rate; FN, false negatives; FP, false positives.

MATERIALS AND METHODS

Identification of STEC Isolates From NCBI Database

E. coli isolates from the NCBI SRA (National Center for Biotechnology Information Sequence Read Archive) in June of 2020 were queried. The keyword "*Escherichia coli*" was used to retrieve SRA accession numbers of *E. coli* isolates. Raw reads were retrieved from ENA (European Nucleotide Archive). The *stx* genes (stx_1, GenBank accession number M19437; stx_2 GenBank accession number X07865) and *ipaH* gene (GenBank accession number M32063) were used to screen *E. coli* reads using Salmon v0.13.0 (Patro et al., 2017). Taxonomic classification for *E. coli* was confirmed by Kraken v1.1.1 (Wood and Salzberg, 2014). Isolates that were positive to any *stx* genes and negative to the *ipaH* gene (the latter to exclude *Shigella* or enteroinvasive *E. coli* [EIEC]) were selected to form the STEC dataset.

A control dataset that represented the sequence types (STs) and ribosomal STs (rSTs) of *stx* negative *E. coli* ("non-STEC") isolates was constructed. STs and rSTs of non-STEC isolates were obtained from the *E. coli/Shigella* database in Enterobase on August 2020 (Zhou et al., 2020). For STs and rSTs with only one isolate, the isolate was selected. For STs and rSTs with more than one isolate, one representative isolate for each ST and rST were randomly selected. In total, 14,126 *stx*-negative *E. coli* isolates representing 4,354 STs and 11,520 rSTs were selected as non-STEC control database.

Genome Assembly and Data Processing

Raw reads were assembled *de novo* using SPADES v3.14.0 assembler with default settings [http://bioinf.spbau.ru/spades] (Bankevich et al., 2012). The metrics of assembled genomes were obtained with QUAST v5.0.0 (Gurevich et al., 2013). Three standard deviations (SD) from the mean for contig number, largest contig, total length, GC, N50 and genes were used as quality filter for assembled genomes.

The STs for isolates in the STEC database were checked using mlst (https://github.com/tseemann/mlst) with the *E. coli* scheme from PubMLST (Jolley and Maiden, 2010). rSTs were extracted from the *E. coli/Shigella* rMLST database in Enterobase on August 2020 (Zhou et al., 2020). Serotyping of *E. coli* O and H antigen types were predicted by using SerotypeFinder v2.0.1 (Joensen et al., 2015). The phylogroups of STEC isolates were obtained using ClermonTyping (Beghain et al., 2018).

Selection of Isolates for STEC Identification Dataset

Representative isolates for each ST, rST and serotype in the STEC dataset were selected to form the identification dataset. For STs, rSTs and serotypes with only one isolate, the isolate was selected. For STs, rSTs and serotypes with more than one isolate, one representative isolate for each ST, rST and serotype was randomly selected. For rSTs in the top six STs, one representative isolate for each rST with two or more isolates was randomly selected. A further 691 isolates including 72 *Escherichia coli* reference (ECOR) strains downloaded from Enterobase, 573 non-STEC *E.coli* isolates representing 573 STs with more than nine genomes, 41 *Shigella* and EIEC isolates representing each cluster identified in our previous study (Zhang et al., 2021), three *E. albertii* isolates and two *E. fergusonii* isolates were used as controls for the identification dataset. The details of the identification dataset are listed in **Table S1**. The remaining STEC isolates in the STEC database were referred to as the validation dataset (**Table S2**).

The identification dataset was used to identify the phylogenetic relationships of STEC isolates and was also used to identify cluster/serotype-specific gene markers. The validation dataset was used to evaluate the performance of cluster/serotype-specific gene markers relative to phylogenetic relationships.

Phylogeny of STEC Isolates Based on WGS

Phylogenetic trees including an identification tree and 15 validation trees were constructed by using Quicktree v1.3 (Hu et al., 2020) with default parameters to identify and confirm the phylogenetic clustering of STEC isolates. The phylogenetic trees were visualised by Grapetree and ITOL v5 (Zhou et al., 2018; Letunic and Bork, 2019).

The identification phylogenetic tree was generated using isolates in the identification dataset for the identification of clusters of STEC isolates. The validation trees were constructed using isolates in the STEC validation dataset and a subset of isolates known to represent each identified cluster from the identification dataset to assign validation dataset isolates to the clusters defined.

Identification of the Cluster/Serotype-Specific Gene Markers

Cluster/serotype-specific gene markers were identified from STEC accessory genomes. The genomes from the identification dataset were annotated using PROKKA v1.13.3 (Seemann, 2014). Pan- and core-genomes were analysed by Roary v3.12.0 (Page et al., 2015) using an 80% sequence identity threshold. The candidate gene markers specific to each cluster/serotype were identified from accessory genes with an in-house python script from our previous study (Zhang et al., 2021). The best performing specific gene marker set was selected from the candidates by using BLASTN to search against the identification dataset.

As in our previous studies (Zhang et al., 2019; Zhang et al., 2021) the genomes from a given cluster containing all specific gene markers for that cluster were termed true positives (TP), the genomes from the same cluster lacking any of those same gene markers were termed false negatives (FN). The genomes from other clusters containing all of those same gene markers were termed false positives (FP). The sensitivity (True positive rate, TPR) of each cluster-specific gene marker was defined as TP/(TP+FN). The specificity (True negative rate, TNR) was defined as TN/(TN+FP).

Validation of the Cluster/Serotype-Specific Gene Markers

The specific gene markers were examined by using BLASTN to search against the validation dataset (**Table S2**) and non-STEC *E. coli* control database for the presence of any of the cluster/

serotype-specific gene markers. The BLASTN thresholds were defined as 80% sequence identity and 50% gene length coverage.

Development of STECFinder, an Automated Pipeline for Molecular Serotyping of STEC

STECFinder was developed for STEC serotyping from either paired end Illumina genome sequencing reads or assembled genomes. The typing reference sequences used for construction of STECFinder included specific gene marker sets identified in this study, established *E. coli* O antigen and H antigen gene sequences collected from SerotypeFinder (Joensen et al., 2015), *stx* subtype sequences collected from VirulenceFinder and three other studies (Joensen et al., 2014; Lacher et al., 2016; Bai et al., 2018; Yang et al., 2020), the *ipaH* gene sequence downloaded from NCBI, and seven House Keeping (HK) genes -*recA*, *purA*, *mdh*, *icd*, *gyrB*, *fumC* and *adk* from the *E. coli* MLST scheme (Jolley and Maiden, 2010) for contamination checking (**Figure 1**). All sequences are available in fasta format at https://github.com/LanLab/STECFinder with cluster specific genes named with the following convention: STEC-cluster-gene_number (i.e. STEC-C1-gene_1 for the first gene in the C1 specific set).

For the analysis of sequence data as raw reads, KMA (*k*-mer alignment) v1.3.15 (Clausen et al., 2018) was used to align the raw reads to the typing reference sequences. KMA utilizes *k*-mer seeding and the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) to accurately align reads to genes of interest. The best-aligning template was chosen from a novel sorting scheme ConClave scheme incorporated into KMA (Clausen et al., 2018). To determine whether the genes were present or absent, the

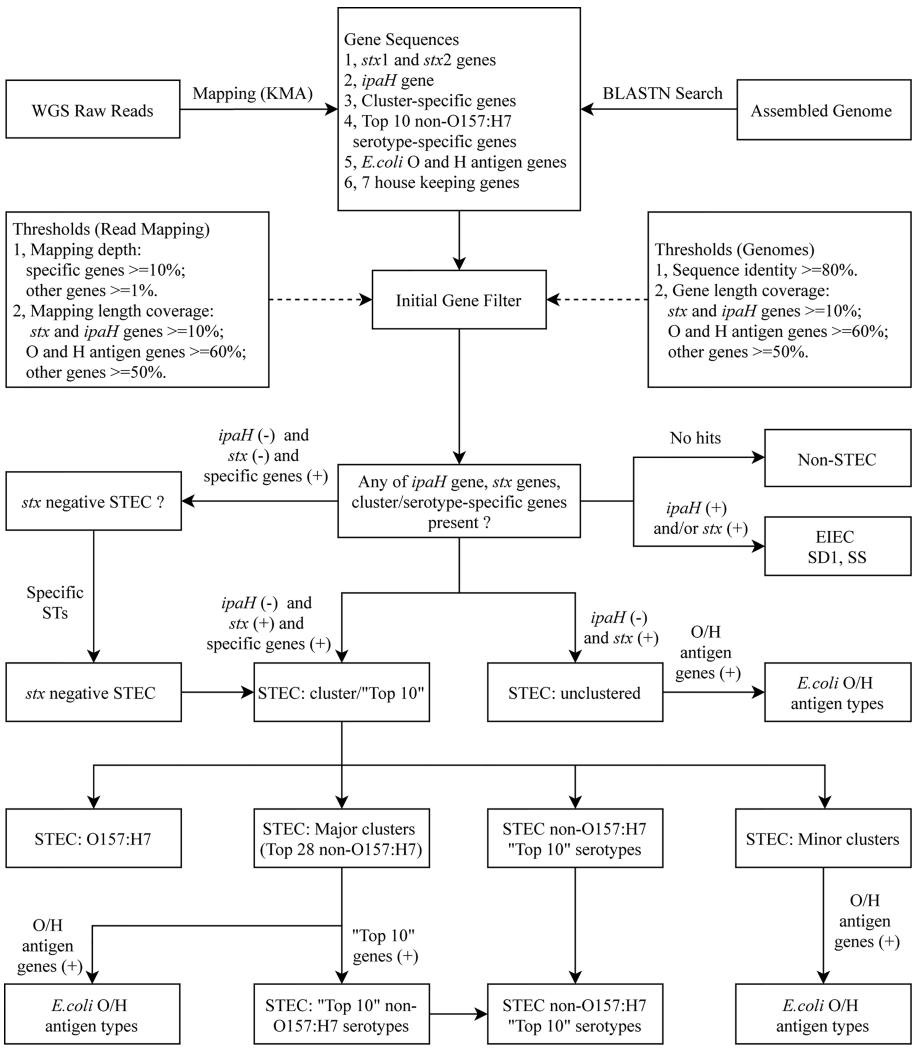

FIGURE 1 | *in silico* serotyping pipeline workflow. Schematic of *in silico* serotyping STEC by cluster/serotype-specific genes combined with the *ipaH* gene, *stx* genes including all available subtypes and *E. coli* O antigen and H antigen genes, implemented in STECFinder. Both assembled genomes and raw reads are accepted as data input.

mapping length coverage and a minimum depth were used as the thresholds.

For the submission of sequence data as assembled genomes, BLASTN v2.9.0 (Camacho et al., 2009) was used to search against the typing reference sequences with 80% sequence identity. The presence or absence of genes was determined by the gene length coverage.

The presence or absence of genes in STECFinder was determined by the cutoff value of gene length coverage for assembled genomes and the mapping length coverage and a minimum mapping depth for raw reads. For assembled genomes, length coverage of 50% for all cluster/serotype-specific genes, 60% for O and H antigen genes and 10% for *ipaH* gene and *stx* genes were used as cutoff value for determination of the presence of genes. For raw reads, mapping length coverage of 50% for all cluster/serotype-specific genes, 60% for O and H antigen genes, 10% for *ipaH* gene and *stx* genes and a minimum depth of 10 for all cluster-specific genes, a minimum depth of one for O and H antigen genes, *ipaH* gene and *stx* genes were used to define the gene as present. In addition, when multiple O and H genes were detected the bitscore was incorporated into STECFinder for filtering and ranking O and H antigen. The highest match was chosen as the O or H antigen present, when multiple O or H variants were present.

The major and minor clusters and top 10 non-O157:H7 STEC serotypes were assigned based on the presence of cluster/serotype-specific gene marker set together with the presence of *stx* subtypes and the absence of *ipaH* gene. All genes in a cluster/serotype-specific gene set must be defined as present for a cluster or serotype to be called. An 'unclustered' was assigned for isolate that cannot be detected by any of cluster-specific gene marker set. Unclustered STEC could be new clusters or isolates that contained all genes in the marker set but one or more genes from marker set did not pass the cutoff value.

Additional subsets of gene marker sets were added to increase the accuracy of clusters and calling of the top 10 non-O157:H7 STEC serotypes. For example, the combination of the specific gene marker set of O157:H7 and AM18 can eliminate the known false presences of AM18 gene set in O157:H7. The isolate is assigned as AM18 if both gene sets are present while the isolate is assigned as O157:H7 if AM18 specific gene set is absent. The subsets of combined gene sets were incorporated into the STECFinder for elimination of false cluster assignment are listed in **Table S6**.

STECFinder was tested with identification dataset. The accuracy and specificity of STECFinder for prediction of clusters and serotypes were evaluated with STEC validation dataset and non-STEC *E.coli* control dataset.

Application of STECFinder in STEC Typing Using Metagenomics Data From STEC Spiked Food Samples

STECFinder can take input from metagenomics sequencing reads for STEC typing. The application of STECFinder in metagenomics analysis was evaluated using 17 metagenomic sequencing read sets from samples published by Buytaers et al. (Buytaers et al., 2020). These 17 shotgun metagenomic sequencing reads (Buytaers et al., 2020) were downloaded from ENA.

RESULTS

Screening Sequenced Genomes for STEC Isolates

The presence of any of *stx* genes and the absence of the *ipaH* gene were used to identify STEC isolates. We examined 140,348 isolates with the species annotation of *E. coli* with paired end Illumina sequencing reads available in ENA database. Of the 140,348 isolates, 43,960 isolates were positive to stx_1 and/or stx_2 genes and negative for the *ipaH* gene. 41,101 of the 43,960 isolates passed taxonomic classification and genome assembly quality filters and were selected to form the STEC dataset.

Isolates in the STEC dataset were typed using MLST, rMLST and SerotypeFinder. MLST typed the 41,101 STEC isolates into 817 STs (202 isolates not typed by MLST) of which 368 STs were represented by a single isolate, 424 STs represented by two to 100 isolates each and accounted for 12% of the STEC isolates, whereas 25 STs contained more than 100 isolates each and encompassed 86.61% of the STEC isolates, of which ST11 is the largest, accounting for 37.12% of the STEC isolates, followed by ST21 (14.71%), ST17 (11.91%), ST16 (6.72%), ST655 (2,71%) and ST32 (2.46%). rMLST divided the 41,101 STEC isolates into 2,911 rSTs (12,208 isolates not typed by rMLST).

Using SerotypeFinder, 38,958 of the 41,101 (94.79%) isolates were assigned to 460 *E. coli* O:H antigen types, 2,039 isolates (4.96%) were not assigned an O antigen and were typed for H antigens only with 38 H antigen types, of which H7, H2, H8, H11 and H21 were the most frequent types, 96 isolates (0.23%) were typed as multiple O:H types and six isolates (0.01%) were untypeable.

The Frequency of STEC Serotypes

The 38,958 STEC O:H antigen typeable isolates belonged to 460 different serotypes including O157:H7 (38.55% of 38,958 typeable isolates) and 459 non-O157:H7 serotypes (61.45% of 38,958 typeable isolates).

Of the 459 non-O157:H7 serotypes, the top 28 serotypes were present in more than 100 isolates each and accounted for 50.8% of 38,958 typeable STEC isolates, of which the 10 most frequent serotypes (41.66% of 38,958 typeable STEC isolates) were O26:H11, O103:H2, O111:H8, O121:H9, O145:H28, O45:H2, O91:H14, O118/O151:H16, O123/O186:H2 and O146:H21. The top 6 serotypes belonged to the well-known "Big six" STEC non-O157 serogroups (Brooks et al., 2005; Hedican et al., 2009; Bosilevac and Koohmaraie, 2011). It should be noted that three serotypes, O103:H11, O103:H25, and O121:H7, belonging to the "Big six" non-O157 STEC serogroups were outside the top 10 serotypes. The 116 serotypes present with 10 to 100 isolates each, belonged to 8.64% of typeable STEC isolates. The remaining 315 serotypes with less than 10 isolates each represented 2% of the typeable STEC isolates (**Figure 2**).

Identification of STEC Clusters

To identify any phylogenetic clusters containing one or more STEC serotypes from the 41,101 STEC isolates, we selected representative isolates to perform phylogenetic analysis as it was impractical to construct a tree with all isolates. The selection was performed on

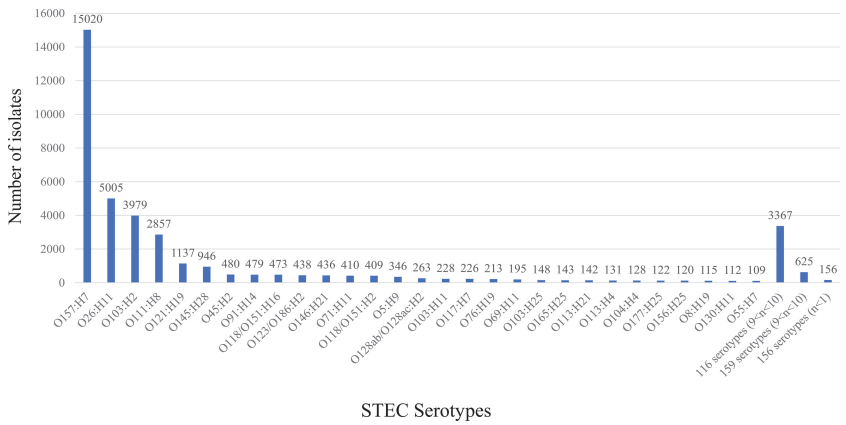

FIGURE 2 | The frequency of 460 STEC serotypes. The graph shows the frequency of 460 STEC serotypes. STEC O157:H7 and top 28 non-O157:H7 serotypes are listed separately. The number on top of each stacked column refers to the number of isolates for each serotype.

the basis of ST, rST and serotype of the 41,101 STEC isolates. One isolate was selected to represent each ST, rST and serotype for a total of 2,567 STEC isolates. Note that in the case that STs or rSTs overlapped with serotype, an isolate was only selected once to avoid duplicates of the same isolate. The selection included 817 STs, 1,413 rSTs, 460 STEC serotypes and 102 partial antigen types (H antigen only and multiple O/H types). A further 691 isolates consisting of 72 ECOR isolates, 573 non-STEC *E.coli* isolates, 41 *Shigella* and EIEC isolates, three *E. albertii* isolates and two *E. fergusonii* isolates were also included. The identification dataset consisted of 3,258 isolates in total. Details are listed in **Table S1**. A phylogenetic tree was constructed using 3,258 isolates in the identification dataset to identify the clusters (**Figure 3**).

The identification of clusters was focused on O157:H7 and the top 28 non-O157:H7 serotypes. A major cluster was defined if the cluster node had a bootstrap value of above 80% and contained STEC isolates belonging to O157:H7 and top 28 non-O157:H7 serotypes. The isolates of O157:H7 were

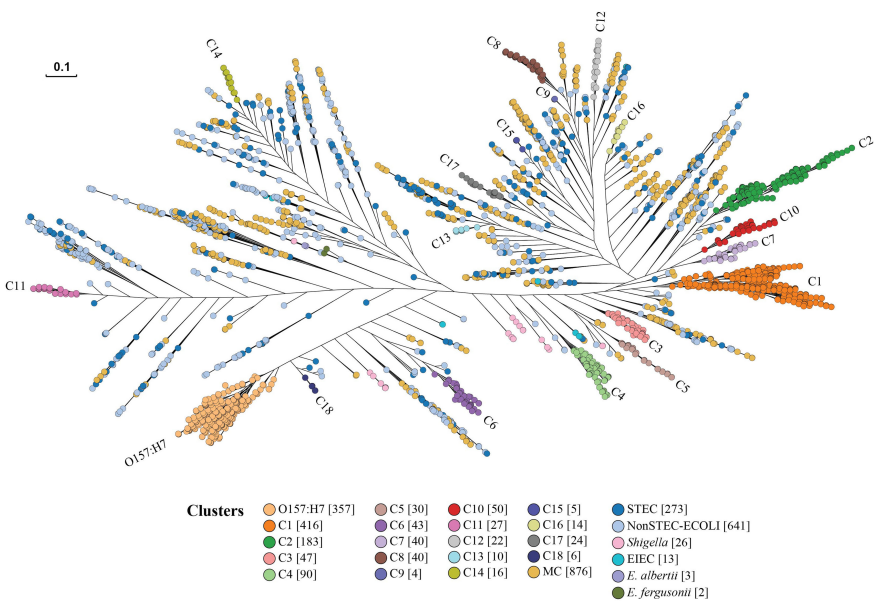

FIGURE 3 | STEC cluster identification phylogenetic tree. Representative isolates from the identification dataset were used to construct the phylogenetic tree by Quicktree v1.3 to identify STEC (Shiga toxin-producing *E. coli*) clusters and visualised using Grapetree. The dendrogram shows the phylogenetic relationships of 2,567 STEC isolates represented in the identification dataset. Branch lengths are log scale for clarity. The scale bar represents 0.1 substitutions per site. STEC clusters are coloured. Numbers in square brackets after cluster name are the number of isolates for each identified cluster. ECOLI is *E. coli.* EIEC is Enteroinvasive *E. coli*. MC indicates a minor STEC cluster.

grouped into one large cluster. A further 18 major clusters (C1-C18) all of which carried only non-O157:H7 serotypes (**Figure 3**; **Table 1** and **Figure S1**), were identified. The isolates of top 28 non-O157:H7 serotypes fell into these 18 major clusters. Of the 2,567 STEC isolates, 1,412 fell within the O157:H7 cluster or one of the 18 non-O157:H7 major STEC clusters.

Of the remaining 1,155 STEC isolates, 877 isolates were grouped into 229 STEC minor clusters with two or more isolates in a cluster, whereas 278 isolates were singletons separated from other clusters by non-STEC *E. coli* isolates. We further typed the isolates from minor clusters using phylogroup typing (Brooks et al., 2005) and each minor cluster was named by phylogroup and lineage number, for example, phylogroup A minor cluster 1 (AM1). Most of the minor clusters belonged to phylogroup B1 (**Table 2**).

In total, 19 major STEC clusters including one O157:H7 and 18 non-O157:H7 clusters (Top 28 non-O157:H7 serotypes) and 229 STEC minor clusters containing other non-O157:H7 serotypes were identified. Of the 19 major clusters, 12 had a single serotype and seven had two or more serotypes. The frequency of non-O157:H7 STEC serotypes in the major clusters are shown in **Figure 4**. For the 229 STEC minor clusters, 103 contained a single serotype, 109 consisted of two or more serotypes and the remaining 17 comprised of isolates with H antigen types only.

Among the top 10 non-O157:H7 serotypes, O121:H19 (C5), O145:H28 (C6), O91:H14 (C7) had a single origin while O146:H21 (C8 and C9) was a paraphyletic serotype. O26:H11 and O118/O151:H16 were grouped into C1. O123/O186:H2 was grouped into C2. O103:H2, O111:H8 and O45:H2 had polyphyletic origins. O103:H2 and O111:H8 were grouped into C2 and B1M118, C1 and B1M119, respectively. O45:H2 had three lineages which were clustered into C2, C3 and AM37. Three serotypes (O128ac:H2, O8:H19 and O113:H21) of the remaining top 28 non-O157:H7 serotypes were polyphyletic serotypes. Thirty non top 28 non-O157:H7 serotypes also had polyphyletic origins.

Apart from STEC isolates, 26 of the 573 *stx* negative *E. coli* isolates from the identification dataset were grouped into clusters. Of the 19 major clusters identified, 12 contained *stx* negative *E. coli* isolates (ST11 in O157:H7; ST765 and ST29 in C1; ST17, and ST376 in C2; ST343 andST300 in C3, ST342 in C4;

TABLE 1 | Major STEC clusters identified in identification dataset.

Cluster	No. of isolates	No. of serotypes	No. of STs	Top 28 non-O157:H7 serotypes*
O157:H7	356	1	83	O157:H7
C1	414	30	97	1-O26:H11, 3-O111:H8, 8-O118/O151:H16, 12-O71:H11, 15-O103:H11, 18-O69:H11
C2	181	16	42	2-O103:H2, 6-O45:H2, 9-O123/O186:H2 11-O118/O151:H2
C3	45	18	12	19-O103:H25, 25-O156:H25, 6-O45:H2
C4	89	14	21	13-O5:H9, 20-O165:H25, 24-O177:H25
C5	29	1	5	4-O121:H19
C6	41	1	6	5-O145:H28
C7	40	2	13	7-O91:H14
C8	40	1	14	10-O146:H21
C9	4	1	1	10-O146:H21
C10	50	2	15	14-O128ab:H2
C11	27	1	6	16-O117:H7
C12	21	1	6	17-O76:H19
C13	10	1	7	21-O113:H21
C14	16	2	2	22-O113:H4
C15	5	1	1	23-O104:H4
C16	14	1	4	26-O8:H19
C17	24	11	7	27-O130:H11
C18	6	1	1	28-O55:H7

**The serotypes in each non-O157:H7 cluster are listed with their rank by isolate frequency for the top 28 non-O157:H7 serotypes followed by the serotype.*

TABLE 2 | Summary of identified STEC minor clusters in identification dataset.

Phylogroup	No. of MC*	Name of MC	No. of isolates	No. of serotypes	No. of STs
A	37	AM1-AM37	139	64	42
B1	126	B1M1-B1M126	519	157	186
B2	14	B2M1-B2M14	35	20	17
C	7	CM1-CM7	17	10	8
D	22	DM1-DM22	67	26	29
E	19	EM1-EM19	73	26	34
G	4	GM1-GM4	27	12	12

**MC, minor clusters.*

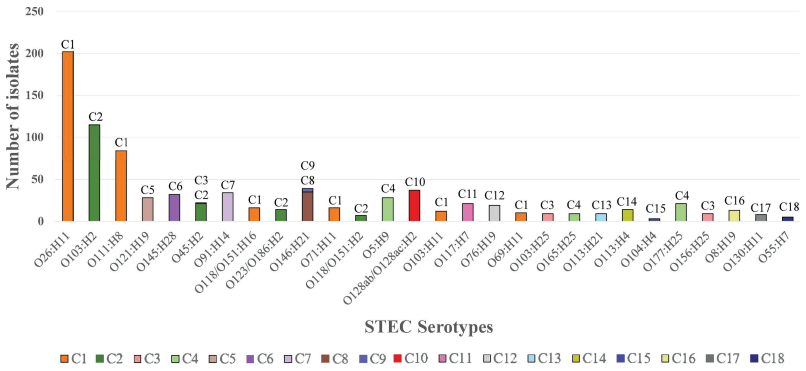

FIGURE 4 | The frequency of the top 28 non-O157:H7 STEC serotypes in STEC major clusters. The graph shows the frequency of top 28 non-O157:H7 serotypes in the 18 STEC major clusters. Clusters are shown per colour legend and also at the top of the bar. X-axis shows the serotype while y-axis shows the number of isolates.

ST655 in C5; ST32 in C6; ST442 and ST1992 in C8, ST335 in C18). These STs containing *stx* negative *E. coli* isolates were the most frequent STs in the STEC database, suggesting these *stx* negative *E. coli* isolates may have lost the *stx* genes. The details of STEC clusters and lineages were listed in **Table S3**.

However, 11 STEC minor clusters also contained *stx* negative *E. coli* isolates. Therefore, we further examined STs with more than two isolates from all minor STEC clusters that were also found within the 14,126 *stx* negative *E. coli* ("non-STEC") isolates. Of the 229 minor STEC clusters, the STs in 58 clusters contained *stx* positive isolates only and the STs in 171 clusters contained both *stx* negative and *stx* positive isolates. Of these 171 minor STEC clusters, the STs in four clusters consisted of *stx* positive isolates and *E. coli* isolates that didn't carry typical pathotype specific genes (data not shown). While STs in the remaining 167 clusters consisted of *stx* positive isolates and isolates that carried pathotype specific genes from other *E. coli* pathotypes (data not shown). Thus, these STEC minor clusters are a mix of STEC and other pathotypes.

Identification of the Cluster/Serotype-Specific Gene Markers

In this study, we used the same definition and approach as used to find the *Shigella*/EIEC cluster specific genes (Zhang et al., 2021). We searched for potential specific gene marker sets for the 19 major and 229 minor clusters using the accessory genomes from the 3,258 identification dataset isolates. Genes associated with STEC O antigen gene clusters were excluded from the analysis to identify O antigen gene independent markers. Multiple candidate cluster/serotype-specific gene marker sets for each of the 19 major STEC clusters and 229 minor STEC clusters were identified. The single gene marker set with 100% sensitive and the highest specificity were then selected from candidate cluster-specific gene marker sets by BLASTN searches against genomes in the identification dataset using 80% sequence identity and 50% gene length threshold.

We also searched for specific gene markers for six of the top 10 non-O157:H7 serotypes (O26:H11, O111:H8, O118/O151:H16, O103:H2, O45:H2 and O123/O186:H2) which were not in a cluster of their own. The best performing gene marker set for each of six of top 10 non-O157:H7 serotypes were identified using the same approach as used to identify and select cluster-specific gene marker sets.

The sensitivity and specificity of each major STEC cluster and six non-O157:H7 serotype specific gene marker set for the identification dataset were listed in **Table 3**. The major STEC cluster and six non-O157:H7 serotype specific gene marker sets were all 100% sensitive and the specificity varied from 99.72% to 100% for major STEC cluster-specific gene marker sets and from 99.41% to 100% for non-O157:H7 serotype-specific gene marker sets. The STEC minor cluster-specific gene marker sets were 100% specific with the exception of 12 minor clusters which had specificity ranging from 99.85% to 99.97% (**Table S4**).

Validation of Cluster/Serotype-Specific Gene Markers

The STEC cluster/serotype-specific gene marker sets were evaluated with 38,534 STEC isolates from the validation dataset and 14,126 isolates from non-STEC *E.coli* control dataset.

The STEC cluster-specific gene marker sets were able to assign 35,464 of 38,534 (92.03%) STEC isolates to the major clusters and 2,703 (7.01%) STEC isolates to minor clusters. In total, 38,155 of 38,534 (99.02%) STEC isolates can be assigned to clusters by cluster-specific gene marker sets, while 150 of the 38,534 (0.39%) STEC isolates were assigned with more than one cluster and 217 of the 38,534 (0.56%) STEC isolates were not assigned to any cluster by STEC cluster-specific gene marker sets.

Validation phylogenetic trees (**Figure S2**) were then constructed to confirm the assignment of cluster-specific gene marker sets. We divided the 38,534 STEC validation isolates into 15 subgroups. Each of the 15 subgroups isolates together with a subset of 476 STEC isolates with known clusters and 691 non-STEC isolates from identification dataset were used to generate validation trees for a total of 15 validation trees. The validation isolates were considered to truly belong to a given cluster if the isolates were found within a branch that only contained identification dataset isolates from that cluster with a bootstrap value of 80% or greater. In total 38,340 (99.5%) validation

TABLE 3 | The sensitivity and specificity of STEC cluster/serotype-specific gene markers.

Clusters	Cluster-specific genes marker sets	Identification dataset (3,258 isolates)		
		No. of isolates	Sensitivity	Specificity*
O157:H7	Set of 6 genes	356	100	99.72
C1	Set of 4 genes	414	100	99.82
C2	Set of 4 genes	181	100	99.97
C3	Set of 3 genes	45	100	100
C4	Set of 3 genes	89	100	99.97
C5	Set of 4 genes	29	100	100
C6	Set of 3 genes	41	100	99.88
C7	Set of 4 genes	40	100	99.97
C8	Set of 5 genes	40	100	99.97
C9	Set of 2 genes	4	100	100
C10	Set of 2 genes	50	100	100
C11	Single gene	27	100	100
C12	Set of 2 genes	21	100	100
C13	Set of 4 genes	10	100	100
C14	Set of 4 genes	16	100	99.97
C15	Set of 2 genes	5	100	100
C16	Set of 4 genes	14	100	99.97
C17	Set of 3 genes	24	100	99.97
C18	Set of 3 genes	6	100	99.97
O26:H11	Set of 6 genes	204	100	99.41
O103:H2	Set of 4 genes	121	100	99.87
O111:H8	Set of 3 genes	96	100	100
O45:H2 (C2)	Set of 5 genes	22	100	99.97
O45:H2 (C3)	Set of 3 genes	1	100	100
O118/O156:H16	Set of 4 genes	17	100	99.94
O123/O186:H2	Set of 3 genes	21	100	100

**The specificity of cluster-specific gene set less than 100% was due to at least one false positive found in that set.*

isolates were assigned to major and minor STEC clusters with 35,574 (92.32%) and 2,766 (7.18%) respectively, while the remaining 194 isolates (0.5%) were not assigned to any clusters.

Compared to cluster assignment by phylogenetic trees as the ground truth, cluster-specific gene marker sets correctly assigned 35,461 validation isolates to major clusters and 2,704 validation isolates to minor clusters. Cluster -specific gene marker sets also correctly identified 191 of the 194 isolates without cluster assignments. In total the accuracy of assignments by cluster -specific gene marker sets were 99.54%. The sensitivity and specificity for each cluster-specific gene marker set for validation dataset were listed in **Table S4**.

The STEC cluster specific gene marker sets were validated on 14,216 non-STEC *E. coli* isolates. The specificity of the STEC cluster-specific gene markers set for major clusters varied from 99.38% to 100% and the specificity of the STEC cluster-specific gene marker sets for minor clusters ranged from 97.25% to 100%. Details are listed in **Table S5**.

STECFinder for Molecular Serotyping of STEC Isolates and Its Accuracy and Specificity

STECFinder was developed for cluster and serotype identification of STEC isolates. Cluster was identified using cluster -specific gene marker sets and serotype was identified using serotype-specific gene markers as well as *E. coli* O and H antigen genes within clusters. Either paired end Illumina genome sequencing reads or assembled genomes can be used. STECFinder is available on github (https://github.com/LanLab/STECFinder).

The accuracy and specificity of STECFinder for STEC typing were tested with 3,258 isolates from the identification dataset. For assembled genomes, all 1,412 STEC isolates belonging to 19 major clusters and all 877 STEC isolates belonging to 229 minor clusters were correctly predicted, while 26 of 573 *stx* negative *E.coli* isolates were assigned to STEC clusters by their corresponding cluster-specific gene marker sets. Eighteen STEC singletons were assigned to clusters or minor clusters. For read mapping, two of 1,412 isolates belonging to the 19 major clusters and 25 of 877 isolates from minor clusters were not detected by cluster-specific gene marker sets, while 26 *stx* negative *E.coli* was assigned to STEC clusters similar to the assignment using the assembled genomes. The accuracy of STECFinder for cluster assignments was 99.45% and 98.5% for assembled genomes and read mapping respectively. The accuracy of cluster assignment for the top 10 non-O157:H7 serotypes was 99.14% and 99.11% for assembled genomes and read mapping, respectively.

STECFinder was validated on 38,534 isolates from the STEC validation dataset. Compared to the ground truth assignments determined using phylogenetic analysis, STECFinder assigned 99.85% and 99.83% of validation isolates correctly to clusters for assembled genomes and read mapping, respectively. The accuracy of cluster assignment for top 10 non-O157:H7 serotypes was 99.72% for assembled genomes and 99.65% for

read mapping. For the 38,534 *stx*-positive isolates from validation dataset, STECFinder demonstrated 100% cluster assignment specificity for both assembled genomes and read mapping. The cluster assignment specificity of STECFinder was further evaluated using the 14,126 *stx*-negative *E. coli* isolates from the "non-STEC" control dataset. The specificity was 87.07% and 85.12% for assembled genomes and read mapping, respectively. Further investigation of the false positive isolates found that 1,074 false positive isolates belonged to the STEC cluster based on phylogenetic analysis. After removing all of these false positive isolates, the specificity was 94.66% and 92.72% for assembled genomes and read mapping respectively.

STECFinder can assign STEC isolates to serotype level within predicted clusters. The comparison of *in silico* serotyping of the total of 41,101 STEC isolates between STECFinder and SerotypeFinder (Joensen et al., 2015) was performed. For assembled genomes, the serotype prediction of 40,912 of 41,101 (99.54%) STEC isolates by STECFinder agreed with that by SerotypeFinder when applying the same cutoff values of 80% sequence identity and 60% length coverage. For the remaining 189 STEC isolates with non-identical serotype prediction, STECFinder predicted serotypes were largely a subset of O:H types predicted by SerotypeFinder. For example, an isolate may be assigned as wzx_O103 and H2 by STECFinder while SerotypeFinder predicted as a mixed wzx_O103/O26 and H2/H11.

There were 40,618 of 41,101 (98.82%) STEC isolates with the same serotype prediction by STECFinder and SerotypeFinder from read mapping. For the remaining 483 cases, STECfinder assigned a full serotype while SerotypeFinder assigned 257 isolates with H antigen only, 117 isolates with O antigen only and 109 isolates with multiple O:H types.

Detection of STEC Clusters and Serotypes Using STECFinder in Spiked Food Samples Using Shotgun Metagenomic Sequencing Reads

The application of STECFinder in metagenomics analysis was evaluated with 17 metagenomic sequencing reads from samples published by Buytaers et al. (Buytaers et al., 2020). The 17 metagenomic samples consisted of nine minced beef meat samples spiked with a STEC O157:H7 isolate, one fresh goat cheese sample each spiked with STEC O145:H28 isolate, O103:H2 isolate and co-spiked with STEC O103:H2 and O145:H28 isolates and five STEC negative control food samples. Samples were spiked with STEC isolates at the lowest infectious dose (<10 CFU for 25 g of food) (Buytaers et al., 2020).

STECFinder assigned the nine samples spiked with STEC O157:H7 to O157:H7 cluster, one sample with STEC O145:H28 to C6 (O145:H8), one sample spiked with STEC O103:H2 to C2 and O103:H2 (O103:H2 is within C2). One sample co-spiked with STEC O103:H2 and O145:H28 was assigned to C2 and O103:H2 (O103:H2 is within C2), and C6 (O145:H8). The cluster/serotype-specific gene marker sets were not detected in the five control samples and STECFinder assigned the five sequenced reads of STEC negative control to "Other-*E. coli*". STECFinder correctly typed the spiked samples using cluster/serotype-specific gene markers.

DISCUSSION

In this study, we performed genomic analysis of more than 41,000 STEC genomes representing 460 different serotypes and identified 19 major phylogenetic clusters including one O157:H7 cluster and 18 non-O157:H7 clusters containing the 28 most frequent non-O157:H7 serotypes, and 229 minor clusters. WGS-based phylogenetic analysis of such a large set of genome data found that STEC had far greater genetic diversity than what has been observed previously with clusters containing one or more serotypes. The close phylogenetic relationship between O26:H11, O111:H8 and O103:H11 in C1, O103:H2 and O45:H2 in C2 agreed with previous studies (González-Escalona and Kase, 2019; Zhang et al., 2020). With the large number of serotypes (460) as well as polyphyletic or paraphyletic origin of 37 serotypes, identification of serotype specific markers for all serotypes was not possible. However, cluster specific markers were identified and used to develop a pipeline, STECFinder, to facilitate cluster and serotype identification of STEC isolates.

STEC infections have a significant impact on public health worldwide (FAO/WHO STEC EXPERT GROUP, 2019). Early detection and differentiation of STEC is vital for food safety surveillance and public health. The initial screening of *stx* genes for STEC serotype detection may lead to misdiagnosis of STEC because *stx* genes can be lost or gained (FAO/WHO STEC EXPERT GROUP, 2019). We identified a small number of *stx*-negative *E. coli* isolates that were grouped into STEC clusters with the corresponding STEC serotypes and STs. Whether these *stx*-negative *E.coli* isolates lost *stx*-containing prophages or were the progenitors of STEC remains unknown. It may also be possible that only a subset of isolates within those STs were *stx* positive due to recent acquisition of *stx*. However, human infections caused by *stx*-negative isolates with typical STEC serotypes have been reported previously (Bielaszewska et al., 2007; Mora et al., 2012; Ferdous et al., 2015). STECFinder will predict these typical STEC serotypes based on cluster/serotype-specific gene markers even if *stx* is absent. It should be noted that STECFinder does not make determination whether a given isolate is an STEC as its key utility is to predict predefined STEC clusters and serotypes. The presence and identity of *stx* genes is also reported to allow the user to make their own determination.

Our analysis found some minor clusters as well as STs contain both *stx* negative and *stx* positive isolates with *stx* negative isolates being of other *E. coli* pathotypes, which suggests that the STEC within those clusters and STs may be hybrid pathogens. Such hybrids have been recognised in recent years including the well-known STEC/EAEC (enteroaggregative *E. coli*) hybrid O104:H4 (ST678) and STEC/UPEC (uropathogenic *E. coli*) hybrid O2:H6 (ST141) (Navarro-Garcia, 2014; Gati et al., 2019). Therefore, for STEC clusters, serotypes or STs that carry isolates with different pathogenicity, a note of caution on the use of STECFinder is required as such clusters identified may not uniquely contain STEC pathogens. More data are needed to determine how many serotypes or STs carry different pathotypes and STECFinder does not attempt to determine other or hybrid pathogenic types. Determining whether an isolate is a hybrid pathogen is often

difficult as some pathogenic types were not well defined by the presence of virulence genes.

Serotyping provides valuable information on identification of potential pathogenic STEC (Gyles, 2007; World Health Organization, 2019). Current serotyping methods focus on well-known O157 and "Big six" non-O157 serogroups which account for about 70% of STEC infections. There are many challenges for the detection of other non-O157:H7 serotypes which cause the remaining 30% of STEC infections (DebRoy et al., 2011; Norman et al., 2012; Zweifel et al., 2013; Smith et al., 2014). In addition, not all STEC can be serotyped *in silico* or predicted based on O or H type genes from genome sequencing data (Joensen et al., 2015; González-Escalona and Kase, 2019). STECFinder can accurately predict STEC serotypes including those most frequently associated with foodborne outbreaks and severe disease. STECFinder can also accurately predict other non-"Big six" non-O157:H7 STEC serotypes. This could be beneficial for identification of the most frequent STEC serotypes for early diagnosis and for clinical management and will better inform the genomic surveillance of STEC serotypes.

We verified the serotype of STEC isolates predicted using STECFinder by phylogenetic cluster assignment and shared STs with STEC isolates of known serotypes. Compared with the existing pipeline for *E. coli in silico* serotyping, SerotypeFinder (Joensen et al., 2015), cluster/serotype-specific gene markers based STECFinder can eliminate the majority of uncertain antigen type calls and provides more accurate STEC serotyping within predicted clusters. STECFinder will be useful for epidemiological and diagnostic investigations as well as providing an alternative *in silico* STEC typing method.

We were unable to validate 43 of the 229 minor cluster-specific gene marker sets as these minor clusters had few isolates and once isolates were included in the identification dataset, no isolates remained for validation. Therefore, markers for these 43 minor clusters are tentative and require future validation when more genomes become available. Genes specific to each of these STEC minor clusters were also based on very small number of genomes and should be used with caution. However, since these minor clusters are rarely isolated, they have relatively little effect on the overall applicability of the cluster-specific gene marker sets to STEC typing.

Culture-independent approaches such as shotgun metagenomic analysis may be used for detection of contaminating STEC directly from food samples or enriched food samples (Leonard et al., 2015; Buytaers et al., 2020). However, it is difficult to determine STEC serotype from food or faecal samples directly as O and H antigen genes cannot uniquely identify a STEC serotype in a mixed sample. We showed that the cluster/serotype-specific gene marker sets of interest were detected in the spiked food samples by STECfinder using shotgun metagenomic sequencing reads from the study of Buytaers et al. (2020). Our cluster or serotype specific genes provide proxy markers to identify these serotypes in original or non-pure culture samples. These gene markers could be adapted for metagenomics based diagnosis and culture independent typing, facilitating rapid identification of known STEC clusters and serotypes.

CONCLUSION

This study analysed 41,101 publicly available genomes of STEC isolates and identified 19 major and 229 minor STEC clusters. Specific gene marker sets for the 19 major and 229 minor clusters were identified and found to be valuable for *in silico* typing. We also identified serotype specific markers for the top 10 non-O157:H7 STEC serotypes. These markers can be used as proxy markers to identify the serotypes. We additionally developed STECFinder, a freely available *in silico* serotyping pipeline incorporating the cluster/serotype specific gene markers to facilitate serotyping of STEC isolates using genome sequences with high specificity and sensitivity. The STECFinder pipeline was tested on published metagenomics samples to determine the serotype of known STECs and the results show that cluster and serotype specific markers have potential for culture independent STEC serotyping.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RL designed the study. XZ, MP, and SK performed the bioinformatic analysis. XZ, MP, and RL analysed the results. XZ drafted the manuscript. MP and RL provided critical revision of the manuscript. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

The authors thank Duncan Smith and Robin Heron from UNSW Research Technology Services for high performance computing assistance. This work was funded in part by a National Health and Medical Research Council project grant (grant number 1129713) and an Australian Research Council Discovery Grant (DP170101917).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcimb.2021.772574/full#supplementary-material

Supplementary Figure S1 | Identification phylogenetic tree. The identification phylogenetic tree was constructed using Quicktree v1.3 as **Figure 3** and was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01 substitutions per site. STEC (Shiga toxin producing *E. coli*) clusters are colored per cluster legend and shown as the ring. The internal branches are colored to represent the bootstrap values per colour legend with

green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. ECOLI is *E. coli*. EIEC is Enteroinvasive *E. coli*. MC is STEC minor clusters.

Supplementary Figure S2 | The representative validation phylogenetic tree.

Supplementary Figure S2-A | The 38,534 STEC (Shiga toxin-producing *E. coli*) validation isolates were divided randomly into fifteen subsets and each subset of the isolates were combined with the identification dataset to construct a phylogenetic tree (tree 1 to 15) using Quicktree v1.3 and visualised using Grapetree's to assign isolates in validation dataset to clusters. The representative tree (tree 1) is shown in detail as an example and all the others are similar. The scalar bar represents 0.02 substitutions per site. Known STEC clusters from identification dataset are coloured. Numbers in square brackets after cluster name are the number of isolates of each identified cluster. Isolates in validation dataset (valdb) are coloured white. An "validation" isolate was assigned to a STEC cluster if the isolate was found within a branch that exclusively contained identification dataset isolates from that cluster and that branch had a bootstrap support value of 80% or greater. ECOLI is *E. coli*. EIEC is Enteroinvasive *E. coli*. MC is STEC minor clusters.

Supplementary Figure S2-B | The same phylogenetic tree as Figure S2-A was visualised using ITOL v5 which allowed bootstrap values to be displayed by colouring the internal nodes. The scalar bar represents 0.01substitutions per site. STEC (Shiga toxin-producing *E. coli*) clusters are coloured per cluster legend and shown as the ring. The internal branches are coloured to represent the bootstrap values per colour legend with green and red indicating the maximum (1) and minimum bootstrap values (0). Each cluster is supported by bootstrap value of 80% or greater. ECOLI is *E. coli*. EIEC is Enteroinvasive *E. coli*. MC is STEC minor clusters.

REFERENCES

Bai, X., Fu, S., Zhang, J., Fan, R., Xu, Y., Sun, H., et al. (2018). Identification and Pathogenomic Analysis of an *Escherichia coli* Strain Producing a Novel Shiga Toxin 2 Subtype. *Sci. Rep.* 8 (1), 6756. doi: 10.1038/s41598-018-25233-x

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., et al. (2012). SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing. *J. Comput. Biol.* 19 (5), 455–477. doi: 10.1089/cmb.2012.0021

Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E., and Clermont, O. (2018). ClermonTyping: An Easy-to-Use and Accurate *in Silico* Method for *Escherichia* Genus Strain Phylotyping. *Microb. Genom* 4 (7). doi: 10.1099/mgen.0.000192

Bélanger, S. D., Boissinot, M., Ménard, C., Picard, F. J., and Bergeron, M. G. (2002). Rapid Detection of Shiga Toxin-Producing Bacteria in Feces by Multiplex PCR With Molecular Beacons on the Smart Cycler. *J. Clin. Microbiol.* 40 (4), 1436–1440. doi: 10.1128/jcm.40.4.1436-1440.2002

Bettelheim, K. A. (2000). Role of non-O157 VTEC. *Symp Ser. Soc. Appl. Microbiol.* 29), 38s–50s. doi: 10.1111/j.1365-2672.2000.tb05331.x

Beutin, L., Strauch, E., and Fischer, I. (1999). Isolation of *Shigella sonnei* Lysogenic for a Bacteriophage Encoding Gene for Production of Shiga Toxin. *Lancet (London England)* 353 (9163), 1498–1498. doi: 10.1016/S0140-6736(99)00961-7

Bielaszewska, M., Köck, R., Friedrich, A. W., von Eiff, C., Zimmerhackl, L. B., Karch, H., et al. (2007). Shiga Toxin-Mediated Hemolytic Uremic Syndrome: Time to Change the Diagnostic Paradigm? *PloS One* 2 (10), e1024. doi: 10.1371/journal.pone.0001024

Bosilevac, J. M., and Koohmaraie, M. (2011). Prevalence and Characterization of non-O157 Shiga Toxin-Producing *Escherichia coli* Isolates From Commercial Ground Beef in the United States. *Appl. Environ. Microbiol.* 77 (6), 2103–2112. doi: 10.1128/aem.02833-10

Brandal, L. T., Tunsjø, H. S., Ranheim, T. E., Løbersli, I., Lange, H., and Wester, A. L. (2015). Shiga Toxin 2a in *Escherichia albertii*. *J. Clin. Microbiol.* 53 (4), 1454–1455. doi: 10.1128/jcm.03378-14

Brian, M. J., Frosolono, M., Murray, B. E., Miranda, A., Lopez, E. L., Gomez, H. F., et al. (1992). Polymerase Chain Reaction for Diagnosis of Enterohemorrhagic *Escherichia coli* Infection and Hemolytic-Uremic Syndrome. *J. Clin. Microbiol.* 30 (7), 1801–1806. doi: 10.1128/jcm.30.7.1801-1806.1992

Brooks, J. T., Sowers, E. G., Wells, J. G., Greene, K. D., Griffin, P. M., Hoekstra, R. M., et al. (2005). Non-O157 Shiga Toxin-Producing *Escherichia coli* Infections in the United State-2002. *J. Infect. Dis.* 192 (8), 1422–1429. doi: 10.1086/466536

Bryan, A., Youngster, I., and McAdam, A. J. (2015). Shiga Toxin Producing *Escherichia coli*. *Clin. Lab. Med.* 35 (2), 247–272. doi: 10.1016/j.cll.2015.02.004

Buvens, G., De Gheldre, Y., Dediste, A., de Moreau, A. I., Mascart, G., Simon, A., et al. (2012). Incidence and Virulence Determinants of Verocytotoxin-Producing *Escherichia coli* Infections in the Brussels-Capital Region, Belgium, in 2008-2010. *J. Clin. Microbiol.* 50 (4), 1336–1345. doi: 10.1128/jcm.05317-11

Buytaers, F. E., Saltykova, A., Denayer, S., Verhaegen, B., Vanneste, K., Roosens, N. H. C., et al. (2020). A Practical Method to Implement Strain-Level Metagenomics-Based Foodborne Outbreak Investigation and Source Tracking in Routine. *Microorganisms* 8 (8), 1191. doi: 10.3390/microorganisms8081191

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. (2009). BLAST+: Architecture and Applications. *BMC Bioinf.* 10:421. doi: 10.1186/1471-2105-10-421

Clausen, P., Aarestrup, F. M., and Lund, O. (2018). Rapid and Precise Alignment of Raw Reads Against Redundant Databases With KMA. *BMC Bioinf.* 19 (1), 307. doi: 10.1186/s12859-018-2336-6

DebRoy, C., Roberts, E., Kundrat, J., Davis, M. A., Briggs, C. E., and Fratamico, P. M. (2004). Detection of *Escherichia coli* Serogroups O26 and O113 by PCR Amplification of the *wzx* and *wzy* Genes. *Appl. Environ. Microbiol.* 70 (3), 1830–1832. doi: 10.1128/aem.70.3.1830-1832.2004

DebRoy, C., Roberts, E., Valadez, A. M., Dudley, E. G., and Cutter, C. N. (2011). Detection of Shiga Toxin-Producing *Escherichia coli* O26, O45, O103, O111, O113, O121, O145, and O157 Serogroups by Multiplex Polymerase Chain Reaction of the *wzx* Gene of the O-Antigen Gene Cluster. *Foodborne Pathog. Dis.* 8 (5), 651–652. doi: 10.1089/fpd.2010.0769

European Food Safety Authority, E.C.f.D.P.C (2011). The European Union Summary Report on Trends and Sources of Zoonoses, Zoonotic Agents and Food-Borne Outbreaks in 2009. *EFSA Journal: Eur. Food Standards Agency* 9 (3), 2090. doi: 10.2903/j.efsa.2011.2090

Feng, P. C., and Reddy, S. (2013). Prevalences of Shiga Toxin Subtypes and Selected Other Virulence Factors Among Shiga-Toxigenic *Escherichia coli* Strains Isolated From Fresh Produce. *Appl. Environ. Microbiol.* 79 (22), 6917–6923. doi: 10.1128/aem.02455-13

Ferdous, M., Zhou, K., Mellmann, A., Morabito, S., Croughs, P. D., de Boer, R. F., et al. (2015). Is Shiga Toxin-Negative *Escherichia coli* O157:H7 Enteropathogenic or Enterohemorrhagic *Escherichia coli*? Comprehensive Molecular Analysis Using Whole-Genome Sequencing. *J. Clin. Microbiol.* 53 (11), 3530–3538. doi: 10.1128/jcm.01899-15

Frank, C., Faber, M. S., Askar, M., Bernard, H., Fruth, A., Gilsdorf, A., et al. (2011a). Large and Ongoing Outbreak of Haemolytic Uraemic Syndrome, Germany, May 2011. *Euro Surveill* 16 (21), 19878. doi: 10.2807/ese.16.21.19878-en

Frank, C., Werber, D., Cramer, J. P., Askar, M., Faber, M., an der Heiden, M., et al. (2011b). Epidemic Profile of Shiga-Toxin-Producing *Escherichia coli* O104:H4 Outbreak in Germany. *N Engl. J. Med.* 365 (19), 1771–1780. doi: 10.1056/NEJMoa1106483

Gati, N. S., Middendorf-Bauchart, B., Bletz, S., Dobrindt, U., and Mellmann, A. (2019). Origin and Evolution of Hybrid Shiga Toxin-Producing and Uropathogenic *Escherichia coli* Strains of Sequence Type 141. *J. Clin. Microbiol.* 58 (1), e01309–19. doi: 10.1128/jcm.01309-19

Gerner-Smidt, P., Hise, K., Kincaid, J., Hunter, S., Rolando, S., Hyytiä-Trees, E., et al. (2006). PulseNet USA: A Five-Year Update. *Foodborne Pathog. Dis.* 3 (1), 9–19. doi: 10.1089/fpd.2006.3.9

Gonzales, T. K., Kulow, M., Park, D. J., Kaspar, C. W., Anklam, K. S., Pertzborn, K. M., et al. (2011). A High-Throughput Open-Array qPCR Gene Panel to Identify, Virulotype, and Subtype O157 and Non-O157 Enterohemorrhagic *Escherichia coli*. *Mol. Cell Probes* 25 (5-6), 222–230. doi: 10.1016/j.mcp.2011.08.004

González-Escalona, N., and Kase, J. A. (2019). Virulence Gene Profiles and Phylogeny of Shiga Toxin-Positive *Escherichia coli* Strains Isolated From FDA Regulated Foods During 2010-2017. *PloS One* 14 (4), e0214620. doi: 10.1371/journal.pone.0214620

Gould, L. H., Demma, L., Jones, T. F., Hurd, S., Vugia, D. J., Smith, K., et al. (2009). Hemolytic Uremic Syndrome and Death in Persons With *Escherichia coli* O157:H7 Infection, Foodborne Diseases Active Surveillance Network Site-2006. *Clin. Infect. Dis.* 49 (10), 1480–1485. doi: 10.1086/644621

Gray, M. D., Lampel, K. A., Strockbine, N. A., Fernandez, R. E., Melton-Celsa, A. R., and Maurelli, A. T. (2014). Clinical Isolates of Shiga Toxin 1a-Producing *Shigella flexneri* With an Epidemiological Link to Recent Travel to Hispañiola. *Emerg. Infect. Dis.* 20 (10), 1669–1677. doi: 10.3201/eid2010.140292

GROUP, F.W.S.E (2019). Hazard Identification and Characterization: Criteria for Categorizing Shiga Toxin-Producing *Escherichia coli* on a Risk Basis(†). *J. Food Protectionv* 82 (1), 7–21. doi: 10.4315/0362-028x.Jfp-18-291

Gupta, S. K., Strockbine, N., Omondi, M., Hise, K., Fair, M. A., and Mintz, E. (2007). Emergence of Shiga Toxin 1 Genes Within *Shigella dysenteriae* Type 4 Isolates From Travelers Returning From the Island of Hispañola. *Am. J. Trop. Med. Hyg* 76 (6), 1163–1165. doi: 10.4269/ajtmh.2007.76.1163

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics* 29 (8), 1072–1075. doi: 10.1093/bioinformatics/btt086

Gyles, C. L. (2007). Shiga Toxin-Producing *Escherichia coli*: An Overview. *J. Anim. Sci.* 85 (13 Suppl), E45–E62. doi: 10.2527/jas.2006-508

Hara-Kudo, Y., Nemoto, J., Ohtsuka, K., Segawa, Y., Takatori, K., Kojima, T., et al. (2007). Sensitive and Rapid Detection of Vero Toxin-Producing *Escherichia coli* Using Loop-Mediated Isothermal Amplification. *J. Med. Microbiol.* 56 (Pt 3), 398–406. doi: 10.1099/jmm.0.46819-0

Hedican, E. B., Medus, C., Besser, J. M., Juni, B. A., Koziol, B., Taylor, C., et al. (2009). Characteristics of O157 Versus Non-O157 Shiga Toxin-Producing *Escherichia coli* Infections in Minnesot-2006. *Clin. Infect. Dis.* 49 (3), 358–364. doi: 10.1086/600302

Hu, D., Liu, B., Wang, L., and Reeves, P. R. (2020). Living Trees: High-Quality Reproducible and Reusable Construction of Bacterial Phylogenetic Trees. *Mol. Biol. Evol.* 37 (2), 563–575. doi: 10.1093/molbev/msz241

Iguchi, A., Iyoda, S., Seto, K., Morita-Ishihara, T., Scheutz, F., and Ohnishi, M. (2015). *Escherichia coli* O-Genotyping PCR: A Comprehensive and Practical Platform for Molecular O Serogrouping. *J. Clin. Microbiol.* 53 (8), 2427–2432. doi: 10.1128/jcm.00321-15

Inouye, M., Dashnow, H., Raven, L. A., Schultz, M. B., Pope, B. J., Tomita, T., et al. (2014). SRST2: Rapid Genomic Surveillance for Public Health and Hospital Microbiology Labs. *Genome Med.* 6 (11):90. doi: 10.1186/s13073-014-0090-6

Joensen, K. G., Scheutz, F., Lund, O., Hasman, H., Kaas, R. S., Nielsen, E. M., et al. (2014). Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic *Escherichia coli*. *J. Clin. Microbiol.* 52 (5), 1501–1510. doi: 10.1128/jcm.03617-13

Joensen, K. G., Tetzschner, A. M., Iguchi, A., Aarestrup, F. M., and Scheutz, F. (2015). Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. *J. Clin. Microbiol.* 53 (8), 2410–2426. doi: 10.1128/jcm.00008-15

Johnson, R. P., Clarke, R. C., Wilson, J. B., Read, S. C., Rahn, K., Renwick, S. A., et al. (1996). Growing Concerns and Recent Outbreaks Involving Non-O157: H7 Serotypes of Verotoxigenic *Escherichia coli*. *J. Food Prot* 59 (10), 1112–1122. doi: 10.4315/0362-028x-59.10.1112

Johnson, K. E., Thorpe, C. M., and Sears, C. L. (2006). The Emerging Clinical Importance of non-O157 Shiga Toxin-Producing *Escherichia coli*. *Clin. Infect. Dis.* 43 (12), 1587–1595. doi: 10.1086/509573

Jolley, K. A., and Maiden, M. C. (2010). BIGSdb: Scalable Analysis of Bacterial Genome Variation at the Population Level. *BMC Bioinf.* 11, 595. doi: 10.1186/1471-2105-11-595

Ju, W., Cao, G., Rump, L., Strain, E., Luo, Y., Timme, R., et al. (2012). Phylogenetic Analysis of non-O157 Shiga Toxin-Producing *Escherichia coli* Strains by Whole-Genome Sequencing. *J. Clin. Microbiol.* 50 (12), 4123–4127. doi: 10.1128/jcm.02262-12

Kaper, J. B., Nataro, J. P., and Mobley, H. L. (2004). Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* 2 (2), 123–140. doi: 10.1038/nrmicro818

Käppeli, U., Hächler, H., Giezendanner, N., Beutin, L., and Stephan, R. (2011). Human Infections With non-O157 Shiga Toxin-Producing *Escherichia coli*, Switzerlan-2009. *Emerg. Infect. Dis.* 17 (2), 180–185. doi: 10.3201/eid1702.100909

Krüger, A., and Lucchesi, P. M. (2015). Shiga Toxins and Stx Phages: Highly Diverse Entities. *Microbiol. (Reading)* 161 (Pt 3), 451–462. doi: 10.1099/mic.0.000003

Lacher, D. W., Gangiredla, J., Patel, I., Elkins, C. A., and Feng, P. C. (2016). Use of the *Escherichia coli* Identification Microarray for Characterizing the Health Risks of Shiga Toxin-Producing *Escherichia coli* Isolated From Foods. *J. Food Prot* 79 (10), 1656–1662. doi: 10.4315/0362-028x.Jfp-16-176

Lentz, E. K., Leyva-Illades, D., Lee, M. S., Cherla, R. P., and Tesh, V. L. (2011). Differential Response of the Human Renal Proximal Tubular Epithelial Cell Line HK-2 to Shiga Toxin Types 1 and 2. *Infection Immun.* 79 (9), 3527–3540. doi: 10.1128/iai.05139-11

Leonard, S. R., Mammel, M. K., Lacher, D. W., and Elkins, C. A. (2015). Application of Metagenomic Sequencing to Food Safety: Detection of Shiga Toxin-Producing *Escherichia coli* on Fresh Bagged Spinach. *Appl. Environ. Microbiol.* 81 (23), 8183–8191. doi: 10.1128/aem.02601-15

Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) V4: Recent Updates and New Developments. *Nucleic Acids Res.* 47 (W1), W256–w259. doi: 10.1093/nar/gkz239

Li, B., Liu, H., and Wang, W. (2017). Multiplex Real-Time PCR Assay for Detection of *Escherichia coli* O157:H7 and Screening for Non-O157 Shiga Toxin-Producing *E. coli*. *BMC Microbiol.* 17 (1), 215. doi: 10.1186/s12866-017-1123-2

Lin, A., Nguyen, L., Lee, T., Clotilde, L. M., Kase, J. A., Son, I., et al. (2011). Rapid O Serogroup Identification of the Ten Most Clinically Relevant STECs by Luminex Microbead-Based Suspension Array. *J. Microbiol. Methods* 87 (1), 105–110. doi: 10.1016/j.mimet.2011.07.019

Liptáková, A., Siegfried, L., Kmetová, M., Birosová, E., Kotulová, D., Bencátová, A., et al. (2005). Hemolytic Uremic Syndrome Caused by Verotoxin-Producing *Escherichia coli* O26. Case Report. *Folia Microbiol. (Praha)* 50 (2), 95–98. doi: 10.1007/bf02931454

Liu, B., Knirel, Y. A., Feng, L., Perepelov, A. V., Senchenkova, S. N., Wang, Q., et al. (2008). Structure and Genetics of *Shigella* O Antigens. *FEMS Microbiol. Rev.* 32 (4), 627–653. doi: 10.1111/j.1574-6976.2008.00114.x

Lozer, D. M., Souza, T. B., Monfardini, M. V., Vicentini, F., Kitagawa, S. S., Scaletsky, I. C., et al. (2013). Genotypic and Phenotypic Analysis of Diarrheagenic *Escherichia coli* Strains Isolated From Brazilian Children Living in Low Socioeconomic Level Communities. *BMC Infect. Dis.* 13:418. doi: 10.1186/1471-2334-13-418

Ludwig, J. B., Shi, X., Shridhar, P. B., Roberts, E. L., DebRoy, C., Phebus, R. K., et al. (2020). Multiplex PCR Assays for the Detection of One Hundred and Thirty Seven Serogroups of Shiga Toxin-Producing *Escherichia coli* Associated With Cattle. *Front. Cell Infect. Microbiol.* 10, 378. doi: 10.3389/fcimb.2020.00378

Majowicz, S. E., Scallan, E., Jones-Bitton, A., Sargeant, J. M., Stapleton, J., Angulo, F. J., et al. (2014). Global Incidence of Human Shiga Toxin-Producing *Escherichia coli* Infections and Deaths: A Systematic Review and Knowledge Synthesis. *Foodborne Pathog. Dis.* 11 (6), 447–455. doi: 10.1089/fpd.2013.1704

McCarthy, T. A., Barrett, N. L., Hadler, J. L., Salsbury, B., Howard, R. T., Dingman, D. W., et al. (2001). Hemolytic-Uremic Syndrome and *Escherichia coli* O121 at a Lake in Connecticu. *Pediatrics* 108 (4), E59. doi: 10.1542/peds.108.4.e59

McDaniel, T. K., and Kaper, J. B. (1997). A Cloned Pathogenicity Island From Enteropathogenic *Escherichia coli* Confers the Attaching and Effacing Phenotype on E. Coli K-12. *Mol. Microbiol.* 23 (2), 399–407. doi: 10.1046/j.1365-2958.1997.2311591.x

Melton-Celsa, A. R. (2014). Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiol. Spectr.* 2 (4), Ehec–0024-2013. doi: 10.1128/microbiolspec.EHEC-0024-2013

Milley, D. G., and Sekla, L. H. (1993). An Enzyme-Linked Immunosorbent Assay-Based Isolation Procedure for Verotoxigenic *Escherichia coli*. *Appl. Environ. Microbiol.* 59 (12), 4223–4229. doi: 10.1128/aem.59.12.4223-4229.1993

Mizutani, S., Nakazono, N., and Sugino, Y. (1999). The So-Called Chromosomal Verotoxin Genes are Actually Carried by Defective Prophages. *DNA Res.* 6 (2), 141–143. doi: 10.1093/dnares/6.2.141

Mora, A., Herrrera, A., López, C., Dahbi, G., Mamani, R., Pita, J. M., et al. (2011). Characteristics of the Shiga-Toxin-Producing Enteroaggregative *Escherichia coli* O104:H4 German Outbreak Strain and of STEC Strains Isolated in Spain. *Int. Microbiol.* 14 (3), 121–141. doi: 10.2436/20.1501.01.142

Mora, A., López, C., Dhabi, G., López-Beceiro, A. M., Fidalgo, L. E., Díaz, E. A., et al. (2012). Seropathotypes, Phylogroups, Stx Subtypes, and Intimin Types of Wildlife-Carried, Shiga Toxin-Producing *Escherichia coli* Strains With the Same Characteristics as Human-Pathogenic Isolates. *Appl. Environ. Microbiol.* 78 (8), 2578–2585. doi: 10.1128/aem.07520-11

Morton, V., Cheng, J. M., Sharma, D., and Kearney, A. (2017). Notes From the Field: An Outbreak of Shiga Toxin-Producing *Escherichia coli* O121 Infections Associated With Flour - Canad-2017. *MMWR Morb Mortal Wkly Rep*. 66 (26), 705–706. doi: 10.15585/mmwr.mm6626a6

Murakami, K., Etoh, Y., Tanaka, E., Ichihara, S., Horikawa, K., Kawano, K., et al. (2014). Shiga Toxin 2f-Producing *Escherichia albertii* From a Symptomatic Human. *Jpn J. Infect. Dis.* 67 (3), 204–208. doi: 10.7883/yoken.67.204

Nataro, J. P., and Kaper, J. B. (1998). Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* 11 (1), 142–201. doi: 10.1128/CMR.11.1.142

Navarro-Garcia, F. (2014). *Escherichia coli* O104:H4 Pathogenesis: An Enteroaggregative *E. coli*/Shiga Toxin-Producing *E. coli* Explosive Cocktail of High Virulence. *Microbiol. Spectr.* 2 (6), 2–6. doi: 10.1128/microbiolspec.EHEC-0008-2013

Needleman, S. B., and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* 48 (3), 443–453. doi: 10.1016/0022-2836(70)90057-4

Norman, K. N., Strockbine, N. A., and Bono, J. L. (2012). Association of Nucleotide Polymorphisms Within the O-Antigen Gene Cluster of *Escherichia coli* O26, O45, O103, O111, O121, and O145 With Serogroups and Genetic Subtypes. *Appl. Environ. Microbiol.* 78 (18), 6689–6703. doi: 10.1128/aem.01259-12

O'Brien, A. D., Marques, L. R., Kerry, C. F., Newland, J. W., and Holmes, R. K. (1989). Shiga-Like Toxin Converting Phage of Enterohemorrhagic *Escherichia coli* Strain 933. *Microb. Pathog.* 6 (5), 381–390. doi: 10.1016/0882-4010(89)90080-6

Ooka, T., Seto, K., Kawano, K., Kobayashi, H., Etoh, Y., Ichihara, S., et al. (2012). Clinical Significance of *Escherichia albertii*. *Emerg. Infect. Dis.* 18 (3), 488–492. doi: 10.3201/eid1803.111401

Paciorek, J. (2002). Virulence Properties of *Escherichia coli* Faecal Strains Isolated in Poland From Healthy Children and Strains Belonging to Serogroups O18, O26, O44, O86, O126 and O127 Isolated From Children With Diarrhoea. *J. Med. Microbiol.* 51 (7), 548–571. doi: 10.1099/0022-1317-51-7-548

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., et al. (2015). Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis. *Bioinformatics* 31 (22), 3691–3693. doi: 10.1093/bioinformatics/btv421

Parsons, B. D., Zelyas, N., Berenger, B. M., and Chui, L. (2016). Detection, Characterization, and Typing of Shiga Toxin-Producing *Escherichia coli*. *Front. Microbiol.* 7, 478. doi: 10.3389/fmicb.2016.00478

Paton, J. C., and Paton, A. W. (1998). Pathogenesis and Diagnosis of Shiga Toxin-Producing *Escherichia coli* Infections. *Clin. Microbiol. Rev.* 11 (3), 450–479. doi: 10.1128/CMR.11.3.450

Paton, A. W., Woodrow, M. C., Doyle, R. M., Lanser, J. A., and Paton, J. C. (1999). Molecular Characterization of a Shiga Toxigenic *Escherichia coli* O113:H21 Strain Lacking Eae Responsible for a Cluster of Cases of Hemolytic-Uremic Syndrome. *J. Clin. Microbiol.* 37 (10), 3357–3361. doi: 10.1128/jcm.37.10.3357-3361.1999

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. (2017). Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression. *Nat. Methods* 14 (4), 417–419. doi: 10.1038/nmeth.4197

Qin, X., Klein, E. J., Galanakis, E., Thomas, A. A., Stapp, J. R., Rich, S., et al. (2015). Real-Time PCR Assay for Detection and Differentiation of Shiga Toxin-Producing *Escherichia coli* From Clinical Samples. *J. Clin. Microbiol.* 53 (7), 2148–2153. doi: 10.1128/jcm.00115-15

Scheutz, F., Teel, L. D., Beutin, L., Piérard, D., Buvens, G., Karch, H., et al. (2012). Multicenter Evaluation of a Sequence-Based Protocol for Subtyping Shiga Toxins and Standardizing Stx Nomenclature. *J. Clin. Microbiol.* 50 (9), 2951–2963. doi: 10.1128/jcm.00860-12

Seemann, T. (2014). Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* 30 (14), 2068–2069. doi: 10.1093/bioinformatics/btu153

Smith, J. L., Fratamico, P. M., and Gunther, N. (2014). Shiga Toxin-Producing *Escherichia coli*. *Adv. Appl. Microbiol.* 86, 145–197. doi: 10.1016/b978-0-12-800262-9.00003-2

Stigi, K. A., Macdonald, J. K., Tellez-Marfin, A. A., and Lofy, K. H. (2012). Laboratory Practices and Incidence of non-O157 Shiga Toxin-Producing *Escherichia coli* Infections. *Emerg. Infect. Dis.* 18 (3), 477–479. doi: 10.3201/eid1803.111358

Tarr, P. I., Gordon, C. A., and Chandler, W. L. (2005). Shiga-Toxin-Producing *Escherichia coli* and Haemolytic Uraemic Syndrome. *Lancet* 365 (9464), 1073–1086. doi: 10.1016/s0140-6736(05)71144-2

Teel, L. D., Daly, J. A., Jerris, R. C., Maul, D., Svanas, G., O'Brien, A. D., et al. (2007). Rapid Detection of Shiga Toxin-Producing *Escherichia coli* by Optical Immunoassay. *J. Clin. Microbiol.* 45 (10), 3377–3380. doi: 10.1128/jcm.00837-07

Teunis, P. F., Ogden, I. D., and Strachan, N. J. (2008). Hierarchical Dose Response of *E. coli* O157:H7 From Human Outbreaks Incorporating Heterogeneity in Exposure. *Epidemiol. Infect.* 136 (6), 761–770. doi: 10.1017/s0950268807008771

Tuttle, J., Gomez, T., Doyle, M. P., Wells, J. G., Zhao, T., Tauxe, R. V., et al. (1999). Lessons From a Large Outbreak of *Escherichia coli* O157:H7 Infections: Insights Into the Infectious Dose and Method of Widespread Contamination of Hamburger Patties. *Epidemiol. Infect.* 122 (2), 185–192. doi: 10.1017/s0950268898001976

Valilis, E., Ramsey, A., Sidiq, S., and DuPont, H. L. (2018). Non-O157 Shiga Toxin-Producing *Escherichia coli*-A Poorly Appreciated Enteric Pathogen: Systematic Review. *Int. J. Infect. Dis.* 76, 82–87. doi: 10.1016/j.ijid.2018.09.002

Verstraete, K., De Reu, K., Van Weyenberg, S., Piérard, D., De Zutter, L., Herman, L., et al. (2013). Genetic Characteristics of Shiga Toxin-Producing *E. coli* O157, O26, O103, O111 and O145 Isolates From Humans, Food, and Cattle in Belgium. *Epidemiol. Infect.* 141 (12), 2503–2515. doi: 10.1017/s0950268813000307.

Wood, D. E., and Salzberg, S. L. (2014). Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments. *Genome Biol.* 15 (3), R46. doi: 10.1186/gb-2014-15-3-r46

World Health Organization (2019). *Shiga Toxin-Producing Escherichia coli (STEC) and Food: Attribution Characterization and Monitoring*. (Rome: World Health Organization).

Yang, X., Bai, X., Zhang, J., Sun, H., Fu, S., Fan, R., et al. (2020). *Escherichia coli* Strains Producing a Novel Shiga Toxin 2 Subtype Circulate in China. *Int. J. Med. Microbiol.* 310 (1):151377. doi: 10.1016/j.ijmm.2019.151377

Zhang, W., Bielaszewska, M., Bauwens, A., Fruth, A., Mellmann, A., and Karch, H. (2012). Real-Time Multiplex PCR for Detecting Shiga Toxin 2-Producing *Escherichia coli* O104:H4 in Human Stools. *J. Clin. Microbiol.* 50 (5), 1752–1754. doi: 10.1128/jcm.06817-11

Zhang, Y., Liao, Y. T., Sun, X., and Wu, V. C. H. (2020). Is Shiga Toxin-Producing *Escherichia coli* O45 No Longer a Food Safety Threat? The Danger is Still Out There. *Microorganisms* 8 (5). doi: 10.3390/microorganisms8050782

Zhang, W., Mellmann, A., Sonntag, A. K., Wieler, L., Bielaszewska, M., Tschäpe, H., et al. (2007). Structural and Functional Differences Between Disease-Associated Genes of Enterohaemorrhagic *Escherichia coli* O111. *Int. J. Med. Microbiol.* 297 (1), 17–26. doi: 10.1016/j.ijmm.2006.10.004

Zhang, X., Payne, M., and Lan, R. (2019). *In Silico* Identification of Serovar-Specific Genes for *Salmonella* Serotyping. *Front. Microbiol.* 10:835. doi: 10.3389/fmicb.2019.00835

Zhang, X., Payne, M., Nguyen, T., Kaur, S., and Lan, R. (2021). Cluster-Specific Gene Markers Enhance *Shigella* and Enteroinvasive *Escherichia coli In Silico* Serotyping. *Microb. Genom.* 7 (12). doi: 10.1099/mgen.0.000704. 2001.2030.428723

Zhou, Z., Alikhan, N. F., Mohamed, K., Fan, Y., and Achtman, M. (2020). The EnteroBase User's Guide, With Case Studies on *Salmonella* Transmissions, Yersinia Pestis Phylogeny, and *Escherichia* Core Genomic Diversity. *Genome Res.* 30 (1), 138–152. doi: 10.1101/gr.251678.119

Zhou, Z., Alikhan, N. F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., et al. (2018). GrapeTree: Visualization of Core Genomic Relationships Among 100,000 Bacterial Pathogens. *Genome Res.* 28 (9), 1395–1404. doi: 10.1101/gr.232397.117

Zweifel, C., Cernela, N., and Stephan, R. (2013). Detection of the Emerging Shiga Toxin-Producing *Escherichia coli* O26:H11/H- Sequence Type 29 (ST29) Clone in Human Patients and Healthy Cattle in Switzerland. *Appl. Environ. Microbiol.* 79 (17), 5411–5413. doi: 10.1128/aem.01728-13

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.