Molecular Characterization of Shiga Toxin-Producing *Escherichia coli*

**Pallavi Singh and Shannon D. Manning**

*Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA*

**Introduction**

Shiga toxin-producing *Escherichia coli* (STEC) is an important foodborne pathogen that contributes to roughly 265,000 infections in the United States each year; an estimated 97,000 of these infections are caused by STEC O157 while the remainder are caused by STEC of other serotypes (non-O157 STEC) (Scallan *et al.*, 2011). An estimated 68 and 82% of serotype O157 and non-O157 infections, respectively, were suggested to be acquired via foodborne transmission (Scallan *et al.*, 2011). Together, STEC O157 and non-O157 infections account for over 3600 hospitalizations and 31 deaths in the United States each year (Scallan *et al.*, 2011). Although STEC causes a wide range of clinical complications, it is best known for its ability to cause diarrhea, hemorrhagic colitis, hemolytic uremic syndrome (HUS), and even death (Lingwood, 1996). Health-related costs associated with STEC infections in 2005 were estimated to be over $400 million per year (Frenzen *et al.*, 2005), while large losses associated with STEC contamination have also been documented in both the agriculture and food production industries.

STEC and enterohemorrhagic *E. coli* (EHEC) are defined by their ability to produce one or more Shiga toxins, which are encoded by distinct prophages and occur in different combinations. Shiga toxins contribute to hemorrhagic
coli, which may lead to severe kidney damage, kidney failure, and death in some individuals, particularly children under the age of 10 (Karmali et al., 1985). EHEC strains represent a subset of STEC (Donnenberg, 2002) that are also capable of producing attaching and effacing lesions on host epithelial cells via genes present on the locus of enterocyte effacement (LEE) pathogenicity island (McDaniel and Kaper, 1997). Specifically, the LEE possesses a type III secretion system as well as the eae gene encoding the adhesin intimin that mediates intimate attachment to the host epithelia via binding to its translocated receptor, Tir (encoded by tir) (Beutin et al., 1994; Nataro and Kaper, 1998; Kaper, Nataro, and Mobley, 2004). EHEC strains containing the LEE have an enhanced ability to colonize the lower gastrointestinal tract relative to other STEC (Beutin et al., 1994; Nataro and Kaper, 1998; Kaper, Nataro, and Mobley, 2004) and were suggested to cause more severe disease (Ethelberg et al., 2004).

The majority of detection methods in use today involve screening suspect samples for the presence of either the Shiga toxin (Stx) or the stx and eae genes prior to culture. After the bacterium is recovered, multiple methods have been developed to characterize the genetically diverse STEC strain population. Such methods have changed considerably over the past decade due to the increased use and decreased cost of sequencing technology. Most sequence-based methods are superior to the traditional phenotypic techniques (e.g., serotyping) that have been in place since the emergence of STEC in 1982. Nevertheless, it is important to note that serotyping is still a useful typing tool, particularly in epidemiological studies and outbreak investigations as it provides a rapid way to exclude strains from an outbreak. Additional typing tools, however, are required to further differentiate among epidemiologically linked strains. Pulsed-field gel electrophoresis (PFGE), for example, is the basis of the PulseNet system established by the Centers for Disease Control and Prevention (CDC) (Swaminathan et al., 2001), and has been instrumental in the detection and control of STEC outbreaks.

While not always feasible in outbreak investigations, genome sequencing can decrease the time needed to detect outbreaks and expedite source identification, particularly for non-O157 STEC that remain difficult to culture despite improved diagnostic tools. In 2011, for example, several food sources including tomatoes, leaf salad, and cucumbers were mistakenly reported to be the source of the Stx-producing E. coli O104:H4 outbreak of HUS in Germany (Frank et al., 2011a). These food sources, however, were implicated in the early stage of the outbreak based on epidemiological data and not the isolation of E. coli O104:H4 from the food items. Because of the economic losses that followed the elimination of suspect food items and the increasing frequency of HUS cases and deaths, the genome of an E. coli O104:H4 isolate was sequenced and published while the outbreak was still in progress (Li et al., 2011; Rohde et al., 2011). Although the source of the outbreak was still not
known, the genome provided a tool that allowed researchers worldwide to examine the strain in silico. Consequently, it was determined that the O104:H4 outbreak strain is a member of the enteroaggregative E. coli (EAEC) family of diarrheagenic E. coli, which interacts with eukaryotic epithelial cells via a different mechanism than EHEC and other E. coli pathotypes (Figure 12.1). The EAEC O104:H4 outbreak strain also possesses the Stx2-converting bacteriophage common in STEC, an antibiotic resistance plasmid, and the EAEC virulence plasmid (pAA) important for colonization and biofilm formation (Sheikh et al., 2001; Bielaszewska et al., 2011; Al Safadi et al., 2012). Although the rapid release and assembly of genome data contributed to some inaccurate sequence calls, knowledge of the O104:H4 genome facilitated the development of detection methods that targeted molecular characteristics unique to the outbreak strain (Bielaszewska et al., 2011; Cheung et al., 2011; Mora et al., 2011; Qin et al., 2011; Grad et al., 2012). As a result, clinical and public health laboratories were able to more rapidly identify cases and screen potential food sources. These practices ultimately led to the identification of

---

**Figure 12.1** Differences in the type of interactions between diarrheagenic E. coli representing the six pathotypes and eukaryotic epithelial cells. Kaper et al. 2004. Reproduced with permission of Nature Publishing Group.
O104:H4-contaminated fenugreek sprouts and control of the outbreak (Grad et al., 2012) that affected over 3800 people (Frank et al., 2011b).

In this chapter, we describe the various molecular subtyping techniques that are currently used to characterize STEC while highlighting useful applications. These methods are classified into three general categories: DNA fingerprinting, sequence-based genotyping, and virotyping. We have also evaluated each method for advantages and limitations to help the readers determine which methods are best suited for their study goals. Some of the techniques described may also be applicable to other diarrheagenic *E. coli* pathotypes including EAEC, enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), and diffusely adherent *E. coli* (DAEC); however, this chapter will focus solely on the molecular characterization of STEC and EHEC.

**DNA fingerprinting**

Multiple DNA fingerprinting methods, which require no knowledge of the genomic sequence, have been developed to characterize STEC. These methods are extremely useful in outbreak investigations to establish a link between individuals and identify the infection source. The majority of fingerprinting methods in use today rely on polymerase chain reaction (PCR) amplification or restriction digestion of a bacterial genome and subsequent visualization by gel electrophoresis. Many researchers use these methods independently or in combination with other methods to enhance discriminatory power and resolution.

**Pulsed-field gel electrophoresis**

PFGE is likely to be the most widely used fingerprinting technique for STEC and other enteric pathogens that are monitored via PulseNet. The technique is based on the restriction digestion of genomic DNA using a rare cutting restriction enzyme (e.g., *XbaI*) and the subsequent resolution of DNA banding patterns, or fingerprints, by electrophoresis. The main disadvantages of PFGE are that discrimination between strains occasionally requires the use of multiple enzymes and rapid genetic changes can result in distinct banding patterns among epidemiologically linked individuals. PFGE is also time consuming, requires specialized equipment and analytical expertise, and is not readily amenable to automation when compared to most sequence-based genotyping approaches (Table 12.1). Nevertheless, PFGE is extremely useful for STEC surveillance, as the procedures are standardized and fingerprint patterns can be compared across geographic locations. PFGE is also useful for detecting genetic events including point mutations, horizontally acquired elements, changes in insertion sequence copy number, chromosomal insertions, transposition, and deletion events, which rapidly alter bacterial
genomes and result in distinct banding patterns. As a result, PFGE can detect differences between closely related strains, thereby enhancing its utility in outbreak investigations.

Indeed, PFGE is frequently used for STEC typing, source tracking, and outbreak investigations (Izumiya et al., 1997; Radu et al., 2001; Scott et al., 2006; Grant et al., 2008), particularly for STEC serotypes O157:H7. The PulseNet system administered by the CDC, for example, uses PFGE to characterize STEC isolated via surveillance in the United States (Swaminathan et al., 2001). Routine surveillance is imperative since it facilitates the rapid identification of clusters that can trigger large-scale outbreak investigations as well as the identification of unrelated or sporadic cases in a given location (Bender et al., 1997). The development of PulseNet International in 2006 has enhanced our ability to identify STEC outbreaks worldwide (Swaminathan et al., 2006) and was instrumental in the rapid detection and confirmation of E. coli O104:H4 cases in the United States. PFGE is commonly performed in conjunction with phenotypic methods including serotyping and antibiotic susceptibility testing, which are straightforward and inexpensive laboratory tests that provide additional data to use during outbreak investigations (Verstraete et al., 2012).

**Amplified fragment length polymorphism**

Amplified fragment length polymorphism (AFLP) is similar to PFGE in that it relies on the digestion of genomic DNA with a combination of rare and frequent cutter enzymes. Contrary to PFGE, adaptor oligonucleotides are
ligated to the DNA fragments, which are amplified by PCR using an unlabeled primer set followed by a radiolabeled primer set. The resulting PCR products are resolved by electrophoresis and visualized via autoradiography. The utilization of different restriction enzymes and primer combinations can generate numerous AFLP fingerprints without any knowledge of the genomic sequence, which is considered a key advantage. Other benefits include the detection of single base mutations, and the ability to determine the genetic basis for fragment differences by excising specific bands from gels, cloning and sequencing the DNA fragments, or performing hybridizations (Hu, Lan, and Reeves, 2002). Similarly, fluorescent AFLP (FAFLP), which uses fluorescent dye-labeled primers instead of radiolabeled primers, has been found to be reproducible, easily standardized across laboratories, and amenable to high throughput analyses using an automated sequencer. In one FAFLP study with *E. coli* O157:H7, digestion with two restriction endonucleases, *Eco*RI and *Msp*I, followed by selective amplification was more discriminating than PFGE (Zhao *et al.*, 2000); however, this finding was not observed in another study (Grif *et al.*, 1998).

**Multilocus variable-number tandem-repeat analysis**

Multilocus variable-number tandem-repeat analysis (MLVA) is a fast and portable method that detects polymorphisms in a subset of genes with a variable number of tandem repeats (VNTRs). These VNTRs are located throughout the genome and can be amplified by PCR targeting conserved regions adjacent to variable regions of a given gene. MLVA was demonstrated to be a useful typing tool for STEC O157 in that it provided higher resolution and was less laborious to perform relative to other fingerprinting techniques (Lindstedt *et al.*, 2003). Hence, it was adopted as an important tool for characterization of STEC O157:H7 due to its ease of use and higher resolution than PFGE (Noller *et al.*, 2003; Kawamori *et al.*, 2008). Additionally, MLVA has been used in epidemiological studies and to characterize non-O157 STEC strains, which could be distinguished from O157 STEC strains via unweighted pair group method with arithmetic mean (UPGMA) clustering analysis (Bustamante *et al.*, 2010). In another study, MLVA was applied to STEC O157 strains recovered from a bovine feedlot to demonstrate that the feedlot, water troughs, pen bars, pen floor, feces, and feed played a role in STEC transmission among cattle (Murphy *et al.*, 2008). Nevertheless, MLVA does have several limitations that are worth noting. Because VNTRs can rapidly evolve, multiple MLVA patterns have been detected in strains originating from a single outbreak (Noller *et al.*, 2003). Mutation rates have also been shown to differ across MLVA sites (Noller *et al.*, 2003; Lindstedt, Vardund, and Kapperud, 2004; Keys, Kemper, and Keim, 2005). Such rapid evolution was suggested to be attributable to
stressed environmental conditions such as elevated temperatures and starvation in one study (Cooley et al., 2010).

**Random amplified polymorphic DNA**

Random amplified polymorphic DNA (RAPD) is a simple and rapid PCR-based molecular typing tool that uses random, short (10-mer) primers that arbitrarily bind to and amplify template DNA. Amplification will result in different banding patterns across genetically distinct strains that can be visualized via gel electrophoresis. RAPD has been demonstrated to be a valuable tool for epidemiological studies of *E. coli* O157 (Grif et al., 1998). In another study of *E. coli* O157 isolates, the PCR fragments generated from RAPD were digested with restriction endonucleases to further enhance discriminatory power (Hopkins and Hilton, 2001). RAPD has also been implemented in the characterization of O157 phages, resulting in marked differentiation with good reproducibility (Dini and De Urraza, 2010). Important disadvantages of RAPD are that intra-laboratory variation is high and the technique allows annealing to and amplification of template with slight mismatches; therefore, sequence differences still result in amplification, which can contribute to misclassification and misinterpretation of band patterns (Hopkins and Hilton, 2001; Kim et al., 2005). Also, previous studies aimed at characterization of *E. coli* O157 isolated from food samples, for example, found RAPD to have less discriminatory power than PFGE and phage-typing (Grif et al., 1998; Radu et al., 2001).

**Repetitive PCR**

Repetitive PCR (rep-PCR) genomic fingerprinting technique is based on highly conserved, repetitive DNA sequences found throughout the *E. coli* chromosome (Lupski and Weinstock, 1992). Such repetitive sequences include the 35–40 bp repetitive extragenic palindromic (REP) sequences, 124–127 bp enterobacterial repetitive intergenic consensus (ERIC) sequences, and 154 bp BOX element sequence (Lupski and Weinstock, 1992; Koeuth, Versalovic, and Lupski, 1995). REP,ERIC,BOX, and polytrinucleotide (GTG) sequences are highly conserved, and primers specific to these regions have been designed to amplify these sequences. The resulting PCR products vary in length and can be visualized by gel electrophoresis. Advantages include affordability, accuracy, simultaneous processing of multiple strains, and reproducibility (Carson et al., 2003; Joerger and Ross, 2005). STEC has been characterized using this method in various studies. Gilbreath et al., for example, employed rep-PCR to demonstrate that interaction between wild ungulates and cattle may result in higher STEC colonization of cattle (Gilbreath et al., 2009). Rep-PCR has also been used for source tracking of pathogenic *E. coli* in a number
of studies, demonstrating that over 78% of the *E. coli* isolates could be tracked accurately to their original sources including chickens, turkeys, wild birds, cattle, and humans (Dombek *et al.*, 2000; Mohapatra, Broersma, and Mazumder, 2007). Rep-PCR has also proven useful as an epidemiological tool. For example, a study aimed at characterizing uropathogenic *E. coli* strains from patients in India and Germany classified strains based on their geographical locations (Jadhav *et al.*, 2011). Another study demonstrated rep-PCR as a complementary tool to RAPD for O157 phage typing (Dini and De Urraza, 2010). Relative to other typing tools, however, rep-PCR has considerably lower discriminatory power with high intra-laboratory variation (Johnson and O’Bryan, 2000).

**Sequence-based genotyping**

Contrary to DNA fingerprinting, sequence-based molecular methods require prior knowledge of the genomic sequence or involve genomic sequencing; several of these methods are useful for STEC characterization. In particular, sequence-based methods are useful in epidemiological and evolutionary studies, as sequencing data are amenable to phylogenetic analyses. Such methods have been used to estimate the genetic diversity and evolutionary relatedness of the STEC strain population and have enhanced our understanding of STEC emergence. In this section, we have highlighted methods that examine short genome sequences as well as those that rely on whole genome sequence data.

**Octamer-based genome scanning**

Octamer-based genome scanning (OBGS) is a PCR-based genomic fingerprinting method that was developed to examine genomic diversity and epidemiological relationships in STEC O157 from humans and bovines (Kim, Nietfeldt, and Benson, 1999). In *E. coli*, octamers are over-represented oligomers (e.g., gagcaggg) on the leading strand that are located near the origin and terminus of replication (Salzberg *et al.*, 1998). Octamer-based fluorescent-labeled PCR primers are used for the leading strand and unlabeled primers complementary to the octamers are used for the lagging strand. Fluorescent-labeled PCR products of different sizes, even for single nucleotide insertions or deletions, can be analyzed with automated sequencers (Kim, Nietfeldt, and Benson, 1999).

The initial OBGS analysis of STEC O157 identified two genetically distinct lineages; lineage I consisted of pathogenic strains from humans and bovine samples and lineage II comprised predominantly bovine and nonpathogenic human-derived isolates (Kim, Nietfeldt, and Benson, 1999). Additional studies utilized the OBGS lineage data to demonstrate that each lineage
SequenCe-BaSed genotyPing comprised strains with distinct antiterminator Q genes on the Stx2 bacteriophage (Lejeune et al., 2004), had differential expression of several LEE genes (McNally et al., 2001), and represented widespread clones identified in other geographic locations (Sharma et al., 2009; Franz et al., 2012). Furthermore, bovine-derived isolates were found to survive better in the presence of acid than human isolates as were lineage II versus lineage I isolates, thereby indicating a relationship between acid resistance and genetic background (Saridakis et al., 2004).

Further analysis of the genotypic and phenotypic differences between the lineages demonstrated the unique physiological and ecological characteristics that influence virulence properties and host association; these unique properties can in turn serve as epidemiological markers (Ziebell et al., 2008). OBGS and lineage-specific polymorphism assay (LSPA) (see the following text), for example, have been used in conjunction with screening for other virulence characteristics such as stx2 variation, to better understand the relationships between virulence, host specificity, and geographical spread (Ziebell et al., 2008).

Further analysis of the genotypic and phenotypic differences between the lineages demonstrated the unique physiological and ecological characteristics that influence virulence properties and host association; these unique properties can in turn serve as epidemiological markers (Ziebell et al., 2008). OBGS and lineage-specific polymorphism assay (LSPA) (see the following text), for example, have been used in conjunction with screening for other virulence characteristics such as stx2 variation, to better understand the relationships between virulence, host specificity, and geographical spread (Ziebell et al., 2008).

**Lineage-specific polymorphism assay**

Developed by Yang et al. (2004), LSPA is a PCR-based high throughput fingerprinting method that can characterize isolates based on polymorphisms in six genetic markers with the allele order folD, Z5935, yhcG, rtcB, rbsB, and arp-iclR. The development of this method required prior knowledge of the gene sequences, as loci were selected for inclusion based on their stability, location, and conservation within a lineage defined by OBGS (Kim, Nietfeldt, and Benson, 1999). LSPA is based on multiplex PCR amplification using a fluorescently labeled forward primer and an unlabeled reverse primer, and the band patterns can be visualized using a denaturing polyacrylamide gel (Yang et al., 2004). Loci with alleles shared among lineage I and I strains were designated with a “1” or “2,” respectively, while novel alleles were designated as allele “3.” Positive control strains are typically run alongside test samples in order to accurately classify the different genotypes. The analysis of banding patterns has been performed using AlphaEaseFC™ or BioNumerics (Applied Maths) software programs, and strains with distinct band sizes relative to the control strains are considered to represent unique alleles for a given locus (Ziebell et al., 2008; Sharma et al., 2009). The LSPA protocol is considered to be a high throughput and reliable method with low inter-laboratory variation (Zhang et al., 2007; Ziebell et al., 2008; Sharma et al., 2009).

Application of LSPA to 1429 human- and bovine-derived STEC O157 strains revealed 41 distinct genotypes, though types 111111, 112111, 211111, 212111, and 222222 accounted for 92% of all strains examined, and a sixth category included all other genotypes (Yang et al., 2004).
which possesses alleles that are specific for lineage I, was associated with human infections, while genotype 212111 was encountered with high frequency among bovine-derived strains (Yang et al., 2004; Ziebell et al., 2008). Evolutionary studies of LSPA polymorphisms suggest that lineage II strains were derived from lineage I ancestral populations (Yang et al., 2004).

**Multilocus sequence typing**

The precursor to multilocus sequence typing (MLST), a technique that characterized STEC based on the electrophoretic mobility of several cellular enzymes (Selander et al., 1986). Because electrophoretic mobility is dependent on the amino acid sequence of each enzyme, variability in mobility patterns can be attributable to variation at the corresponding gene. In general, the combination of electromorphs over the total number of enzymes examined dictates the electrophoretic type of a given bacterial strain, which was used to estimate the genetic diversity of the different enzymes across strains. More specifically, a landmark MLEE study suggested that STEC O157:H7 emerged from an enteropathogenic *E. coli* (EPEC) O55:H7 ancestor via the successive acquisition of mobile DNA virulence elements (Feng et al., 1998). Although MLEE has been replaced by MLST, it was considered an important typing tool that enhanced our knowledge of STEC diversity and evolution; similar conclusions regarding STEC evolution have been made using MLST (Reid et al., 2000).

The MLST scheme was initially developed by Maiden et al. for *Neisseria meningitidis* in 1998 (Maiden et al., 1998) and a scheme specific for *E. coli* (EcMLST) was established soon thereafter (Reid et al., 2000; Qi et al., 2004). EcMLST was developed by sequencing fragments of the following seven housekeeping genes: *aspC* (aspartate aminotransferase), *clpX* (ATP-dependent Clp protease), *fadD* (cyl-CoA synthetase), *icdA* (isocitrate dehydrogenase), *lysP* (lysine-specific permease), *mdh* (malate dehydrogenase), and *uidA* (beta-d-glucuronidase) (Table 12.2). Sequence data generated from a set of 130 *E. coli* strains representing 20 serotypes of the different phylogenetic groups identified via MLEE were used to refine the primers for use with this method. Following sequence alignment, allelic variation in MLST loci can be determined by uploading sequence data into the EcMLST database for allele and sequence type (ST) assignments and for identification of novel alleles and STs. MLST protocols and the EcMLST database are available via the STEC Center at Michigan State University (www.shigatox.net).

The first EcMLST database, which contained query tools for determining allelic profiles of sequences specific for the seven genes, was developed in 2004 and released online on July 14, 2005 (Qi et al., 2004). Additional MLST genes have also been added to the system for a subset of strains so users can access the 7-gene or 15-gene typing scheme. The EcMLST database contains
MLST gene, allelic profile, and strain tables with molecular and epidemiological information on each strain if available. In 2012, there were data available for over 3500 strains for 7 genes and over 150 strains for 15 genes yielding a total of 1003 STs and 129 STs, respectively. Additional Escherichia coli typing schemes are also available that utilize a different set of loci (Wirth et al., 2006; Jaureguy et al., 2008).

Because most of the variation that is revealed by MLST is selectively neutral, it is useful for estimating diversity, identifying evolutionary relationships, and elucidating the history of divergence of the Escherichia coli chromosomal background (Maiden et al., 1998). MLST is also useful for epidemiological studies and examining changes in the global distribution of strains over time. Additional advantages include sensitivity in that it can detect single nucleotide changes in the conserved MLST loci, reproducibility, and ease by which the data can be compared across laboratories. The factors that determine the discriminatory power of MLST include the type number of genes and length of the sequence; however, these factors also contribute to limited resolution. For example, MLST is less useful for determining whether certain strains are part of an outbreak.

### Comparative genomic hybridization

Microarray-based comparative genomic hybridization (CGH) allows investigators to determine the presence or absence of specific genes as compared to a reference genome. In 2000, the technique was applied to a set of pathogenic...
E. coli, including Sakai O157:H7, and a commensal E. coli strain, W3110, to identify open reading frames (ORFs) specific to each (Fukiya et al., 2004). In all, the study defined 96 ORFs to be present in W3110 and absent in all 16 pathogenic strains, enhancing our understanding of the ORF profile of pathogenic versus commensal E. coli.

The use of microarray-based CGH has also provided additional support for the evolution of EHEC O157:H7 from EPEC O55:H7 via the identification of shared genes (Wick et al., 2005). Here, multigenome arrays based on E. coli strains Sakai (Hayashi et al., 2001), K-12 (Blattner et al., 1997), and EDL-933 (Perna et al., 2001) were hybridized with nine strains representing each putative step in the stepwise evolution of O157:H7 from EPEC O55:H7. This study was important for identifying which genes were lost or gained relative to strains Sakai, K-12, and EDL-933 (Wick et al., 2005). Interestingly, 82.6% of 5121 genes were present in all nine strains, whereas 11.3% genes were variably absent or present (VAP) or highly divergent. Among the VAP identified, roughly 85% represented phage-derived segments that were dispersed throughout the Sakai and EDL933 genomes and were suggested to contribute to virulence (Ohnishi, Kurokawa, and Hayashi, 2001; Perna et al., 2001; Wick et al., 2005). A subsequent study applied CGH to identify genetic features unique to LSPA lineages I and II relative to strains Sakai, K-12, and EDL-933 (Zhang et al., 2007). This study identified a third intermediate lineage (I/II) with LSPA genotype 211111 and detected differences in the distribution of virulence genes across lineages (Zhang et al., 2007). To further examine VAP regions, a follow-up study utilizing PCR-based assays conducted to subtype 34 VAP regions identified in the STEC O157 CGH analyses, thereby enhancing the discriminatory power of the technique (Laing et al., 2008).

Another comparative hybridization method, suppression subtractive hybridization (SSH), has also been developed to identify and detect genes that are unique to a given strain representing a specific genotype. SSH involves the hybridization of a control strain against a tester strain; the unhybridized DNA fragments can be amplified by PCR, cloned, and sequenced. SSH has been applied to STEC O157 strains representing LSPA lineages I and II to identify putative virulence genes that are more common in lineage I, comprised predominantly of human-derived strains. It was found that 12 conserved regions, which contained genes encoding a hemolysin-activating protein and an iron transport system, were present in most lineage I strains, but were absent in most lineage II strains (Steele et al., 2007, 2009).

In summary, CGH represents a robust, high resolution molecular typing technique for STEC O157 that was suggested to have higher discriminatory power than PFGE (Zhang et al., 2007; Laing et al., 2008). CGH, however, is limited in that the lack of a hybridization signal is difficult to interpret without additional analyses.
Single nucleotide polymorphism typing

Single nucleotide polymorphism (SNP) genotyping represents a high-throughput technique to characterize strains based on the distribution of SNPs located in a subset of genes. Typically, most of the genes selected for genotyping are highly conserved to more accurately reflect phylogenetic relationships. Several SNP genotyping schemes have been developed for STEC O157 and utilize qPCR, mass spectrophotometry, or pyrosequencing. The development of such schemes can be tedious because of the time and effort required to create and test the primer sets; however, the screening of large strain collection is rapid after the method has been developed. Some companies have developed SNP genotyping kits that can be designed for 48 or more strains. The Illumina VeraCode technology, for example, relies on the BeadXpress Reader, a high-throughput, dual-color laser detection system that can scan different multiplexed assays, thereby making the assays faster to run and more cost-effective.

One STEC O157 SNP genotyping scheme utilized hairpin-shaped primers and real-time PCR to detect 96 SNPs located in 83 genes (Manning et al., 2008). This scheme classified over 500 strains into 39 distinct genotypes that clustered into 9 clades. Further analysis demonstrated that a minimum subset of 32 of the 96 SNPs could differentiate the strains into all 39 genotypes. After strains were assigned to SNP genotypes and clades, epidemiological data from 333 patients with STEC O157 infections in Michigan between 2001 and 2006 (Manning et al., 2007) were examined to identify associations between genetic background and clinical symptoms. Multivariate logistic regression analyses demonstrated that younger patients (0–18 years of age) and patients with HUS were significantly more likely to be infected with a clade 8 strain relative to the other clades (Manning et al., 2007).

A second SNP genotyping scheme based on 178 SNP loci was validated by matrix-assisted laser desorption-ionization time-of-flight (MALDI-TOF) and was used to subtype 261 STEC O157 strains of bovine and human origin (Clawson et al., 2009). A total of 42 genotypes were resolved and a minimum subset of 32 polymorphisms could differentiate among the genotypes. Notably, genotypes were separated into clades containing strains that originated from bovines or humans: 34 and 16 genotypes were of cattle and human origin, respectively, while an additional 8 genotypes were observed in strains of both human and bovine origin (Clawson et al., 2009).

In a follow-up study of 25 clinical and environmental STEC O157:H7 strains, 1225 SNPs were examined to better understand the genetic diversity and relatedness of three produce-associated outbreak strains that occurred during the same time period (Eppinger et al., 2011). A total of 356 synonymous, 647 nonsynonymous, and 222 intergenic SNPs were examined; these were randomly distributed throughout the genome and none were located
within mutational hotspots. Comparisons between the 1225-SNP and LSPA, MLVA, and the 96-SNP scheme demonstrated concordance across typing methods, though the 1225-SNP scheme could better differentiate the most closely related lineages (Eppinger et al., 2011).

SNP genotyping represents a useful typing tool for characterizing STEC O157, which cannot be differentiated by certain other sequence-based genotyping methods such as MLST (Noller et al., 2003). Screening for SNPs in both human- and bovine-derived O157 strains has revealed a high level of genetic diversity and has been used to identify important epidemiological associations. It is important to note, however, that screening large strain collections for too few SNPs can result in the misclassification of phylogenetic lineages as inclusion of more SNPs in any given assay will result in a more accurate representation of a genome. Screening for a small subset of SNPs is likely to be more useful in the clinical laboratory setting and could potentially allow health care practitioners to more rapidly assess the virulence potential of an infecting strain (Riordan et al., 2008).

**Whole genome sequencing**

With the advent of next generation sequencing, it has become considerably cheaper and more cost-effective to sequence entire genomes than perform other sequence-based typing methods. Other techniques including PFGE, OBGs, and CGH are also based on the entire genome, yet they do not cover all of it and are not sensitive enough to detect SNPs, thereby limiting the discriminatory power. As a result, many studies of the “pan-genome,” which includes both the core genome that is present in all strains and the dispensable genome that is variably present (Tettelin et al., 2005), have emerged. Pangenomic calculations for pathogenic and commensal *E. coli* species revealed an open pangenome with a reservoir of over 13000 genes that contribute to the organisms’ genomic diversity (Rasko et al., 2008). Compared to a subset of 16 genomes, only 94 genes were unique to the commensal *E. coli* HS strain, while roughly 2200 conserved genes were present in all genomes (Rasko et al., 2008). In a subsequent study, Panseq software was used to predict the pangenome and identify genomic islands specific for STEC O157:H7 and *E. coli* K-12 (Laing et al., 2010).

As the cost of whole genome sequencing continues to go down and more bioinformatics tools become available, whole genome sequencing will be considered the method of choice to examine the biology, evolution, and epidemiology of STEC. While sequencing is extremely useful for emerging infectious disease agents such as *E. coli* O104:H4, it is important to note that the analysis of genome data is not trivial. Indeed, sequence analysis can occasionally result in data bottlenecks and must be considered when an investigator chooses to examine genome data instead of another sequence-based method.
Virotyping

STEC are also routinely characterized based on the presence or absence of specific virulence genes as well as the degree of genetic variation within these genes. By determining the type of Shiga toxin genes (stx) present or the molecular serotype, for example, investigators can rapidly assess whether a given set of strains are similar or different before performing more sensitive genotyping or fingerprinting tests. Examining the distribution and diversity of virulence genes will also provide clues regarding the relative virulence of a given strain and can be used in the early stages of an outbreak investigation. For the most part, virulence gene characterization, or virotyping, of STEC should be performed in conjunction with other typing techniques.

Shiga toxin and virulence gene profiling

The genes encoding the Shiga toxins are well characterized and can provide clues regarding the virulence of a given strain based on previously described associations. For instance, prior studies have observed a correlation between the presence of stx2 and/or stx2c and HUS (Friedrich et al., 2002; Bielaszewska et al., 2006; Persson et al., 2007).

To characterize stx, PCR and PCR-based restriction fragment length polymorphism (RFLP) methods are frequently used; both are easy, quick, and inexpensive and can be performed on crude DNA extracted directly from complex matrices such as feces and food. Multiplex PCR protocols have also been developed to allow investigators to screen both stx1 and stx2 and other virulence genes such as eae and the plasmid-associated enterohemolysin gene, hlyA, which is specific for O157 (Paton and Paton, 1998). PCR followed by restriction digestion has been used to differentiate stx2 from the highly similar stx2-variants including stx2c (Manning et al., 2008), which is present on a distinct bacteriophage (Strauch, Schaudinn, and Beutin, 2004). Another RFLP-based method for stx2 and variants involve digesting the stx2 gene with four restriction enzymes (PvuII, HaeIII, HincII, and AccI). This method, named PHIA, demonstrated adequate discriminatory ability as the strains with similar digestion patterns clustered together (De Baets et al., 2004). Indeed, the application of PHIA to 36 and 27 strains of bovine and human origin, respectively, revealed increased variation in the bovine strains and also identified novel variants.

In a recent study conducted by Scheutz et al., 285 stx sequences representing all stx variants were analyzed in an effort to evaluate prior subtyping schemes and develop a more sensitive PCR-based subtyping assay. Following phylogenetic analysis of the stx sequence data, the stx1 strains clustered into three different groups with 13 unique subtypes, whereas the stx2 sequences clustered into five different groups with 93 unique subtypes (Scheutz et al.,
2012). These data illustrate the high degree of genetic variation of stx across STEC strains. Furthermore, mapping restriction enzyme sites onto stx sequence data demonstrated that some stx variants may have been mistyped in prior RFLP studies. The new subtyping method took these factors into consideration as well as the high likelihood of cross reactivity between stx2a, stx2c, and stx2d; higher annealing temperatures were found to be useful (Scheutz et al., 2012).

**Shiga toxin-encoding bacteriophage typing**

As determined by sequencing studies, the Shiga toxin-encoding prophages comprise six regions. Regions I–IV contain repeat sequences, while regions V and VI, located upstream and downstream of stx, are polymorphic and serve as useful targets for molecular typing studies (Sato et al., 2003; Shima et al., 2004; Sugimoto et al., 2011). Shima et al., for instance, developed a Shiga toxin-encoding bacteriophage typing method that involves PCR amplification and the subsequent restriction digestion of region V with BglII or EcoRV. The application of this method to 202 STEC strains resulted in 24 distinct RFLP patterns (Shima et al., 2004). As mentioned previously, PCR-based RFLP assays are quick and allow investigators to rapidly analyze multiple strains simultaneously. Nonetheless, the discriminatory power is lower than PFGE (Shima et al., 2004).

Characterizing the genomic location of the Shiga toxin-encoding phage(s) has also been described as a useful tool to differentiate among STEC strains. Such methods have also been utilized to examine bacteriophage integration, excision, or recombination events. The primary insertion sites for these bacteriophages include yehV, which encodes a protein that regulates curli production (Brown et al., 2001) and wrbA that encodes a multimeric flavodoxin-like protein (Grandori et al., 1998). Bacteriophage encoding stx1 and stx2 typically integrate at yehV and wrbA, respectively (Makino et al., 1999; Plunkett et al., 1999; Yokoyama et al., 2000), though integration at other sites including argW and sbcB have been described (Shaikh and Tarr, 2003; Kotewicz et al., 2008).

PCR has been used to amplify the regions spanning the length of the Shiga toxin-encoding bacteriophages as well as the right and left junctions of the insertion sites. In one study, 362 STEC O157:H7 isolates recovered from bovines and humans were screened by bacteriophage insertion site genotyping to determine the distribution of occupied and intact integration sites. Three major clusters were identified that varied based on the bacteriophage occupancy profile. Clusters 1 and 2 included strains with the Stx2 encoding bacteriophage inserted into a site other than wrbA or in wrbA, respectively, while cluster 3 included strains with the Stx1 encoding bacteriophage inserted into yehV and the Stx2 encoding bacteriophage inserted into wrbA (Shaikh and Tarr, 2003). A follow-up study also demonstrated that the bacteriophage
insertion site profiles varied across strains of different origins (Besser et al., 2007). Specifically, 95% of human-derived and 51% of bovine-derived strains belonged to one of the three clusters, suggesting that the STEC strains originating from cattle are considerably more diverse.

**Screening for informative SNPs**

Several polymorphisms have been shown to differentiate STEC O157:H7 strains. For example, Bono et al. analyzed 108 human-derived strains for SNPs in \( eae \) and \( tir \). The 1627 bp \( tir \) gene, which is present on the LEE island, has five polymorphic sites that classify the gene into 10 different alleles that can be used in epidemiological studies (Frankel et al., 1998; Hayward et al., 2006; Bono et al., 2007). SNP locus 255, which represents a T or A, varies across strains of different origins. The T, for example, was overrepresented in human-derived strains (99%) when compared to bovine strains (Bono et al., 2007).

Another polymorphic marker, \( q \), is located on the Stx-encoding bacteriophage and encodes an anti-terminator that regulates the production of late bacteriophage genes and toxin production (Friedman and Court, 1995). It was demonstrated that distinct \( q \) alleles were detected in human-derived strains belonging to the OBGS lineage I compared to the bovine-derived strains of lineage II (Lejeune et al., 2004).

Some SNPs have also been identified that can be used to differentiate phylogenetic lineages defined by a specific genotyping technique. One study, for instance, described four SNPs in genes ECs2357 at position 539, ECs2521 at position 1060, ECs3881 at position 438, and ECs4130 at position 630 that can detect one of the four predominant clades (Riordan et al., 2008) identified by SNP genotyping in a prior study (Manning et al., 2008). These four SNPs could most accurately differentiate STEC O157 strains belonging to clade 8, the lineage that was found to be associated with HUS cases in Michigan (Manning et al., 2008). An additional set of four SNPs in \( rhsA \), the gene encoding the RhsA protein with an unknown function, at positions 3468, 3478, 3479, and 3516 could also specifically identify clade 8 strains (Liu, Knabel, and Dudley, 2009; Hartzell et al., 2011). The goal of these studies was to develop a method that facilitates rapid detection of strains associated with more severe disease.

In a recent study, a subset of STEC O157 isolates from the Netherlands were characterized by LSPA, bacteriophage insertion site genotyping, \( stx \) profiling, and SNPs for \( tir \), \( q \), and \( rhsA \) locus 3468 (Franz et al., 2012). This study aimed to compare the distribution of clade 8 strains, as determined by the \( rhsA \) SNP assay, among STEC strains of diverse origins. Restriction digestion of the 635 bp \( rhsA \) amplicon with \( HaeII \) and \( Sau96I \) resulted in two (525 and 110 bp) and three (296, 273, and 66 bp) fragments, respectively, for clade 8 strains, and three \( HaeII \) (269, 256, and 110 bp) and two \( Sau96I \) (339 and
296 bp) fragments for strains of other clades. It was further demonstrated that the frequency of clade 8 was similar among STEC strains recovered from bovines, food, and humans (Franz et al., 2012). However, because clade 8 strains were not associated with a specific LSPA lineage as was shown previously (Eppinger et al., 2011), it is possible that classifying strains based on one polymorphism may not be ideal for lineage determination.

**Plasmid typing**

Most plasmid characterization studies involve comparing the size, number, and sequence of different plasmids. Plasmid pO157, which encodes an α-hemolysin (E-hly) (Schmidt, Karch, and Beutin, 1994), extracellular protease (espP), a type II secretion system (etpC to etpO), and toxin (toxA and toxB) (Burland et al., 1998), is present in virtually all STEC O157 strains (Levine et al., 1987; Ostroff et al., 1989). The role of pO157 in pathogenesis, however, is not fully understood (Johnson and Nolan, 2009), though recent studies have suggested an association with biofilm formation (Puttamreddy, Cornick, and Minion, 2010). Along with serotyping, plasmid typing was used to characterize the first STEC O157:H7 outbreak in Michigan and Oregon in 1982 and helped confirm the link between contaminated hamburger patties and disease cases (Wells et al., 1983).

Although plasmid profiling is not currently used for STEC characterization because of its lower discriminatory power relative to genome-based DNA fingerprinting and sequence-based methods, it can provide clues about pathogenesis, strain relationships, and virulence when used in conjunction with other typing methods and serotyping (Ratnam et al., 1988; Ostroff et al., 1989; Louie et al., 1999). In fact, additional types of plasmid have been discovered in STEC, which is important because the presence of a given plasmid can result in dramatic changes in virulence. The emergent STEC O104:H4 outbreak strain from Germany, for example, possesses an antibiotic resistance plasmid and the EAEC virulence plasmid (pAA) that facilitates colonization and biofilm formation (Sheikh et al., 2001; Bielaszewska et al., 2011; Al Safadi et al., 2012). Follow-up studies have since been conducted to screen STEC strain collections and environmental samples for the pAA and other O104:H4 virulence genes by PCR (Willford et al., 2011). Similarly, a study of pO157-positive STEC O157:H7 outbreak strains from China revealed the presence of a novel 37,785 bp conjugative plasmid (pO157_Sal) with 94% identity to a *Salmonella enterica* serovar Agona plasmid (Wang et al., 2011). PCR assays targeting four unique plasmid genes were utilized among 314 Chinese isolates and 12 isolates from different locations to determine the frequency of strains containing pO157_Sal. A total of 15 Chinese STEC isolates originating from a 1999 Xuzhou outbreak (Zheng et al., 2005) were positive for pO157_Sal; none of the screened US isolates were positive for the plasmid (Wang et al., 2011).
Molecular serotyping

STEC O157:H7 is the serotype that has been most commonly associated with human infections and outbreaks; however, there are hundreds of non-O157 STEC serotypes that have been identified. The primary non-O157 serotypes associated with human diseases include O26, O45, O103, O111, O121, and O145 (Johnson, Thorpe, and Sears, 2006). Conventional serotyping, the phenotypic expression of a specific type of somatic (O) antigen and the flagellar (H) antigen, has traditionally been performed by agglutination tests using anti-
*E. coli* polyclonal antiserum (Nataro and Kaper, 1998; Bettelheim and Beutin, 2003). These assays, however, can be time consuming, laborious, costly, and are typically restricted to large reference laboratories (Durso, Bono, and Keen, 2005), while cross-reactivity among serotypes is also problematic and can lead to misclassification. Hence, to decrease the time it takes to serotype STEC, many researchers have begun to utilize sequence data to determine the molecular serotype of a given strain. While molecular serotyping allows researchers to cut back on costs, labor, and time, the method is limited in that sequence data are not available for all known *E. coli* serotypes.

Genes involved in O antigen synthesis are located in the *rfb* cluster, which can range in size from 6 to 19 genes as well as in organization (Clermont et al., 2007). Interestingly, SNPs have been described, which are specific for different O-types and can serve as targets for molecular serotyping schemes. One recent study, for example, identified 22 informative SNPs from 164 STEC strains representing serotypes O26, O45, O103, O111, O121, and O145 that can be detected using MALDI-TOF assays (Norman, Strockbine, and Bono, 2012). At least one SNP was determined to be unique to each O-antigen type examined and thus this assay may represent a rapid way to detect these serotypes in the future.

Long-range PCR has previously been used to amplify the entire *rfb* cluster. RFLP analysis of amplicons following digestion with multiple restriction enzymes (e.g., MboII) was found to adequately characterize 148 serotypes and facilitated the identification of novel types (Coimbra et al., 2000). One problem encountered with this method, however, was that similar digestion patterns were observed if the *rfb* cluster was too similar, thereby requiring restriction digestion with two additional endonucleases, *Hha*I and *Sau*I (Coimbra et al., 2000). Long-range PCR is more expensive than conventional PCR and in some cases it can result in complex banding patterns that require analysis using specialized software such as the web-based molecular serotyping tool (MST) program (Coimbra et al., 2010). MST compares user-generated *rfb*-RFLP patterns to its database comprising over 170 *Shigella* and *E. coli* RFLP patterns to facilitate the identification of O-types.

Multiplex PCR has also been developed to distinguish among multiple serotypes. To detect serotypes O113, O111, and O157, for example, the
amplification of different regions of \( rfb \) results in fragment sizes of 593, 406, and 259 bp, respectively (Paton and Paton, 1999). Because multiplex PCR is easy to perform and is relatively inexpensive, these assays can be used to screen complex samples such as feces or foods, for the presence or absence of specific O-types. In addition to \( rfb \), the \( wzx \) gene encoding the O-antigen flippase and the \( wzy \) encoding the O-antigen polymerase have been targeted in PCR assays (DebRoy et al., 2004). For instance, the amplification of \( wzx \) in 50 O26 and O113 strains resulted in amplicons that varied in size, with \( wzx \) yielding a 152 bp fragment for O26 and a 771 bp fragment for O113, and \( wzy \) yielding fragments of 276 bp for O26 and 491 bp for O113. These assays were found to be highly specific and no fragments were amplified in a set of 178 non-O26 and -O113 strains (DebRoy et al., 2004) as were other PCR assays targeting both O26:H11 (Durso, Bono, and Keen, 2005) and O111:H8 (Durso, Bono, and Keen, 2007) strains. Similarly, a rapid 5’-nuclease PCR amplification assay that utilizes TaqMan probes and targets O-antigen-specific genes has also been developed (Perelle et al., 2004). This assay enables the identification of O157 (\( rfbE \)), O111 (\( wbdI \)), O26 (\( wzx \)), O113 (\( wzy \)), O91 (\( wzy \)), O55 (\( wbgN \)), O145 (\( ihpI \)), O103 (\( eae \)), and the flagellar H7 antigen gene (\( fliC \) H7). Among 137 strains representing 44 serotypes, the assay was accurate in all but two strains, as cross reactivity was observed in non-O103 strains when the O103-specific \( eae \) marker was examined.

Although these techniques provide accurate and rapid method for serotyping STEC O157 and a subset of non-O157 STEC, many investigators are still faced with the challenge of serotyping STEC that cannot be typed by current phenotypic or molecular serotyping methods. Because serotyping data are historically important and have been linked to disease severity, it is imperative to obtain more sequencing data for O-antigen-specific genes representing all STEC serotypes that have previously been implicated in human infections.

**Conclusions**

In this chapter, we have highlighted the molecular techniques that are readily available for characterizing the diverse STEC strain population. Although there are advantages and disadvantages associated with each technique, the selection of a given method is entirely dependent on the research question as well as the cost and feasibility. Genome sequencing costs continue to decrease and sequence and data analysis tools continue to improve. These trends suggest that we will likely see many changes and improvements in the type and application of STEC typing tools that will be used in the near future.
References


REFERENCES


