Multiple evolutionary trajectories for non-O157 Shiga toxigenic *Escherichia coli*

Nabil-Fareed Alikhan[1,2,6*], Nathan L. Bachmann[1,2,3], Nouri L. Ben Zakour[1,2,7], Nicola K. Petty[1,2,4], Mitchell Stanton-Cook[1,2], Jayde A. Gawthorne[1,2], Rowland Cobbold[5], Mark A. Schembri[1,2], Scott A. Beatson[1,2].

[1]School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland, Australia. [2]Australian Infectious Diseases Research Centre, The University of Queensland, Brisbane, Queensland, Australia.[3] Present address: The University of Sydney. [4] Present address: iThree Institute, University of Technology Sydney. [5]School of Veterinary Science, The University of Queensland, Brisbane, Queensland. [6]Present address: Quadram Institute Bioscience, Norwich, United Kingdom. [7]Present address: The Westmead Institute for Medical Research and The University of Sydney.

*Corresponding author:

Nabil-Fareed Alikhan, Quadram Institute Bioscience, Norwich, United Kingdom, NR4 7UA;

Telephone +44 1603 255 000; email nabil-fareed.alikhan@quadram.ac.uk

Acknowledgements

This project was supported by the Queensland State Government's Smart Futures Fund

National and International Research Alliances Program.

Abstract

Shiga toxigenic *Escherichia coli* (STEC) is an emerging global pathogen and remains a major cause of food-borne illness with more severe symptoms including hemorrhagic colitis and hemolytic-uremic syndrome. Since the characterization of the archetypal STEC serotype, *E. coli* O157:H7, more than 250 STEC serotypes have been defined. Many of these non-O157 STEC are associated with clinical cases of equal severity as O157. In this study, we utilize whole genome sequencing of 44 STEC strains from eight serogroups associated with human infection to establish their evolutionary relationship and contrast this with their virulence gene profile and established typing methods. Our phylogenomic analysis delineated these STEC strains into seven distinct lineages, each with a characteristic repertoire of virulence factors. Some lineages included commensal or other *E. coli* pathotypes. Multiple independent acquisitions of the Locus for Enterocyte Effacement were identified, each associated with a distinct repertoire of effector genes. Lineages were inconsistent with O-antigen typing in several instances, potentially indicating lateral gene transfer of the O-antigen region. STEC lineages could be defined by the conservation of clustered regularly interspaced short palindromic repeats (CRISPRs), however, no CRISPR profile could differentiate STEC from other *E. coli* strains. Six genomic regions (ranging from 500 bp - 10 kbp) were found to be conserved across all STEC in this dataset and may dictate interactions with Stx phage lysogeny. The genomic analyses reported here present non-O157 STEC as a diverse group of pathogenic *E. coli* emerging from multiple lineages that independently acquired mobile genetic elements that promote pathogenesis.

Introduction

Food-borne pathogens persist as a major cause of clinical infection within the industrialized world (Newell et al., 2010, Scharff, 2012). Shiga toxigenic *E. coli* (STEC) is one such emerging global food-borne pathogen responsible for severe human disease with symptoms of hemorrhagic colitis (HC) and hemolytic-uremic syndrome (HUS) (Riley et al., 1983, Karmali et al., 1983). Shiga toxins (Stx) are lambdoid phage-encoded verocytotoxins that are homologous to *Shigella dysenteriae* type 1 toxins and cause HUS in humans (O'Brien et al., 1982, Konowalchuk et al., 1977). Stx has been classified into two main groups (Stx1 and Stx2), with Stx1 and Stx2 both further divided into a number of more closely related subtypes (Scheutz et al., 2012).

Enterohemorrhagic *E. coli* (EHEC), a subset of STEC associated with human disease, were first identified in 1982 from an outbreak of contaminated beef (Centers for Disease, 1982). Since then, the serotype associated with that outbreak, O157:H7, has been linked to many other major outbreaks across the globe, including the 1996 Sakai outbreak (Michino et al., 1999). The features used to serotype *E. coli* include the O-antigen of the lipopolysaccharide (LPS) and the H-antigen of the flagella (Levine et al., 1984). More than 150 different O antigens have been recognized and each one defines a specific serogroup (Kaper et al., 2004). A combination of the O and H antigens is used to define a serotype (Nataro and Kaper, 1998). An additional 250 STEC serotypes have also been associated with human disease, often referred to collectively as non-O157 (Johnson et al., 2006).

STEC are a heterogeneous group of *E. coli* that exhibit a high degree of genomic and

phenotypic diversity. Only certain Shiga toxin harboring serogroups have been associated with human disease (i.e. are EHEC), and within these serogroups there are also differences in their association with human disease outbreaks (Brooks et al., 2005, Johnson et al., 2006). In the United States, these differences have been formalized into a 'Seropathotyping' scheme, which ranks serogroups by degree of pathogenesis based on outbreak prevalence and epidemiological studies. This scale places O157:H7 as the most prevalent STEC serotype followed by O26, O111, O103, O121 and O145, and finally O91, O104 and O113 (Karmali et al., 2003). Furthermore, seropathotypes A-C include serotypes associated with severe disease, either hemorrhagic colitis or HUS, where seropathotype D does not. Seropathotype E includes those strains not associated with human disease. This ranking does not apply globally, as in some jurisdictions, O157 and these key non-O157 strains have varying relative impacts on public health, whilst additional serogroups, including O45, O128 and O174,contribute significantly to disease burden in other regions (Johnson et al., 2006, Smith et al., 2013). As yet, it is unclear how much of the variation observed between serogroups is linked to differences in the host response to infection, or is due to variations in virulence gene content between strains.

Other than encoding at least one *stx* gene, some STEC, including O157, also carry the Locus of Enterocyte Effacement (LEE) pathogenicity island, which encodes a Type III Secretion System (T3SS), the adhesion Intimin and its translocated receptor Tir(McDaniel et al., 1995, Deng et al., 2004). Many of the genes encoded within the LEE pathogenicity island are responsible for the attaching and effacing phenotype induced by EHEC and Enteropathogenic *E. coli* (EPEC) strains (Jerse et al., 1990). LEE encodes the structural genes of the *E. coli* T3SS and up to six effector proteins (McDaniel et al., 1995). In addition, numerous other effector genes have been found scattered throughout the chromosome, many in clusters referred to as 'exchangeable effector loci' (EELs) and they are often associated with mobile genetic elements such as prophages. In the *E. coli* O157:H7 Sakai strain, it was determined that phage associated EELs are located at the 3' end of prophage regions immediately adjacent to tail fiber genes and can be easily distinguished from the genomic backbone by a bias towards low GC content (Tobe et al., 2006).

A number of LEE-negative STEC are also associated with human infection. These strains have other genes that contribute to virulence, including *epeA, sab, and subAB(Newton et al., 2009, Herold et al., 2009, Paton and Paton, 2005)*, and appear to adhere to epithelial cells via alternative mechanisms. This is most dramatically demonstrated in the case of the 2011 German outbreak strain O104:H4, which appears to have descended from an enteroaggregative *E. coli* (EAEC) progenitor strain and acquired the genes encoding aggregative adherence fimbriae that are associated with this pathotype (Mellmann et al., 2011, Rasko et al., 2011, Rohde et al., 2011). STEC pathogenesis is also mediated by additional virulence factors carried by mobile genetic elements. A number of non-LEE encoded effectors have been identified; current evidence indicates the genes encoded by these effectors are mobilized through transduction by bacteriophage (Tobe et al., 2006). Furthermore, a number of plasmid-encoded virulence genes have been identified in STEC, including those encoding hemolysin (*hlyA*), proteases (*espP* and *katP)* and cytotoxins (*subA) (Paton and Paton, 2005, Bosilevac and Koohmaraie, 2011, Paton et al., 2004)*.

Despite extensive screening, no virulence factor or marker, other than the *stx* genes themselves, are universally conserved across all STEC although the combination of stx_2 and eae is associated with more severe clinical outcomes associated with EHEC infection (Smith et al., 2013).

STEC strains also exhibit diversity in their phylogeny, and are represented across the *E. coli* species. The O157 serotype is found within *E. coli* phylogroup E, having evolved from an ancestral O55:H7 EPEC strain(Feng et al., 1998). In contrast, the dominant non-O157 STEC serotypes are generally found within the B1 phylogroup (Mora et al., 2012). It has been established that non-O157 STEC, like other pathogenic *E. coli*, is potentially comprised of multiple independent lineages that acquired key virulence factors through stepwise evolution (Wick et al., 2005, Donnenberg and Whittam, 2001, Croxen and Finlay, 2010). This concept of parallel evolution was first presented through phylogenetic analysis of seven housekeeping genes (Reid et al., 2000) and confirmed by whole genome sequencing of O26, O103 and O111 STEC strains (Ogura et al., 2009). These studies demonstrated that STEC pathogenicity is derived from the acquisition of virulence factors via multiple independent events, and mediated via the acquisition of mobile genetic elements.

To establish the evolutionary relationship of non-O157 STEC, and provide a phylogenomic framework for investigating differences in virulence gene profile and established typing methods, we performed whole genome sequencing of forty-four genetically diverse STEC strains from eight serotypes commonly associated with human disease (O26, O111, O91, O128, O103, O113, O121 and O45). The strains were obtained from Australia and the United States, with the majority of strains representing the most clinically relevant serotypes O26 and O111. Overall, this study offers an overview of non-O157 STEC, including both LEE-negative and LEE-positive strains of clinical significance, presenting STEC as a diverse group of E. coli from at least seven distinct lineages with varying virulence profiles. Through the high granularity granted by next-generation sequencing, we were able to contrast STEC phylogeny and established typing methods, including O-antigen typing and EcMLST, to test whether these approaches prove valid against a whole genome phylogeny. We were also able to determine whether a previously published CRISPR typing method for LEE-positive non-O157 strains is applicable to STEC at large.

Methods

Bacterial strains

Forty-four STEC strains from collections held by Queensland Health Forensic and Scientific Services (QHFSS) (n=16), The Commonwealth Scientific and Industrial Research Organisation (CSIRO) (n=23) and Washington State University (WSU) (n=5) were used in this study. These strains encompass the major non-O157 serogroups and include strains from human, animal and contaminated food sources (Table 1).

Genome sequencing and annotation

All strains were sequenced on the Illumina HiSeq2000 platform. The resulting paired end 100 base pair reads (average of 302bp insert size with standard deviation of 108bp) were filtered using PRINSEQ-lite (Schmieder and Edwards, 2011) . Reads were trimmed at both ends to achieve a mean quality cut-off of Q20 and minimum read length of 80 base pairs.

Filtered reads were assembled using SPAdes version 2.5.0 (Bankevich et al., 2012) with default kmers (21, 33 and 55) and with inbuilt read and scaffold correction, and the "--careful" flag. The resulting assemblies included a subset of low coverage scaffolds, which were artifacts of the sequencing and assembly process. These low coverage scaffolds could be partitioned from scaffolds with an expected coverage, and were filtered out using a coverage cut-off calculated independently for each assembly. The cut-off was based on scaffolds average coverage while adjusting for GC bias and had an average of 10 with a standard deviation of 5. In contrast, the average read coverage for each genome was 252 with a standard deviation of 134. The filtered, assembled scaffolds were ordered using Mauve ContigMover (mauve_snapshot_2012-06-07) (Rissman et al., 2009) against the published *E. coli* O111:H- strain, 11128 (Accession no. AP010960), and annotated using Prokka version 1.5.2 (Seemann, 2014).

Phylogenetic analysis and recombination testing

To compile a set of core gene sequences for subsequent phylogenetic analysis, we first retrieved all predicted gene sequences from the published complete genome of *E. coli* O111:H- str. 11128 (Accession No. AP010960), and identified homologs in available complete *E. coli* genomes from the B1 Phylogroup, non-O157 STEC and O157:H7 Sakai using the Basic Local Alignment Search Tool (BLAST) (version 2.2.26+) (Camacho et al., 2009). Putative homologs for each reference gene were defined as the predicted gene with the best scoring BLAST alignment match that had greater than 90% nucleotide identity over 90% gene length to the respective reference gene from strain 11128. Each cluster of homologous genes was subsequently aligned using MUSCLE (Edgar, 2004)and then concatenated into a single alignment sequence. Variable sites were extracted from this alignment to produce a concatenated and aligned sequence of SNPs for each taxa in PHYLIP format for phylogenetic analysis. This approach was implemented through an in-house script Dryad (http://github.com/happykhan/Dryad-SA). The aligned SNP sequence was used in PhyML (v20120412) (Guindon et al., 2010) to infer a maximum-likelihood phylogram using the HKY85 substitution model and 400 bootstraps.

To assess if recombination within core gene families impacted on the topology of the tree, a strict filter was imposed such that ortholog groups were removed if they exhibited significant evidence for recombination ($p < 0.05$) for at least two out of three recombination tests (NSS, $MaxX^2$ and PHI) implemented in PhiPack (Bruen et al., 2006). Variable sites were extracted and a phylogenetic tree was determined using the same methodology as described above.

Sequence type and serogroup determination.

In silico E. coli MLST (EcMLST) was performed using the sequences of seven housekeeping genes, *aspC*, *clpX*, *fadD*, *icdA*, *lysP*, *mdh* and *uid* as defined within the EcMLST database (Weihong et al., 2004). Only exact matches using nucleotide-to-nucleotide BLAST (BLASTn v2.2.26+) (Camacho et al., 2009) were used to identify each allele variant. Allele profiles were queried against the EcMLST database to determine sequence type. The EcMLST 15 locus scheme, which extends the 7 locus scheme and also included *arcA*, *aroE*, *cyaA*, *dnaG*, *grpE*, *mltD*, *mutS* and *rpoS*, was applied to strains that could not be distinguished with 7 locus alone. The serogroup of each strain was confirmed *in silico* by comparing the nucleotide sequences, using BLAST (BLASTn

v2.2.26+), of a region within the LPS biosynthesis locus (between *hisG* and *yegH*) for O-antigen typing and the *fliC* gene for flagella typing. Comparisons of the O-antigen biosynthesis region were visualized using the Artemis Comparison Tool (Carver et al., 2005) and EasyFig (Sullivan et al., 2011).

Stx, LEE and effector profiling

Shiga toxin encoding genes were detected through amino acid and nucleotide sequence comparison (BLASTp and BLASTn) to Stx2a (protein id: CAA71748) and Stx1a (protein ID: AAG57228) sequences from *E. coli* O157:H7 str. EDL993. Detected *stx* subunit B gene sequences were compared with BLASTn to the GenBank nucleotide non-redundant database and required an identical nucleotide sequence match to previously sequenced *stx* genes from STEC strains. *stx* subtypes can vary by as little as a single SNP and isolates harboring multiple *stx2* genes can have these genes interpreted as a collapsed repeat in assembled contigs. To address this, *stx* copy number and type was determined through the analysis of SNPs from the *stx* genes generated from mapping reads onto the O157:H7 Sakai genome (Accession no. BA000007) with BWA (Li and Durbin, 2009). *stx* copy number and type was also confirmed with Mapsembler (Peterlongo and Chikhi, 2012). The subtype for matching strains was defined in accordance with existing literature, where the designations Stx1a(protein ID: AAG57228), Stx2a(protein id: CAA71748), $Stx2d_{activatable}$(protein id: CCA61220), Stx2c(protein id: ABB36585) and Stx2d(protein id: AE779208) were derived from (Scheutz et al., 2012),whereas Stx1c(protein id: CAC88707) was retrieved from (Brett et al., 2003).

The LEE insertion site for sequenced strains was determined through genome comparison to known LEE insertion sites as defined in (Bertin et al., 2004) using the Artemis Comparison Tool (Carver et al., 2005); *yicK* to *selC*, observed in O157:H7 Sakai (Accession no. BA000007), *yghD* to *pheV,* observed in O111:H- 11128 (Accession no. AP010960) and O103:H212009 (Accession no. AP010958), and *cadC* to *pheU* observed in O26:H1111368 (AP010953).

Effector repertoires were annotated in each draft genome using the EffectorFAM database of profile HMMs built from confirmed effector families (https://github.com/NathanBachmann/EffectorFam). Genomic context of each effector was carried out by ACT comparisons with representative complete genomes from each LEE+ lineage (i.e. O26:H11 str. 11368 for ST106A, O111:H- str. 1128 for ST106B, and O103:H2 str. 12009 for ST118).

Virulence profile

Virulence factor profiles across *E. coli* genomes were generated using SeqFindr (http://github.com/mscook/seqfindr). Virulence factors were considered present with greater than 80% average nucleotide identity across the total reference gene length. Comparisons between individual genomes and verification of SeqFindr results were performed using BLAST+ (v2.2.26+) (Camacho et al., 2009), Artemis Comparison Tool (Carver et al., 2005) and EasyFig(Sullivan et al., 2011).

CRISPR detection

Genomes were interrogated for CRISPR spacer sequences using PILER-CR (Edgar, 2007) and verified with CRISPRFinder (Grissa et al., 2007) across the whole genome.

CRISPR loci defined in (Diez-Villasenor et al., 2010)were also inspected using nucleotide-to-nucleotide BLAST (BLASTn v2.2.26+), Artemis Comparison Tool (Carver et al., 2005) and EasyFig(Sullivan et al., 2011). The distribution of unique spacer sequences was visualized using binCrisp, a custom python script developed as part of this study (https://github.com/happykhan/binCrisp). The source code for this script has been included in Appendix 3.

Whole genome comparison

The sequences of forty-four non-O157 STEC genomes sequenced as part of this study and 11 representative strains from the major *E. coli* pathotypes and phylogroups for which the complete genome was available were aligned using Mugsy (Angiuoli and Salzberg, 2011) version 1.3. Representatives included STEC strains: O26:H11 11368 (Accession no. AP010953), O157:H7 Sakai (Accession no. BA000007), O103:H212009(Accession no AP010958) and O111:H- 11128 (Accession no AP010960); commensal strains: K12 MG1665 (Accession no. U00096), IAI1 (Accession no CU928160), HS (Accession CP000802)and BL21 (Accession no. CP001665); Enterotoxigenic *E. coli* (ETEC) E24377A (Accession no CP000800); Enteroaggregative *E. coli* (EAEC) 55989 (Accession no CU928145); and Uropathogenic *E. coli* (UPEC) CFT073 (Accession no. AE014075).

Alignment blocks from the whole genome alignment were filtered to identify putative STEC-specific regions. Alignment blocks required no corresponding match in K12 MG1665 or HS but were conserved in twelve STEC genomes chosen as a cross-section of the STEC lineages defined in this study. These included published genomes O26:H11 11368 (Accession no. AP010953), O157:H7 Sakai (Accession no. BA000007), O103:H212009 (Accession no. AP010958) and O111:H- 11128 (Accession no AP010960) and strains sequenced as part of this study; n10 (O111:H8), n01 (O26:H11), n02 (O91:H10), n28 (O91:H21), n15 (O113:H21), n17 (O103:H21), n16 (O128:H2) and n43 (O121). BLAST (BLASTn v2.2.26+) comparisons were used to verify that regions were conserved across all STEC strains(Camacho et al., 2009).

Results

Genome assembly

Forty-four STEC strains sourced from Australia and the United States, and originating from human, animal and contaminated food origins, were investigated in this study. These strains represented 20 different serotypes from serogroups associated with human disease including O26, O111, O91, O128, O103, O113, O121 and O45. The average read coverage for each genome for the 44 STEC strains (Table 1) was 252±134 times the total genome size. Each genome assembly had an average total length of 5,389,684±239,389 bp, and an average N50 scaffold size of 90,958 ±30,439 bp. The number of scaffolds within the assemblies ranged from 121 to 979 scaffolds (mean of 291).

Phylogenetic analysis

To obtain a high-resolution overview of the non-O157 STEC genomes, a maximum-likelihood tree based on 2,153 aligned gene sequences (including 48,912 variable sites) was generated for the 44 strains, as well as 7 previously published *E. coli* reference genomes from the B1 phylogroup and *E. coli* O157 Sakai as an out-group (Figure 1). The

phylogram revealed that STEC strains form multiple lineages that have evolved in parallel and are distinct from O157. Individual lineages could be classified according to a particular sequence type according to the EcMLST seven allele scheme (Figure 1 and Table 1).

To enable a scalable nomenclature we refer to clades by their EcMLST. Of the 44 STEC strains sequenced in this study, three strains, n19, n32 (O91:NM) and n43 (O121), did not cluster with any other strains. Strain n43 (ST182) in particular was highly divergent to all B1 strains (Figure 1). Notably, all O26 and O111 strains were ST106, but fell into two distinct sub-lineages that are referred to here as ST106A and ST106B. This distinction was also supported by the extended 15 allele EcMLST scheme (Weihong et al., 2004), which defined ST106A as ST41 and ST106B as ST39, respectively. The defining sequence in the 15 allele EcMLST scheme was *mutS*, whereby ST106A and ST106B strains were *mutS* alleles 18 and 13, respectively. ST106A encompassed all flagella type 11 (H11) strains from this study (including both O26:H11 and O111:H11), while ST106B included all O111 strains other than O111:H11.

Phylogroup B1 STEC strains formed a clade structure that was polyphyletic and intermingled with other non-STEC *E. coli* from the B1 phylogroup including EAEC, ETEC and commensal strains. Notably, ST461 strains were most closely related to ETEC strain E24377A rather than other STEC strains, indicating a common ancestor. In general, STEC clades defined in this study included a mixture of strains isolated from clinical samples, contaminated food or ruminant livestock.

To verify the lineages defined in the whole genome phylogeny, SNP counts were generated through a pair-wise comparison of all strains using the dnadiff script from the MUMmer package (Kurtz et al., 2004). STEC strains from different lineages differed by an average of 34,119 ±2,750 SNPs, whereas strains within the same lineage differed by less than 7,000 SNPs. As a frame of reference, B1 Phylogroup STEC diverged by, on average, 81,434 ±1,806 SNPs from O157:H7 Sakai. ST106A and ST106B differed on average by 15,859 SNPs (standard deviation 1,141). ST106A and ST106B strains differed by 3,733±1,595 and 1,625±878 SNPs on average, respectively.

To ensure that the lineages defined in this study were not the product of rapid evolution due to recombination we also determined a recombination-free tree in which core genes were removed if there was significant evidence for recombination (Supplementary Figure 1 [Figure 9]). The recombination-free tree included 1,136 genes and exhibited the same phylogenetic topology for major STEC lineges (ST106A, ST106B, ST118, ST379,ST234, ST89 and ST461) as observed in Figure 5, with nodes displaying >90% bootstrap support.

Shiga Toxins and the Locus of Enterocyte Effacement

Shiga toxin type was heterogeneous across strains within this study suggesting that the acquisition and loss of *stx* genes is dynamic and has occurred in multiple independent events. Stx type (summarized in Table 1 and Figure 2) was consistent with several of the defined lineages, including ST106A, ST106B, ST461 and ST379. ST461 and ST379 strains carried *stx2d*, as well as *stx1c* and $stx2d_{act}$, respectively. ST106A strains differed from ST106B strains; whereas ST106B strains carried both *stx1a* and *stx2a*, ST106A strains only carried *stx1a*. This suggests that *stx2a* was gained in ST106B, or alternately, was lost in ST106A after the two lineages diverged. There was also evidence for variation in *stx* content within the defined lineages examined. For example, n30 (ST89) contains

stx2d$_{act}$ and *stx1a* while all other strains within this lineage only contain *stx2d*$_{act}$. Similarly, strains n20 and n21 (ST234) were *stx2a* and *stx2d* positive, however other ST234 strains contained either *stx2a* or *stx2d* (but not both), with acquisition of two *stx* prophages within the lineage followed by deletion within individual strains being the most parsimonious explanation. Due to the repetitive nature of the Stx encoding prophage, Stx genes did not assemble along with their cognate phage or the insertion site. Attempts to resolves this through mapping the underlying reads showed no clear link between prophage and the Stx genes. As such, it was not possible to determine the Stx-encoding phage insertion site. This could be addressed with re-sequencing with sequencing technologies boasting a longer read length.

The Locus of Enterocyte Effacement (LEE) was present in all strains within ST106A, ST106B, ST118 and ST182, but absent from all other phylogroup B1 lineages in this study. The core LEE regions (encoding the T3SS) from the 44 strains examined in this study shared greater than 95% nucleotide sequence conservation with the corresponding LEE region from *E. coli* 11368 (O26:H11; Figure 3). The site of integration of the LEE was determined by sequence comparison to known insertion sites, namely *pheU*, *pheV* and *selC* as previously observed in *E. coli* strains 11368 (O26:H11), 11128 (O111:H-) and Sakai (O157:H7), respectively (Bertin et al., 2004). Among the subset of strains for which the LEE insertion site could be determined; n37 and n48 (ST118) contained the LEE carrying the Intimin epsilon variant inserted into *pheV*; strains n10 and n11 (ST106B) contained the LEE carrying the Intimin theta variant inserted into *pheV*; and strains n1, n3, n4, n5, n6, n12, n13, n38, n39, n40 and n42 (ST106A) and n43 (ST182) contained the LEE inserted into *pheU*. The latter two cases likely represent independent events given the evolutionary distance between these two lineages and the differences in the Intimin type (ST106A: Intimin beta; ST182: Intimin epsilon) (Figure 1). The LEE insertion site could not be unequivocally determined in strains n8, n9, n17, n36, n37 and n44 from the draft sequence data alone, although the likely insertion sites could be predicted given the position of the LEE within other strains from the same lineage (Figure 1). Taken together, the data indicate that Stx and/or LEE acquisition has occurred multiple independent times among strains in the B1 phylogroup to give rise to different STEC lineages.

Type III secreted effectors

All non-O157 LEE+ STEC were found to encode the six known LEE encoded effector genes (*espG, espZ, espH, map, tir, espF*). In addition, 21 different phage-associated exchangeable effector loci (EELs) were found within the genomes of strains that encoded the LEE, with some variation in effector repertoire between and within lineages (Supplementary Table 1). A number of EELs were consistent with previously defined effector loci (Ogura et al., 2009) and these designations have been included in Supplementary Table 1 to maintain a standard nomenclature. Notably, O121 strain n43 encoded an effector loci with a gene order not previously observed in (Ogura et al., 2009), encoding NleO, NleN, NleM. The largest number of effectors (>30 per genome) were found to be encoded in the ST106 and ST118 lineages, with approximately half this number identified in the O121 strain n43 (Supplementary Table 1). Each EEL (EEL1-EEL21) differs in effector gene content, order and number of effectors. Six of the ten STEC O26:H11 strains within the ST106A lineage are missing one or two EELs when compared to the reference O26:H11 str. 11368 genome. Likewise, the O111:H11 strains within

lineage ST106A shared a similar EEL profile as the O26:H11 strains (Supplementary Table 2). In contrast, the O111 strains from lineage ST106B contained several different EELs compared with the ST106A strains (Supplementary Table X3) consistent with the acquisition of different prophages after their acquisition of LEE. Interestingly, some EELs (EEL01, 02, 05, 10) were shared by both ST106A and ST106B lineages, albeit with some minor differences indicative of lineage-specific loss of individual genes in the case of EEL05 (Supplementary Figure 2). Examination of other LEE+ STEC strains in our collection identified other EELs shared between phylogenetically distributed strains, including the O121 strain n43 from the ST182 lineage (indicating independent acquisition of the same EEL).

STEC virulence factors

A number of other virulence factors have been associated with STEC pathogenesis in previous studies. We queried the strains examined in this study as well as selected strains from the B1 phylogroup using BLAST and read-mapping for a range of STEC virulence factors, including genes encoding adhesins, autotransporter proteins, fimbriae, cytotoxins and genes from plasmids origin defined pathogenicity islands (including the Yersinia high pathogenicity island [HPI], O-island 112 and O-island 43/48; Figure 2)

The *iha* and *ehaA* genes have been previously described as well conserved in STEC (Toma et al., 2004, Tarr et al., 2000, Easton et al., 2011). Within this study, all STEC strains were positive for *ehaA,* however *ehaA* was also present in non-STEC strains IAI1, 59899 and E24377A. The *iha* gene was conserved in all strains except for ST234 strains and strains n39, n38 and n51 (ST106A). The distribution of other virulence factors varied, even among strains within the same serotype or sequence type. The O-islands 122 and 43/48 were conserved among LEE positive strains, however *pagC* from the O-island 122 was absent in ST106A and ST118 strains. The *fyuA* gene, used as a marker for the *Yersinia* HPI, was conserved in LEE positive ST106A strains and ST379 strains, but not in ST106B strains or strain n43 (O121).

Several virulence factors have been associated exclusively with LEE negative strains; *saa* (Paton et al., 2001), *sab* (Herold et al., 2009), *epeA* (Leyton et al., 2003) and *subAB* (Cergole-Novella et al., 2007).The *saa* gene is often associated with LEE negative strains that originate from ruminants, but has not been associated with clinical STEC identified from HC or HUS patients. The *saa* gene was absent from all strains examined in this study. The *sab*, *epeA* and *subAB* genes were originally identified on a large plasmid carried by an STEC O113 strain; these genes are conserved in all O113 strains (ST234). These genes were also identified in strain n47 (ST650), suggesting that n47 carries a similar plasmid.

Genes associated with the pSAK virulence plasmid, namely *espP*, *katP*, and *hlyA,* were present in many ST106A and ST106B strains with some exceptions. For example, the pSAK associated genes, *espP, katP*, and *hlyA,* are present in n02 (ST461 O91:10) and n48 (ST118 O103:H2) but absent from other strains of the same sequence type. Conversely, all strains within ST106B possessed the same plasmid-associated genes except n10 (O111:H8).This plasmid profile is consistent with an ancestral acquisition of a pSAK-like plasmid prior to divergence of ST106A and ST106B lineages, with subsequent sporadic strain-specific loss. However, we cannot rule out multiple independent acquisition

without complete sequencing of the plasmid content of these strains. All strains were negative for the *aggA*, *aggC* and *aggR* genes identified in the plasmid carried by German outbreak O104:H4 strains.

Comparison of serogroup and phylogenomic approaches

O-antigen serotype was lineage specific in ST106B (O111), ST379 (O128) and ST234 (O113) strains (Figure cross-references). In contrast, O-antigen serotype was inconsistent with whole genome phylogenomics and sequence typing in three instances involving O91, O103/O45,and O111/O26 strains, indicating wide-spread lateral gene transfer (LGT) of the O-antigen biosynthesis genes. O91 strains clustered into three separate lineages in the whole genome phylogenetic tree (Figure 1): (i) O91:H10 strains belonged to the ST461 lineage;(ii) O91:H21 strains belonged to the ST89 lineage which also contained a distinct O-antigen untypeable H21 strain (n41); (iii) the O91:NM strain n32 was typed as ST815 and did not cluster with any other STEC strains.

While O45 and O103 have been described as distinct members of the 'top-six' non-O157 serogroups by the Centre of Disease Control and Prevention (Paddock et al., 2012), this work classifies these two serogroups within a single lineage (ST118) (Figure 1). This finding was validated through pair-wise SNP counts calculated using dnadiff from the MUMmer package (Kurtz et al., 2004), which showed that strain n44 (O45) is closely related to O103 strains. Strain n44 differed on average by 5,763 SNPs to O103 strains within ST118, which was within one standard deviation (761 SNPs) of the mean SNP difference (5,494 SNPS) for all strains in ST118 regardless of serotype. The ST106A sub-lineage includedO111:H11 and O26:H11 strains, and was distinct from the O111:H8,O111:NM, and O111:H- strains that comprised the ST106B sub-lineage. This indicates that O111:H11 and O26:H11 strains share a common ancestry, and that the O26 O-antigen region was most likely acquired by lateral gene transfer in ST106A.

Serotype was determined by nucleotide alignment of the region between the *gnd* and *galF* genes in the O antigen biosynthesis locus. Regions with the same O-antigen serotype possessed >98% sequence conservation over the O-antigen biosynthesis region, whereas regions with different O-antigen sequences shared no significant nucleotide conservation and very low (7-26%) amino acid identity. This was also observed between the most closely related strains that showed evidence of O-antigen lateral gene transfer, namely O26 and O111 strains from ST106A and ST106B (Figure 4).

O-antigen typing was also validated by sequence comparison of the *gnd* gene, which encodes 6-phosphogluconate dehydrogenase, and is located immediately upstream of the O-antigen locus. This approach has previously been used for molecular-based serogrouping of STEC (Gilmour et al., 2007), where *gnd* sequences with >99% nucleotide conservation define the same O-antigen type (Gilmour et al., 2009). Indeed, a Maximum-likelihood consensus tree based on the *gnd* gene sequence (Figure 6) corresponded with O-antigen type (Table 1) for most of the STEC strains examined. The exception was n32 (O91:NM), which did not cluster with other O91 strains or any other STEC strain. BLAST comparisons showed that the *gnd* sequence from n32 shared, on average, 96.38% nucleotide conservation between other O91 strains and 95.52% nucleotide conservation with other STEC. Comparison of the sequence of the entire O-antigen region in n32 to other O91 and STEC showed the same level of conservation (Figure 4). These data suggest that n32 may have acquired the genes encoding the O91 O-antigen region through lateral gene transfer independent of the *gnd* gene. However, this would suggest

that *gnd* typing may not be suitable for STEC as described previously (Gilmour et al., 2007), which would hamper the utility of using *gnd* typing as a proxy for O-antigen typing.

CRISPR diversity within non-O157 STEC

CRISPRs have been used previously to characterize genomic diversity within a species. The diversity of CRISPR arrays within non-O157 STEC strains was explored to determine the suitability of CRISPRs as a typing method (Delannoy et al., 2012). *E. coli* and *Salmonella* can have two CRISPR loci, with the CRISPR associated (*cas*) genes at each locus variable and classified as either Ypest or *E. coli cas* subtypes (Randau et al., 2010). Up to two CRISPR arrays flank each *cas* loci at a specific insertion: for *E. coli* CRISPR/*cas* these are designated CRISPR1 (between *cysD-cysH*) and CRISPR2 (between *cysH-ygcF*); for Ypest CRISPR/*cas* subtypes these are designated CRISPR3 and CRISPR4 (between *clpS-aat*).

The conservation of spacer sequences was examined in the 44 STEC strains as well as the representative *E. coli* strains (Table 4). CRISPR spacers from the B1 phylogroup STEC strain spacer repertoire were also found in *E. coli* strains E24377A, B REL 606 and IAI1, but were not found in other *E. coli* (data not shown). In turn, spacer sequences from other *E. coli* did not appear in B1 phylogroup STEC strains.

Each locus varied in gene content and spacer sequences for all STEC strains, even among strains within the same lineage. All 44 STEC strains contained a CRISPR1 locus, with a total of 115 unique spacers identified across the B1 phylogroup. The majority of these were localized in clusters of 3 to 25 spacers, with a median of 8 spacers per CRISPR1 locus (Figure 7). There was little similarity in spacer content between STEC strains from different lineages. For example, no spacer sequences from LEE-positive strains were found in LEE-negative strains. Some variation of spacers within a particular lineage was observed and could be due to strain or lineage specific deletion of spacers. Some spacer sequences were consistent across strains of the same lineage and may represent potential genotyping targets to identify individual STEC lineages).

CRISPR2 arrays were detected in the majority of strains in this study, localized blocks comprising multiple clusters of 2 to 4 spacers, with a median value of 9 (Figure 8). The diversity of spacer sequences was variable within different lineages; strains n05, n01, n04, n38 (ST106A) contained common spacers but were distinct from other ST106A strains n03, n06, n50, n39, n40, n12 and n13. Strain n51 possessed a set of 9 spacer sequences that were not present in any other ST106A strains. Similarly, ST106B was separated into two groups, with strains n37 and n08 containing one set of spacers and strains n9, n10, n11 and n36 containing a different (but conserved) set of three spacers. Spacers were conserved within ST118, except for strain n17, which carried two spacers that were not detected in any other B1 phylogroup strain examined in this study. Some lineages showed a high degree of conservation of CRISPR2 spacer sequences with examples of step-wise spacer acquisition (ST234) or deletion (ST379). Further resolution of individual lineages will assist in determining the significance of spacer diversity within the STEC population.

In terms of CRISPR3 and CRISPR4, CRISPR3 was only detected in the n34 genome, with two arrays of 13 and 22 spacer sequences identified within the region between *clpA* and *infA*. Strain n34 also carried a full set of Ypest *cas* genes. Instead of the CRISPR3 array, strains n1, n4, n38, n39, n40, n12 and n13 carried a mobile genetic element encoding an

integrase, a number of genes encoding hypothetical proteins and genes encoding a YeeU/YeeB toxin/anti-toxin system. No CRISPR array was detected in the CRISPR3 loci for any other B1 phylogroup STEC strains examined in this study. Instead, these strains contained genes encoding a tRNA-Ser and a Translation initiation factor IF-1 at this locus. All non-O157 strains examined in this study lacked CRISPR4 and no novel CRISPR regions were found using whole genome detection of CRISPR arrays with either PILER-CR or CRISPRFinder.

Genomic features shared by STEC strains

Whole genome alignment was utilized to identify genomic regions conserved among STEC genomes. We identified 22 genomic regions that were significantly associated with STEC strains, eight of which were determined from the whole genome alignment as conserved across all STEC strains used in this dataset. These regions are summarized in Table 3, together with details of their location relative to the published 11368 (O26:H11) and EDL933 (O157:H7) genomes. Of the 22 regions identified by this analysis, three large regions (>5kb) were present across all twelve representative STEC strains, but not present in K12 MG1665 (Table 3, bold): (i) a 10,117bp region encoding the Type VI Secretion system as part of O-island 7; (ii) genes encoding CRISPR associated genes associated with the CRISPR1 region, and (iii) the second cryptic T3SS (ETT2) within O-island 115 (Ren et al., 2004).

The CRISPR1 *cas* loci could be classified into three variant alleles according to its similarity with representative regions from CFT073 (from which it is absent), MG1655 (K12), or 11368 (O26:H11) (Table 4). The *cas* genes in MG1655 and 11368 were divergent at the amino acid level, with 14% amino acid identity on average between orthologous proteins encoded by *casABCDE* (Supplementary Figure 3). The STEC O26:H11 *cas* variant was found in all strains within B1 phylogroup, including B1 phylogroup STEC and other *E. coli* such as ETEC E24377A, EAEC strain 55989 and commensal strains SE11 and IAI1. The STEC O26:H11 *cas* variant was also conserved in O157:H7 strains and n43, the ST182/O121 STEC strain that belongs to a divergent STEC lineage.

Discussion

Non-O157 STEC are increasingly recognized as an important food-borne pathogen responsible for global outbreaks (Bettelheim, 2007). While O157:H7 remains the dominant serotype in the United States, United Kingdom and Japan, STEC serogroups including O26, O111, O103, O121, O45,O128, O91 and O113 have also been associated with human disease (Gould et al., 2013, Johnson et al., 2006). Strains from these serogroups include both LEE positive and LEE negative variants (Newton et al., 2009, Mellmann et al., 2008). The 2011 outbreak, caused by an O104 LEE negative strain, led to more than 3,000 cases and more than 30 fatalities (Rohde et al., 2011). This particular strain appears to have descended from an EAEC progenitor strain, hence, it is sometimes referred to as Shiga toxigenic Enteroaggregative *E. coli* (STEAEC) (Mora et al., 2011). This outbreak highlights the need for wider studies into STEC other than the dominant O157:H7 serotype. To address this, we have analyzed the whole genome sequence of 44 non-O157 STEC strains from different origins to explore their overall diversity.

We aimed to examine the whole genome phylogeny of the most prevalent non-O157 STEC serogroups including LEE positive and LEE negative strains. It has been previously noted that STEC is comprised of multiple lineages that have evolved from parallel paths (Reid et al., 2000, Ogura et al., 2009, Steyert et al., 2012). The phylogenomic analyses presented here placed all strains within the B1 phylogroup (Girardeau et al., 2005) and confirmed the previously reported distinction between LEE positive O26, O111, and O103 strains (Ogura et al., 2009). We also demonstrate that LEE negative serogroups of clinical importance, namely O91, O113 and O128, form separate lineages within the B1 phylogroup. Our analysis of the LEE positive O121 strain showed that it was distinct from all other STEC strains, consistent with previous studies of O121 (Tarr et al., 2002).

Defined STEC lineages were largely concordant with EcMLST typing (Weihong et al., 2004)and we referred to lineages in accordance to that scheme to allow for a scalable nomenclature. ST106 correspond to the previously defined EHEC-2 clonal group, ST118 corresponds to the STEC-2 clonal group and ST182 correspond to the distant STEC-3 clonal group (Tarr et al., 2002, Iguchi et al., 2012, Abu-Ali et al., 2009). However, ST106 strains comprised two individual sub-lineages (Figure 1) and were designated as ST106A, which comprised H11 strains of both O26 and O111, and ST106B, which included O111 strains of flagella types other than H11. In this case, the housekeeping gene *mutS* from the EcMLST 15 allele scheme was able to distinguish between these two lineages. A number of single STEC lineages were also identified in this study with many more lineages expected as the genomes of further non-O157 STEC are sequenced.

In order to determine that the defined STEC lineages reflected the true ancestry of STEC and was not subject to recombination, gene sequences used to generate the phylogram were filtered using recombination tests implemented in PHIpack. PHIpack implements three different recombination tests, neighbour similarity score (NSS), maximal chi-squared (MaxChi) and the pairwise homoplasy index (PHI), each with different sensitivities and accuracies. In general, MaxChi and PHI are more accurate than NSS, however NSS is more sensitive (Chan et al., 2007). In order to dampen false calling by one particular method, a consensus of at least two methods was required to potentially show evidence of recombination (Chan et al., 2011). This approach reduced the number of genes from 2,153 to 1,136 genes. This implies that half the genes conserved across STEC were not of vertical descent. An avenue of further work would be to compare the effectiveness of recombination detection methods and determine where these genes are indeed recombinant by comparing these results with another detection method such as ClonalFrameML (Didelot and Wilson, 2015). Additionally, as horizontal gene transfer plays a major role in bacterial evolution it is possible that some of these genes are important to STEC evolution.

Serotyping is a standard classification method for *E. coli,* particular for STEC/EHEC. Typing schemes such as seropathotyping (Karmali et al., 2003) have linked disease potential of EHEC strains to particular serotypes based on epidemiological and prevalence studies. Previous studies have shown that certain serogroups (such as O174) are distributed across different phylogenetic backgrounds, consistent with lateral gene transfer of the O-antigen region (Tarr et al., 2008). Within our study of 44STEC strains, we observed several instances where O-antigen was inconsistent with the defined phylogeny, suggesting that lateral transfer of the O-antigen region may be more prevalent than first

suspected. In all cases, O-antigen typing was verified through sequence comparison of the O-antigen regions. Most notably, we could distinguish O111 strains based on flagella type that placed O111:H11 strains (n12 and n13) within the ST106A lineage, separate to other O111 strains in ST106B. The O111 antigen locus has already been associated with lateral transfer between O35 in *Salmonella (Wang and Reeves, 2000)*, and the mobility of this region would explain conflicting virulence profile results within O111 strains. Recent work has shown that other O26:H11 and O111:H11 strains share common ancestry, based on genomic content, (Ju et al., 2014). O103:H11 strains were also found to be more closely related to O26:H11 rather than other O103 strains (Iguchi et al., 2012), highlighting the need to examine the phylogenetic background of O111, O103 and O26 strains when comparing virulence factor profiles.

We also found that O91 strains in our study belonged to separate lineages within the determined phylogeny. A previous MLST-based study has shown that O91 strains of different sequence types had differential associations with HC or HUS (Mellmann et al., 2009). Our phylogenomic analysis shows that the acquisition of genes encoding O91 antigen has occurred through multiple independent events rather than acquisition by a common ancestor. Similarly, an O45 serogroup strain was found to be almost indistinguishable at the core genome level to three O103 strains, including the reference strain 12009 (Ogura et al., 2009).These results highlight the importance of a sequence-based genotyping approach combining lineage and virulence gene content for the routine identification of STEC strains.

Mobility of the O-antigen locus has been observed in interspecies comparisons between *Salmonella* and *E. coli (Wang and Reeves, 2000)*, which showed O-antigen synthesis regions in *Salmonella enterica* O35 and *E. coli* O111 have identical gene order and 78-88% nucleotide conservation. O-antigen mobility has also been observed in *Vibrio splendidus (Wildschutte et al., 2010)* and has been suggested to occur between *E. fergusonii* and *E. coli* O157:H7 (Fegan et al., 2006). Our data suggest that O-antigen mobility may also occur in STEC.

The T3SS encoded on the LEE is essential for the production of attaching and effacing lesions by EHEC and EPEC, and its presence is associated with clinically dominant STEC strains (Karmali et al., 2003).The acquisition of the LEE in EHEC and EPEC is considered a key evolutionary event for both pathotypes (Reid et al., 2000). LEE positive STEC strains with in this study were distributed across four lineages with differential insertion sites and Intimin subtypes (ST106A: Beta-*pheU*, ST106B: Theta-*pheV*, ST118: Epsilon-*pheV* and ST182: Epsilon-*pheU*). LEE positive STEC O26 was suggested to have arisen through step-wise evolution from atypical EPEC O26 strains, whereby the LEE was inserted into *pheU,* and the*stx1* or *stx2*genes were gained through phage integration on the chromosome (Bielaszewska et al., 2007). In contrast, O111 strains are thought to have independently acquired the LEE in *pheV(Tarr and Whittam, 2002, Rumer et al., 2003)*. Our data suggest that the LEE has been acquired within STEC on at least five individual occasions (O157, ST106A, ST106B, ST118 and ST182).The effector sequence profile of complete genomes from LEE positive strains suggest that LEE acquisition was accompanied by phage-mediated lateral gene transfer of a distinct, lineage-specific effector repertoire (Tobe et al., 2006, Ogura et al., 2009). Phage-associated effectors were not identified in any LEE negative strains, suggesting that LEE is necessary for selection

and maintenance of EELs. Intriguingly, we were able to discern several EELs in common between the different lineages, with different locations suggesting independent acquisition of common genetic elements. Furthermore, the observation of EELs with common locations between ST106A and ST106B which have separate LEE insertion sites suggest that the LEE may have been acquired prior to divergence of these lineages, followed by displacement with an independently acquired LEE pathogenicity island in one of these lineages after divergence. The future sequencing of multiple STEC genomes using long-read technologies such as PacBio should enable a detailed analysis of LGT events that have occurred along each STEC lineage.

Stx type was heterogeneous within ST89, ST234 and ST118 lineages. It is possible that the *stx* genes had been lost during cell culture or infection as observed in previous studies (Tarr et al., 1990, Mellmann et al., 2005). $Stx2d_{activatable}$ (elastase-cleaved), was only found in LEE negative STEC strains within the ST89 and ST379 lineages. In other studies, $Stx2d_{act}$ has been found to be prevalent within LEE negative strains (Tasara et al., 2008, Gobius et al., 2003).

CRISPR have been identified in a number of bacterial species (Jansen et al., 2002) and their distribution has been linked with phylogenetic grouping in *E. coli (Touchon et al., 2011)*. Recently, CRISPR loci were proposed as a method to distinguish between highly virulent STEC serotypes (O157:H7, O26:H11, O145:H28, O103:H2, O111:H8, O121:H19, and O45:H2), as the spacer sequences within these loci are unique to these STEC(Delannoy et al., 2012). However, *E. coli* pathovars and lineages in general have shown little association to CRISPR content (Jansen et al., 2002, Touchon et al., 2011). This was also reflected in this study, which showed that spacer repertoire does not distinguish STEC from non-STEC strains, suggesting that any CRISPR based scheme to define STEC from other *E. coli* would not be feasible without an additional method such as *stx*-typing. However, we did find that spacers within CRISPR1 could differentiate individual STEC clonal groups. Therefore, it may be possible to develop a scheme to differentiate individual STEC clonal groups using CRISPR1 loci.

In contrast to the diversity observed within *E. coli* CRISPR spacer sequences, the CRISPR associated genes (*cas*) within CRISPR1-2 were conserved across all but one STEC strain (STEC_7v) examined in this study, as well as previously published STEC complete and draft genomes from a range of *E. coli* phylogroups (E, B1, B2, D) (Table 4). Cas genes were *cas* subtype I-E based on gene order and chromosomal location, but closer inspection with sequence comparisons showed variation within *cas* genes such that *E. coli* strains could be defined as having either K-12-like, O26:H11-like or absent subtype I-E *cas* genes. CRISPR have been shown to have a distinct role in phage defense in *E. coli (Brouns et al., 2008)*, and the modification of CRISPRS/*cas* genes can alter susceptibility to infection in some species (Barrangou et al., 2007). Comparisons within *E. coli* have yielded little correlation between CRISPR arrays and resistance to foreign elements (Diez-Villasenor et al., 2010), but this does not incorporate the sequence variations noted here within the *cas* complex. Our results present the intriguing possibility that the type of *cas* genes carried by STEC is directly related to the observation that only particular *E. coli* lineages are able to acquire phage that carry the *stx* gene or genes encoding Type III effectors.

In conclusion, while pathogenic bacteria is usually discussed in the context of lineage

specific acquisition, here we present STEC as a set of diverse lineages evolving in parallel with independent acquisition of virulence factors. Using the phylogenomic analysis presented here, we have compared alternative typing methods such as serotyping, EcMLST and CRISPR typing, and found that each of these approaches have their own shortcomings in representing the divergent nature of STEC. Given, the varied distribution of STEC virulence, we also attempted to define the CRISPR-cas locus as a common genomic element that could provide the genomic background for Stx-encoding phage acquisition. The recombination analysis presented here suggests a high level of recombination amongst STEC and other E. coli, highlighting this as an avenue of inquiry to determine the genomic background that facilitates parallel evolution. While this work presents many broad findings, a more in-depth study of variation within and between defined STEC lineages is required to determine the distinction between categories such as host association and continental variation.

References

ABU-ALI, G. S., LACHER, D. W., WICK, L. M., QI, W. & WHITTAM, T. S. 2009. Genomic diversity of pathogenic Escherichia coli of the EHEC 2 clonal complex. *BMC genomics,* 10**,** 296.

ANGIUOLI, S. V. & SALZBERG, S. L. 2011. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics,* 27**,** 334-42.

ARCHER, C. T., KIM, J. F., JEONG, H., PARK, J. H., VICKERS, C. E., LEE, S. Y. & NIELSEN, L. K. 2011. The genome sequence of E. coli W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of E. coli. *BMC Genomics,* 12**,** 9.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol,* 19**,** 455-77.

BARRANGOU, R., FREMAUX, C., DEVEAU, H., RICHARDS, M., BOYAVAL, P., MOINEAU, S., ROMERO, D. A. & HORVATH, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science,* 315**,** 1709-12.

BERTIN, Y., BOUKHORS, K., LIVRELLI, V. & MARTIN, C. 2004. Localization of the insertion site and pathotype determination of the locus of enterocyte effacement of shiga toxin-producing *Escherichia coli* strains. *Appl Environ Microbiol,* 70**,** 61-8.

BETTELHEIM, K. A. 2007. The non-O157 shiga-toxigenic (verocytotoxigenic) *Escherichia coli*; under-rated pathogens. *Crit Rev Microbiol,* 33**,** 67-87.

BIELASZEWSKA, M., PRAGER, R., KOCK, R., MELLMANN, A., ZHANG, W., TSCHAPE, H., TARR, P. I. & KARCH, H. 2007. Shiga toxin gene loss and transfer in vitro and in vivo during enterohemorrhagic *Escherichia coli* O26 infection in humans. *Appl Environ Microbiol,* 73**,** 3144-50.

BLATTNER, F. R., PLUNKETT, G., 3RD, BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B. & SHAO, Y. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science,* 277**,** 1453-62.

BOSILEVAC, J. M. & 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**,** 2103-12.

BRETT, K. N., RAMACHANDRAN, V., HORNITZKY, M. A., BETTELHEIM, K. A.,

WALKER, M. J. & DJORDJEVIC, S. P. 2003. stx1c Is the most common Shiga toxin 1 subtype among Shiga toxin-producing Escherichia coli isolates from sheep but not among isolates from cattle. *J Clin Microbiol,* 41**,** 926-36.

BROOKS, J. T., SOWERS, E. G., WELLS, J. G., GREENE, K. D., GRIFFIN, P. M., HOEKSTRA, R. M. & STROCKBINE, N. A. 2005. Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983-2002. *J Infect Dis,* 192**,** 1422-9.

BROUNS, S. J., JORE, M. M., LUNDGREN, M., WESTRA, E. R., SLIJKHUIS, R. J., SNIJDERS, A. P., DICKMAN, M. J., MAKAROVA, K. S., KOONIN, E. V. & VAN DER OOST, J. 2008. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science,* 321**,** 960-4.

BRUEN, T. C., PHILIPPE, H. & BRYANT, D. 2006. A simple and robust statistical test for detecting the presence of recombination. *Genetics,* 172**,** 2665-81.

BRZUSZKIEWICZ, E., THURMER, A., SCHULDES, J., LEIMBACH, A., LIESEGANG, H., MEYER, F. D., BOELTER, J., PETERSEN, H., GOTTSCHALK, G. & DANIEL, R. 2011. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch Microbiol,* 193**,** 883-91.

CAMACHO, C., COULOURIS, G., AVAGYAN, V., MA, N., PAPADOPOULOS, J., BEALER, K. & MADDEN, T. L. 2009. BLAST+: architecture and applications. *BMC Bioinformatics,* 10**,** 421.

CARVER, T. J., RUTHERFORD, K. M., BERRIMAN, M., RAJANDREAM, M. A., BARRELL, B. G. & PARKHILL, J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics,* 21**,** 3422-3.

CENTERS FOR DISEASE, C. 1982. Isolation of *E. coli* O157:H7 from sporadic cases of hemorrhagic colitis - United States. *MMWR Morb Mortal Wkly Rep,* 31**,** 580, 585.

CERGOLE-NOVELLA, M. C., NISHIMURA, L. S., DOS SANTOS, L. F., IRINO, K., VAZ, T. M., BERGAMINI, A. M. & GUTH, B. E. 2007. Distribution of virulence profiles related to new toxins and putative adhesins in Shiga toxin-producing *Escherichia coli* isolated from diverse sources in Brazil. *FEMS Microbiol Lett,* 274**,** 329-34.

CHAN, C. X., BEIKO, R. G. & RAGAN, M. A. 2007. A two-phase strategy for detecting recombination in nucleotide sequences: reviewed article. *South African Computer Journal,* 38**,** 20-27.

CHAN, C. X., BEIKO, R. G. & RAGAN, M. A. 2011. Lateral Transfer of Genes and Gene Fragments in Staphylococcus Extends beyond Mobile Elements. *Journal of Bacteriology,* 193**,** 3964-3977.

CHEN, S. L., HUNG, C. S., XU, J., REIGSTAD, C. S., MAGRINI, V., SABO, A., BLASIAR, D., BIERI, T., MEYER, R. R., OZERSKY, P., ARMSTRONG, J. R., FULTON, R. S., LATREILLE, J. P., SPIETH, J., HOOTON, T. M., MARDIS, E. R., HULTGREN, S. J. & GORDON, J. I. 2006. Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proc Natl Acad Sci U S A,* 103**,** 5977-82.

CLARKE, D. J., CHAUDHURI, R. R., MARTIN, H. M., CAMPBELL, B. J., RHODES, J. M., CONSTANTINIDOU, C., PALLEN, M. J., LOMAN, N. J., CUNNINGHAM, A. F., BROWNING, D. F. & HENDERSON, I. R. 2011. Complete genome sequence of the Crohn's disease-associated adherent-invasive Escherichia coli strain HM605. *J Bacteriol,* 193**,** 4540.

CROXEN, M. A. & FINLAY, B. B. 2010. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol,* 8**,** 26-38.

DELANNOY, S., BEUTIN, L. & FACH, P. 2012. Use of clustered regularly interspaced short palindromic repeat sequence polymorphisms for specific detection of enterohemorrhagic *Escherichia coli* strains of serotypes O26:H11, O45:H2, O103:H2, O111:H8, O121:H19, O145:H28, and O157:H7 by real-time PCR. *J Clin Microbiol,* 50**,** 4035-40.

DENG, W., PUENTE, J. L., GRUENHEID, S., LI, Y., VALLANCE, B. A., VAZQUEZ, A.,

BARBA, J., IBARRA, J. A., O'DONNELL, P., METALNIKOV, P., ASHMAN, K., LEE, S., GOODE, D., PAWSON, T. & FINLAY, B. B. 2004. Dissecting virulence: systematic and functional analyses of a pathogenicity island. *Proc Natl Acad Sci U S A,* 101**,** 3597-602.

DIDELOT, X. & WILSON, D. J. 2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput Biol,* 11**,** e1004041.

DIEZ-VILLASENOR, C., ALMENDROS, C., GARCIA-MARTINEZ, J. & MOJICA, F. J. 2010. Diversity of CRISPR loci in *Escherichia coli*. *Microbiology,* 156**,** 1351-61.

DONNENBERG, M. S. & WHITTAM, T. S. 2001. Pathogenesis and evolution of virulence in enteropathogenic and enterohemorrhagic *Escherichia coli*. *J Clin Invest,* 107**,** 539-48.

DOWD, S. E., CRIPPEN, T. L., SUN, Y., GONTCHAROVA, V., YOUN, E., MUTHAIYAN, A., WOLCOTT, R. D., CALLAWAY, T. R. & RICKE, S. C. 2010. Microarray analysis and draft genomes of two *Escherichia coli* O157:H7 lineage II cattle isolates FRIK966 and FRIK2000 investigating lack of Shiga toxin expression. *Foodborne Pathog Dis,* 7**,** 763-73.

DURFEE, T., NELSON, R., BALDWIN, S., PLUNKETT, G., 3RD, BURLAND, V., MAU, B., PETROSINO, J. F., QIN, X., MUZNY, D. M., AYELE, M., GIBBS, R. A., CSORGO, B., POSFAI, G., WEINSTOCK, G. M. & BLATTNER, F. R. 2008. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. *J Bacteriol,* 190**,** 2597-606.

EASTON, D. M., TOTSIKA, M., ALLSOPP, L. P., PHAN, M. D., IDRIS, A., WURPEL, D. J., SHERLOCK, O., ZHANG, B., VENTURINI, C., BEATSON, S. A., MAHONY, T. J., COBBOLD, R. N. & SCHEMBRI, M. A. 2011. Characterization of EhaJ, a New Autotransporter Protein from Enterohemorrhagic and Enteropathogenic *Escherichia coli*. *Front Microbiol,* 2**,** 120.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

EDGAR, R. C. 2007. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics,* 8**,** 18.

EPPINGER, M., MAMMEL, M. K., LECLERC, J. E., RAVEL, J. & CEBULA, T. A. 2011. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. *Proc Natl Acad Sci U S A,* 108**,** 20142-7.

FABICH, A. J., LEATHAM, M. P., GRISSOM, J. E., WILEY, G., LAI, H., NAJAR, F., ROE, B. A., COHEN, P. S. & CONWAY, T. 2011. Genotype and phenotypes of an intestine-adapted *Escherichia coli* K-12 mutant selected by animal passage for superior colonization. *Infect Immun,* 79**,** 2430-9.

FEGAN, N., BARLOW, R. S. & GOBIUS, K. S. 2006. *Escherichia coli* O157 somatic antigen is present in an isolate of *E. fergusonii*. *Curr Microbiol,* 52**,** 482-6.

FENG, P., LAMPEL, K. A., KARCH, H. & WHITTAM, T. S. 1998. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. *J Infect Dis,* 177**,** 1750-3.

FERENCI, T., ZHOU, Z., BETTERIDGE, T., REN, Y., LIU, Y., FENG, L., REEVES, P. R. & WANG, L. 2009. Genomic sequencing reveals regulatory mutations and recombinational events in the widely used MC4100 lineage of *Escherichia coli* K-12. *J Bacteriol,* 191**,** 4025-9.

FRICKE, W. F., WRIGHT, M. S., LINDELL, A. H., HARKINS, D. M., BAKER-AUSTIN, C., RAVEL, J. & STEPANAUSKAS, R. 2008. Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. *J Bacteriol,* 190**,** 6779-94.

GARMENDIA, J., FRANKEL, G. & CREPIN, V. F. 2005. Enteropathogenic and Enterohemorrhagic *Escherichia coli* Infections: Translocation, Translocation, Translocation. *Infection and Immunity,* 73**,** 2573-2585.

GILMOUR, M. W., CHUI, L., CHIU, T., TRACZ, D. M., HAGEDORN, K., TSCHETTER, L., TABOR, H., NG, L. K. & LOUIE, M. 2009. Isolation and detection of Shiga toxin-producing Escherichia coli in clinical stool samples using conventional and molecular

methods. *J Med Microbiol,* 58**,** 905-11.
GILMOUR, M. W., OLSON, A. B., ANDRYSIAK, A. K., NG, L. K. & CHUI, L. 2007. Sequence-based typing of genetic targets encoded outside of the O-antigen gene cluster is indicative of Shiga toxin-producing *Escherichia coli* serogroup lineages. *J Med Microbiol,* 56**,** 620-8.
GIRARDEAU, J. P., DALMASSO, A., BERTIN, Y., DUCROT, C., BORD, S., LIVRELLI, V., VERNOZY-ROZAND, C. & MARTIN, C. 2005. Association of virulence genotype with phylogenetic background in comparison to different seropathotypes of Shiga toxin-producing *Escherichia coli* isolates. *J Clin Microbiol,* 43**,** 6098-107.
GOBIUS, K. S., HIGGS, G. M. & DESMARCHELIER, P. M. 2003. Presence of activatable Shiga toxin genotype (stx(2d)) in Shiga toxigenic *Escherichia coli* from livestock sources. *J Clin Microbiol,* 41**,** 3777-83.
GOULD, L. H., MODY, R. K., ONG, K. L., CLOGHER, P., CRONQUIST, A. B., GARMAN, K. N., LATHROP, S., MEDUS, C., SPINA, N. L., WEBB, T. H., WHITE, P. L., WYMORE, K., GIERKE, R. E., MAHON, B. E., GRIFFIN, P. M. & EMERGING INFECTIONS PROGRAM FOODNET WORKING, G. 2013. Increased recognition of non-O157 Shiga toxin-producing Escherichia coli infections in the United States during 2000-2010: epidemiologic features and comparison with E. coli O157 infections. *Foodborne Pathog Dis,* 10**,** 453-60.
GRAD, Y. H., LIPSITCH, M., FELDGARDEN, M., ARACHCHI, H. M., CERQUEIRA, G. C., FITZGERALD, M., GODFREY, P., HAAS, B. J., MURPHY, C. I., RUSS, C., SYKES, S., WALKER, B. J., WORTMAN, J. R., YOUNG, S., ZENG, Q., ABOUELLEIL, A., BOCHICCHIO, J., CHAUVIN, S., DESMET, T., GUJJA, S., MCCOWAN, C., MONTMAYEUR, A., STEELMAN, S., FRIMODT-MOLLER, J., PETERSEN, A. M., STRUVE, C., KROGFELT, K. A., BINGEN, E., WEILL, F. X., LANDER, E. S., NUSBAUM, C., BIRREN, B. W., HUNG, D. T. & HANAGE, W. P. 2012. Genomic epidemiology of the Escherichia coli O104:H4 outbreaks in Europe, 2011. *Proc Natl Acad Sci U S A,* 109**,** 3065-70.
GRISSA, I., VERGNAUD, G. & POURCEL, C. 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res,* 35**,** W52-7.
GUINDON, S., DUFAYARD, J. F., LEFORT, V., ANISIMOVA, M., HORDIJK, W. & GASCUEL, O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol,* 59**,** 307-21.
HAYASHI, T., MAKINO, K., OHNISHI, M., KUROKAWA, K., ISHII, K., YOKOYAMA, K., HAN, C. G., OHTSUBO, E., NAKAYAMA, K., MURATA, T., TANAKA, M., TOBE, T., IIDA, T., TAKAMI, H., HONDA, T., SASAKAWA, C., OGASAWARA, N., YASUNAGA, T., KUHARA, S., SHIBA, T., HATTORI, M. & SHINAGAWA, H. 2001. Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res,* 8**,** 11-22.
HEROLD, S., PATON, J. C. & PATON, A. W. 2009. Sab, a Novel Autotransporter of Locus of Enterocyte Effacement-Negative Shiga-Toxigenic *Escherichia coli* O113:H21, Contributes to Adherence and Biofilm Formation. *Infection and Immunity,* 77**,** 3234-3243.
HOCHHUT, B., WILDE, C., BALLING, G., MIDDENDORF, B., DOBRINDT, U., BRZUSZKIEWICZ, E., GOTTSCHALK, G., CARNIEL, E. & HACKER, J. 2006. Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536. *Mol Microbiol,* 61**,** 584-95.
HUSON, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics,* 14**,** 68-73.
IGUCHI, A., IYODA, S., OHNISHI, M. & GROUP, E. S. 2012. Molecular characterization reveals three distinct clonal groups among clinical shiga toxin-producing *Escherichia coli* strains of serogroup O103. *J Clin Microbiol,* 50**,** 2894-900.
IGUCHI, A., THOMSON, N. R., OGURA, Y., SAUNDERS, D., OOKA, T., HENDERSON, I. R., HARRIS, D., ASADULGHANI, M., KUROKAWA, K., DEAN, P., KENNY, B., QUAIL,

M. A., THURSTON, S., DOUGAN, G., HAYASHI, T., PARKHILL, J. & FRANKEL, G. 2009. Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. *J Bacteriol,* 191**,** 347-54.

JANSEN, R., EMBDEN, J. D., GAASTRA, W. & SCHOULS, L. M. 2002. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol,* 43**,** 1565-75.

JEONG, H., BARBE, V., LEE, C. H., VALLENET, D., YU, D. S., CHOI, S. H., COULOUX, A., LEE, S. W., YOON, S. H., CATTOLICO, L., HUR, C. G., PARK, H. S., SEGURENS, B., KIM, S. C., OH, T. K., LENSKI, R. E., STUDIER, F. W., DAEGELEN, P. & KIM, J. F. 2009. Genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3). *J Mol Biol,* 394**,** 644-52.

JERSE, A. E., YU, J., TALL, B. D. & KAPER, J. B. 1990. A genetic locus of enteropathogenic *Escherichia coli* necessary for the production of attaching and effacing lesions on tissue culture cells. *Proc Natl Acad Sci U S A,* 87**,** 7839-43.

JOHNSON, K. E., THORPE, C. M. & SEARS, C. L. 2006. The emerging clinical importance of non-O157 Shiga toxin-producing Escherichia coli. *Clin Infect Dis,* 43**,** 1587-95.

JU, W., RUMP, L., TORO, M., SHEN, J., CAO, G., ZHAO, S. & MENG, J. 2014. Pathogenicity islands in Shiga toxin-producing *Escherichia coli* O26, O103, and O111 isolates from humans and animals. *Foodborne Pathog Dis,* 11**,** 342-5.

KAPER, J. B., NATARO, J. P. & MOBLEY, H. L. 2004. Pathogenic escherichia coli. *Nature Reviews Microbiology,* 2**,** 123-140.

KARMALI, M. A., MASCARENHAS, M., SHEN, S., ZIEBELL, K., JOHNSON, S., REID-SMITH, R., ISAAC-RENTON, J., CLARK, C., RAHN, K. & KAPER, J. B. 2003. Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease. *J Clin Microbiol,* 41**,** 4930-40.

KARMALI, M. A., STEELE, B. T., PETRIC, M. & LIM, C. 1983. Sporadic cases of haemolytic-uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools. *Lancet,* 1**,** 619-20.

KONOWALCHUK, J., SPEIRS, J. I. & STAVRIC, S. 1977. Vero response to a cytotoxin of *Escherichia coli*. *Infect Immun,* 18**,** 775-9.

KOTEWICZ, M. L., MAMMEL, M. K., LECLERC, J. E. & CEBULA, T. A. 2008. Optical mapping and 454 sequencing of *Escherichia coli* O157 : H7 isolates linked to the US 2006 spinach-associated outbreak. *Microbiology,* 154**,** 3518-28.

KULASEKARA, B. R., JACOBS, M., ZHOU, Y., WU, Z., SIMS, E., SAENPHIMMACHAK, C., ROHMER, L., RITCHIE, J. M., RADEY, M., MCKEVITT, M., FREEMAN, T. L., HAYDEN, H., HAUGEN, E., GILLETT, W., FONG, C., CHANG, J., BESKHLEBNAYA, V., WALDOR, M. K., SAMADPOUR, M., WHITTAM, T. S., KAUL, R., BRITTNACHER, M. & MILLER, S. I. 2009. Analysis of the genome of the Escherichia coli O157:H7 2006 spinach-associated outbreak isolate indicates candidate genes that may enhance virulence. *Infect Immun,* 77**,** 3713-21.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biol,* 5**,** R12.

L'ABEE-LUND, T. M., JORGENSEN, H. J., O'SULLIVAN, K., BOHLIN, J., LIGARD, G., GRANUM, P. E. & LINDBACK, T. 2012. The highly virulent 2006 Norwegian EHEC O103:H25 outbreak strain is related to the 2011 German O104:H4 outbreak strain. *PLoS One,* 7**,** e31413.

LEVINE, M. M., BLACK, R. E., CLEMENTS, M. L., YOUNG, C. R., CHENEY, C. P., SCHAD, P., COLLINS, H. & BOEDEKER, E. C. 1984. Prevention of enterotoxigenic Escherichia coli diarrheal infection in man by vaccines that stimulate anti-adhesion (anti-pili) immunity. *Attachment of organisms to the gut mucosa,* 2**,** 223-244.

LEYTON, D. L., SLOAN, J., HILL, R. E., DOUGHTY, S. & HARTLAND, E. L. 2003. Transfer

Region of pO113 from Enterohemorrhagic *Escherichia coli*: Similarity with R64 and Identification of a Novel Plasmid-Encoded Autotransporter, EpeA. *Infection and Immunity,* 71**,** 6307-6319.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LUKJANCENKO, O., WASSENAAR, T. M. & USSERY, D. W. 2010. Comparison of 61 sequenced Escherichia coli genomes. *Microb Ecol,* 60**,** 708-20.

MCDANIEL, T. K., JARVIS, K. G., DONNENBERG, M. S. & KAPER, J. B. 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. *Proceedings of the National Academy of Sciences,* 92**,** 1664-1668.

MELLMANN, A., BIELASZEWSKA, M., KOCK, R., FRIEDRICH, A. W., FRUTH, A., MIDDENDORF, B., HARMSEN, D., SCHMIDT, M. A. & KARCH, H. 2008. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic Escherichia coli. *Emerg Infect Dis,* 14**,** 1287-90.

MELLMANN, A., BIELASZEWSKA, M., ZIMMERHACKL, L. B., PRAGER, R., HARMSEN, D., TSCHAPE, H. & KARCH, H. 2005. Enterohemorrhagic Escherichia coli in human infection: in vivo evolution of a bacterial pathogen. *Clin Infect Dis,* 41**,** 785-92.

MELLMANN, A., FRUTH, A., FRIEDRICH, A. W., WIELER, L. H., HARMSEN, D., WERBER, D., MIDDENDORF, B., BIELASZEWSKA, M. & KARCH, H. 2009. Phylogeny and disease association of Shiga toxin-producing *Escherichia coli* O91. *Emerg Infect Dis,* 15**,** 1474-7.

MELLMANN, A., HARMSEN, D., CUMMINGS, C. A., ZENTZ, E. B., LEOPOLD, S. R., RICO, A., PRIOR, K., SZCZEPANOWSKI, R., JI, Y., ZHANG, W., MCLAUGHLIN, S. F., HENKHAUS, J. K., LEOPOLD, B., BIELASZEWSKA, M., PRAGER, R., BRZOSKA, P. M., MOORE, R. L., GUENTHER, S., ROTHBERG, J. M. & KARCH, H. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One,* 6**,** e22751.

MICHINO, H., ARAKI, K., MINAMI, S., TAKAYA, S., SAKAI, N., MIYAZAKI, M., ONO, A. & YANAGAWA, H. 1999. Massive outbreak of Escherichia coli O157:H7 infection in schoolchildren in Sakai City, Japan, associated with consumption of white radish sprouts. *Am J Epidemiol,* 150**,** 787-96.

MORA, A., HERRRERA, A., LOPEZ, C., DAHBI, G., MAMANI, R., PITA, J. M., ALONSO, M. P., LLOVO, J., BERNARDEZ, M. I., BLANCO, J. E., BLANCO, M. & BLANCO, J. 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**,** 121-41.

MORA, A., LOPEZ, C., DHABI, G., LOPEZ-BECEIRO, A. M., FIDALGO, L. E., DIAZ, E. A., MARTINEZ-CARRASCO, C., MAMANI, R., HERRERA, A., BLANCO, J. E., BLANCO, M. & BLANCO, J. 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**,** 2578-85.

NATARO, J. P. & KAPER, J. B. 1998. Diarrheagenic escherichia coli. *Clinical microbiology reviews,* 11**,** 142-201.

NEWELL, D. G., KOOPMANS, M., VERHOEF, L., DUIZER, E., AIDARA-KANE, A., SPRONG, H., OPSTEEGH, M., LANGELAAR, M., THREFALL, J., SCHEUTZ, F., VAN DER GIESSEN, J. & KRUSE, H. 2010. Food-borne diseases - the challenges of 20 years ago still persist while new ones continue to emerge. *Int J Food Microbiol,* 139 Suppl 1**,** S3-15.

NEWTON, H. J., SLOAN, J., BULACH, D. M., SEEMANN, T., ALLISON, C. C., TAUSCHEK, M., ROBINS-BROWNE, R. M., PATON, J. C., WHITTAM, T. S., PATON, A. W. & HARTLAND, E. L. 2009. Shiga toxin-producing *Escherichia coli* strains negative for locus of enterocyte effacement. *Emerg Infect Dis,* 15**,** 372-80.

NIE, H., YANG, F., ZHANG, X., YANG, J., CHEN, L., WANG, J., XIONG, Z., PENG, J., SUN,

L., DONG, J., XUE, Y., XU, X., CHEN, S., YAO, Z., SHEN, Y. & JIN, Q. 2006. Complete genome sequence of *Shigella flexneri* 5b and comparison with *Shigella flexneri* 2a. *BMC Genomics,* 7**,** 173.

O'BRIEN, A. D., LAVECK, G. D., THOMPSON, M. R. & FORMAL, S. B. 1982. Production of Shigella dysenteriae type 1-like cytotoxin by Escherichia coli. *J Infect Dis,* 146**,** 763-9.

OGURA, Y., OOKA, T., IGUCHI, A., TOH, H., ASADULGHANI, M., OSHIMA, K., KODAMA, T., ABE, H., NAKAYAMA, K., KUROKAWA, K., TOBE, T., HATTORI, M. & HAYASHI, T. 2009. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. *Proc Natl Acad Sci U S A,* 106**,** 17939-44.

OSHIMA, K., TOH, H., OGURA, Y., SASAMOTO, H., MORITA, H., PARK, S. H., OOKA, T., IYODA, S., TAYLOR, T. D., HAYASHI, T., ITOH, K. & HATTORI, M. 2008. Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. *DNA Res,* 15**,** 375-86.

PADDOCK, Z., SHI, X., BAI, J. & NAGARAJA, T. G. 2012. Applicability of a multiplex PCR to detect O26, O45, O103, O111, O121, O145, and O157 serogroups of *Escherichia coli* in cattle feces. *Vet Microbiol,* 156**,** 381-8.

PATON, A. W. & PATON, J. C. 2005. Multiplex PCR for direct detection of Shiga toxigenic *Escherichia coli* strains producing the novel subtilase cytotoxin. *J Clin Microbiol,* 43**,** 2944-7.

PATON, A. W., SRIMANOTE, P., TALBOT, U. M., WANG, H. & PATON, J. C. 2004. A new family of potent AB(5) cytotoxins produced by Shiga toxigenic *Escherichia coli. J Exp Med,* 200**,** 35-46.

PATON, A. W., SRIMANOTE, P., WOODROW, M. C. & PATON, J. C. 2001. Characterization of Saa, a novel autoagglutinating adhesin produced by locus of enterocyte effacement-negative Shiga-toxigenic *Escherichia coli* strains that are virulent for humans. *Infect Immun,* 69**,** 6999-7009.

PERNA, N. T., PLUNKETT, G., 3RD, BURLAND, V., MAU, B., GLASNER, J. D., ROSE, D. J., MAYHEW, G. F., EVANS, P. S., GREGOR, J., KIRKPATRICK, H. A., POSFAI, G., HACKETT, J., KLINK, S., BOUTIN, A., SHAO, Y., MILLER, L., GROTBECK, E. J., DAVIS, N. W., LIM, A., DIMALANTA, E. T., POTAMOUSIS, K. D., APODACA, J., ANANTHARAMAN, T. S., LIN, J., YEN, G., SCHWARTZ, D. C., WELCH, R. A. & BLATTNER, F. R. 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature,* 409**,** 529-33.

PETERLONGO, P. & CHIKHI, R. 2012. Mapsembler, targeted and micro assembly of large NGS datasets on a desktop computer. *BMC Bioinformatics,* 13**,** 48.

RANDAU, L., TOUCHON, M. & ROCHA, E. P. C. 2010. The Small, Slow and Specialized CRISPR and Anti-CRISPR of *Escherichia* and *Salmonella*. *PLoS ONE,* 5**,** e11126.

RASKO, D. A., ROSOVITZ, M. J., MYERS, G. S., MONGODIN, E. F., FRICKE, W. F., GAJER, P., CRABTREE, J., SEBAIHIA, M., THOMSON, N. R., CHAUDHURI, R., HENDERSON, I. R., SPERANDIO, V. & RAVEL, J. 2008. The pangenome structure of Escherichia coli: comparative genomic analysis of E. coli commensal and pathogenic isolates. *J Bacteriol,* 190**,** 6881-93.

RASKO, D. A., WEBSTER, D. R., SAHL, J. W., BASHIR, A., BOISEN, N., SCHEUTZ, F., PAXINOS, E. E., SEBRA, R., CHIN, C. S., ILIOPOULOS, D., KLAMMER, A., PELUSO, P., LEE, L., KISLYUK, A. O., BULLARD, J., KASARSKIS, A., WANG, S., EID, J., RANK, D., REDMAN, J. C., STEYERT, S. R., FRIMODT-MOLLER, J., STRUVE, C., PETERSEN, A. M., KROGFELT, K. A., NATARO, J. P., SCHADT, E. E. & WALDOR, M. K. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. *N Engl J Med,* 365**,** 709-17.

REEVES, P. R., LIU, B., ZHOU, Z., LI, D., GUO, D., REN, Y., CLABOTS, C., LAN, R., JOHNSON, J. R. & WANG, L. 2011. Rates of mutation and host transmission for an

Escherichia coli clone over 3 years. *PLoS One,* 6**,** e26907.
REID, S. D., HERBELIN, C. J., BUMBAUGH, A. C., SELANDER, R. K. & WHITTAM, T. S. 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature,* 406**,** 64-7.
REN, C. P., CHAUDHURI, R. R., FIVIAN, A., BAILEY, C. M., ANTONIO, M., BARNES, W. M. & PALLEN, M. J. 2004. The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. *J Bacteriol,* 186**,** 3547-60.
RILEY, L. W., REMIS, R. S., HELGERSON, S. D., MCGEE, H. B., WELLS, J. G., DAVIS, B. R., HEBERT, R. J., OLCOTT, E. S., JOHNSON, L. M., HARGRETT, N. T., BLAKE, P. A. & COHEN, M. L. 1983. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N Engl J Med,* 308**,** 681-5.
RISSMAN, A. I., MAU, B., BIEHL, B. S., DARLING, A. E., GLASNER, J. D. & PERNA, N. T. 2009. Reordering contigs of draft genomes using the Mauve aligner. *Bioinformatics,* 25**,** 2071-3.
ROHDE, H., QIN, J., CUI, Y., LI, D., LOMAN, N. J., HENTSCHKE, M., CHEN, W., PU, F., PENG, Y., LI, J., XI, F., LI, S., LI, Y., ZHANG, Z., YANG, X., ZHAO, M., WANG, P., GUAN, Y., CEN, Z., ZHAO, X., CHRISTNER, M., KOBBE, R., LOOS, S., OH, J., YANG, L., DANCHIN, A., GAO, G. F., SONG, Y., LI, Y., YANG, H., WANG, J., XU, J., PALLEN, M. J., WANG, J., AEPFELBACHER, M., YANG, R. & CONSORTIUM, E. C. O. H. G. A. C.-S. 2011. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. *N Engl J Med,* 365**,** 718-24.
RUMER, L., JORES, J., KIRSCH, P., CAVIGNAC, Y., ZEHMKE, K. & WIELER, L. H. 2003. Dissemination of pheU- and pheV-located genomic islands among enteropathogenic (EPEC) and enterohemorrhagic (EHEC) E. coli and their possible role in the horizontal transfer of the locus of enterocyte effacement (LEE). *Int J Med Microbiol,* 292**,** 463-75.
RUMP, L. V., STRAIN, E. A., CAO, G., ALLARD, M. W., FISCHER, M., BROWN, E. W. & GONZALEZ-ESCALONA, N. 2011. Draft genome sequences of six Escherichia coli isolates from the stepwise model of emergence of Escherichia coli O157:H7. *J Bacteriol,* 193**,** 2058-9.
SAHL, J. W., STEINSLAND, H., REDMAN, J. C., ANGIUOLI, S. V., NATARO, J. P., SOMMERFELT, H. & RASKO, D. A. 2011. A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. *Infect Immun,* 79**,** 950-60.
SCHARFF, R. L. 2012. Economic burden from health losses due to foodborne illness in the United States. *J Food Prot,* 75**,** 123-31.
SCHEUTZ, F., TEEL, L. D., BEUTIN, L., PIERARD, D., BUVENS, G., KARCH, H., MELLMANN, A., CAPRIOLI, A., TOZZOLI, R., MORABITO, S., STROCKBINE, N. A., MELTON-CELSA, A. R., SANCHEZ, M., PERSSON, S. & O'BRIEN, A. D. 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. *J Clin Microbiol,* 50**,** 2951-63.
SCHMIEDER, R. & EDWARDS, R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics,* 27**,** 863-4.
SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-9.
SHEN, S., MASCARENHAS, M., RAHN, K., KAPER, J. B. & KARMALI, M. A. 2004. Evidence for a hybrid genomic island in verocytotoxin-producing *Escherichia coli* CL3 (serotype O113:H21) containing segments of EDL933 (serotype O157:H7) O islands 122 and 48. *Infect Immun,* 72**,** 1496-503.
SHEPARD, S. M., DANZEISEN, J. L., ISAACSON, R. E., SEEMANN, T., ACHTMAN, M. & JOHNSON, T. J. 2012. Genome sequences and phylogenetic analysis of K88- and F18-positive porcine enterotoxigenic *Escherichia coli*. *J Bacteriol,* 194**,** 395-405.
SMITH, J. L., FRATAMICO, P. M. & GUNTHER, N. 2013. Shiga Toxin-Producing Escherichia coli. *Advances in applied microbiology,* 86**,** 145.

STAMATAKIS, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

STEYERT, S. R., SAHL, J. W., FRASER, C. M., TEEL, L. D., SCHEUTZ, F. & RASKO, D. A. 2012. Comparative genomics and stx phage characterization of LEE-negative Shiga toxin-producing Escherichia coli. *Front Cell Infect Microbiol,* 2**,** 133.

SULLIVAN, M. J., PETTY, N. K. & BEATSON, S. A. 2011. Easyfig: a genome comparison visualizer. *Bioinformatics,* 27**,** 1009-10.

TAN, C., XU, Z., ZHENG, H., LIU, W., TANG, X., SHOU, J., WU, B., WANG, S., ZHAO, G. P. & CHEN, H. 2011. Genome sequence of a porcine extraintestinal pathogenic *Escherichia coli* strain. *J Bacteriol,* 193**,** 5038.

TARR, C. L., LARGE, T. M., MOELLER, C. L., LACHER, D. W., TARR, P. I., ACHESON, D. W. & WHITTAM, T. S. 2002. Molecular characterization of a serotype O121:H19 clone, a distinct Shiga toxin-producing clone of pathogenic Escherichia coli. *Infect Immun,* 70**,** 6853-9.

TARR, C. L., NELSON, A. M., BEUTIN, L., OLSEN, K. E. & WHITTAM, T. S. 2008. Molecular characterization reveals similar virulence gene content in unrelated clonal groups of *Escherichia coli* of serogroup O174 (OX3). *J Bacteriol,* 190**,** 1344-9.

TARR, C. L. & WHITTAM, T. S. 2002. Molecular evolution of the intimin gene in O111 clones of pathogenic *Escherichia coli*. *J Bacteriol,* 184**,** 479-87.

TARR, P. I., BILGE, S. S., VARY, J. C., JR., JELACIC, S., HABEEB, R. L., WARD, T. R., BAYLOR, M. R. & BESSER, T. E. 2000. Iha: a novel *Escherichia coli* O157:H7 adherence-conferring molecule encoded on a recently acquired chromosomal island of conserved structure. *Infect Immun,* 68**,** 1400-7.

TARR, P. I., NEILL, M. A., CLAUSEN, C. R., WATKINS, S. L., CHRISTIE, D. L. & HICKMAN, R. O. 1990. *Escherichia coli* O157:H7 and the hemolytic uremic syndrome: importance of early cultures in establishing the etiology. *J Infect Dis,* 162**,** 553-6.

TASARA, T., BIELASZEWSKA, M., NITZSCHE, S., KARCH, H., ZWEIFEL, C. & STEPHAN, R. 2008. Activatable Shiga toxin 2d (Stx2d) in STEC strains isolated from cattle and sheep at slaughter. *Vet Microbiol,* 131**,** 199-204.

TOBE, T., BEATSON, S. A., TANIGUCHI, H., ABE, H., BAILEY, C. M., FIVIAN, A., YOUNIS, R., MATTHEWS, S., MARCHES, O., FRANKEL, G., HAYASHI, T. & PALLEN, M. J. 2006. An extensive repertoire of type III secretion effectors in *Escherichia coli* O157 and the role of lambdoid phages in their dissemination. *Proc Natl Acad Sci U S A,* 103**,** 14941-6.

TOMA, C., MARTINEZ ESPINOSA, E., SONG, T., MILIWEBSKY, E., CHINEN, I., IYODA, S., IWANAGA, M. & RIVAS, M. 2004. Distribution of putative adhesins in different seropathotypes of Shiga toxin-producing *Escherichia coli*. *J Clin Microbiol,* 42**,** 4937-46.

TOTSIKA, M., BEATSON, S. A., SARKAR, S., PHAN, M. D., PETTY, N. K., BACHMANN, N., SZUBERT, M., SIDJABAT, H. E., PATERSON, D. L., UPTON, M. & SCHEMBRI, M. A. 2011. Insights into a multidrug resistant *Escherichia coli* pathogen of the globally disseminated ST131 lineage: genome analysis and virulence mechanisms. *PLoS One,* 6**,** e26578.

TOUCHON, M., CHARPENTIER, S., CLERMONT, O., ROCHA, E. P., DENAMUR, E. & BRANGER, C. 2011. CRISPR distribution within the *Escherichia coli* species is not suggestive of immunity-associated diversifying selection. *J Bacteriol,* 193**,** 2460-7.

TOUCHON, M., HOEDE, C., TENAILLON, O., BARBE, V., BAERISWYL, S., BIDET, P., BINGEN, E., BONACORSI, S., BOUCHIER, C., BOUVET, O., CALTEAU, A., CHIAPELLO, H., CLERMONT, O., CRUVEILLER, S., DANCHIN, A., DIARD, M., DOSSAT, C., KAROUI, M. E., FRAPY, E., GARRY, L., GHIGO, J. M., GILLES, A. M., JOHNSON, J., LE BOUGUENEC, C., LESCAT, M., MANGENOT, S., MARTINEZ-JEHANNE, V., MATIC, I., NASSIF, X., OZTAS, S., PETIT, M. A., PICHON, C., ROUY, Z., RUF, C. S., SCHNEIDER, D., TOURRET, J., VACHERIE, B., VALLENET, D., MEDIGUE, C., ROCHA, E. P. & DENAMUR, E. 2009. Organised genome dynamics in the

Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet,* 5**,** e1000344.
TURNER, S. M., CHAUDHURI, R. R., JIANG, Z. D., DUPONT, H., GYLES, C., PENN, C. W., PALLEN, M. J. & HENDERSON, I. R. 2006. Phylogenetic comparisons reveal multiple acquisitions of the toxin genes by enterotoxigenic *Escherichia coli* strains of different evolutionary lineages. *J Clin Microbiol,* 44**,** 4528-36.
WANG, L. & REEVES, P. R. 2000. The Escherichia coli O111 and Salmonella enterica O35 gene clusters: gene clusters encoding the same colitose-containing O antigen are highly conserved. *J Bacteriol,* 182**,** 5256-61.
WEIHONG, Q., LACHER, D. W., BUMBAUGH, A. C., HYMA, K. E., OUELLETTE, L. M., LARGE, T. M., TARR, C. L. & WHITTAM, T. S. 2004. EcMLST: an online database for multi locus sequence typing of pathogenic escherichia coli. 499-500.
WELCH, R. A., BURLAND, V., PLUNKETT, G., 3RD, REDFORD, P., ROESCH, P., RASKO, D., BUCKLES, E. L., LIOU, S. R., BOUTIN, A., HACKETT, J., STROUD, D., MAYHEW, G. F., ROSE, D. J., ZHOU, S., SCHWARTZ, D. C., PERNA, N. T., MOBLEY, H. L., DONNENBERG, M. S. & BLATTNER, F. R. 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A,* 99**,** 17020-4.
WICK, L. M., QI, W., LACHER, D. W. & WHITTAM, T. S. 2005. Evolution of genomic content in the stepwise emergence of Escherichia coli O157:H7. *J Bacteriol,* 187**,** 1783-91.
WILDSCHUTTE, H., PREHEIM, S. P., HERNANDEZ, Y. & POLZ, M. F. 2010. O-antigen diversity and lateral transfer of the wbe region among Vibrio splendidus isolates. *Environ Microbiol,* 12**,** 2977-87.
XIONG, Y., WANG, P., LAN, R., YE, C., WANG, H., REN, J., JING, H., WANG, Y., ZHOU, Z., BAI, X., CUI, Z., LUO, X., ZHAO, A., WANG, Y., ZHANG, S., SUN, H., WANG, L. & XU, J. 2012. A novel *Escherichia coli* O157:H7 clone causing a major hemolytic uremic syndrome outbreak in China. *PLoS One,* 7**,** e36144.
YANG, F., YANG, J., ZHANG, X., CHEN, L., JIANG, Y., YAN, Y., TANG, X., WANG, J., XIONG, Z., DONG, J., XUE, Y., ZHU, Y., XU, X., SUN, L., CHEN, S., NIE, H., PENG, J., XU, J., WANG, Y., YUAN, Z., WEN, Y., YAO, Z., SHEN, Y., QIANG, B., HOU, Y., YU, J. & JIN, Q. 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res,* 33**,** 6445-58.
YANG, H., LIAO, Y., WANG, B., LIN, Y. & PAN, L. 2011a. Draft genome sequence of Escherichia coli XH001, a producer of L-threonine in industry. *J Bacteriol,* 193**,** 6406-7.
YANG, H., LIAO, Y., WANG, B., LIN, Y. & PAN, L. 2011b. Genome sequence of *Escherichia coli* XH140A, which produces L-threonine. *J Bacteriol,* 193**,** 6090-1.
YI, H., CHO, Y. J., HUR, H. G. & CHUN, J. 2011. Genome sequence of *Escherichia coli* AA86, isolated from cow feces. *J Bacteriol,* 193**,** 3681.
ZHOU, Z., LI, X., LIU, B., BEUTIN, L., XU, J., REN, Y., FENG, L., LAN, R., REEVES, P. R. & WANG, L. 2010. Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. *PLoS One,* 5**,** e8700.

Table 1. STEC strains characterized in this study

N#	ST[1]	Provider	Source	O-type	H Type	*stx1* type	*stx2* type	LEE	*eae* type
1	106	QHFSS	human	O26	H11	*stx1a*	--	+	Beta
2	461	QHFSS	human	O91	H10	--	*stx2d*	--	--
3	106	QHFSS	human	O26	H11	*stx1a*	--	+	Beta
4	106	QHFSS	human	O26	H11	*stx1a*	--	+	Beta
5	106	QHFSS	human	O26	H11	*stx1a*	--	+	Beta
6	106	QHFSS	human	O26	H11	*stx1a*	--	+	Beta
7	461	QHFSS	human	O91	H10	--	*stx2d*	--	--
8	106	QHFSS	human	O111	H-	*stx1a*	*stx2a*	+	Theta
9	106	QHFSS	human	O111	H-	*stx1a*	*stx2a*	+	Theta
10	106	QHFSS	human	O111	H8	*stx1a*	*stx2a*	+	Theta
11	106	QHFSS	human	O111	H8	*stx1a*	*stx2a*	+	Theta
12	106	QHFSS	human	O111	H11	*stx1a*	--	+	Beta
13	106	QHFSS	human	O111	H11	*stx1a*	--	+	Beta
15	234	QHFSS	human	O113	H21	--	*stx2d*	--	--
16	379	QHFSS	human	O128	H2	*stx1c*	*stx2d*$_{act}$	--	--
17	118	QHFSS	human	O103	H2	--	*stx2a*	+	Epsilon
19	130	CSIRO	bovine feces	O?	H2	--	*stx2c*	--	--
20	234	CSIRO	retail beef	O113	H21	--	*stx2a* + *stx2d*	--	--
21	234	CSIRO	bovine feces	O113	H21	--	*stx2a* + *stx2d*	--	--
22	234	CSIRO	Milk	O113	H21	--	*stx2d*	--	--
23	379	CSIRO	ovine feces	O128	H2	*stx1c*	*stx2d*$_{act}$	--	--
24	379	CSIRO	retail beef	O128	H2	*stx1c*	*stx2d*$_{act}$	--	--
25	379	CSIRO	ovine carcass	O128	H2	*stx1c*	*stx2d*$_{act}$	--	--
26	379	CSIRO	ovine feces	O128	H2	*stx1c*	*stx2d*$_{act}$	--	--
27	379	CSIRO	ovine carcass	O128	NM	*stx1c*	*stx2d*$_{act}$	--	--
28	89	CSIRO	retail beef	O91	H21	--	*stx2d*$_{act}$	--	--
29	89	CSIRO	bovine feces	O91	H21	--	*stx2d*$_{act}$	--	--
30	89	CSIRO	bovine feces	O91	H21	*stx1a*	*stx2d*$_{act}$	--	--
32	815	CSIRO	retail beef	O91	NM	*stx1a*	*stx2d*$_{act}$	--	--
34	650	CSIRO	bovine feces	O?	H7	--	*stx2a* + *stx2d*	--	--
36	106	CSIRO	bovine feces	O111	NM	*stx1a*	*stx2a*	+	Theta
37	106	CSIRO	bovine feces	O111	H8	*stx1a*	*stx2a*	+	Theta
38	106	CSIRO	bovine feces	O26	H11	*stx1a*	--	+	Beta
39	106	CSIRO	bovine feces	O26	H11	*stx1a*	--	+	Beta
40	106	CSIRO	bovine feces	O26	H11	*stx1a*	--	+	Beta
41	89	CSIRO	bovine hide	O?	H21	--	*stx2d*$_{act}$	--	--
42	106	CSIRO	bovine feces	O26	NM	*stx1a*	--	+	Beta
43	182	CSIRO	human	O121	H?[2]	*stx1a*	--	+	Epsilon
44	118	CSIRO	human	O45	H2	*stx1a*	--	+	Epsilon

46	234	WSU	bovine feces	O113	H21	--	*stx2a*	--	--
47	888	WSU	bovine feces	O?	H?[2]	*stx1a*	*stx2d*$_{act}$	--	--
48	118	WSU	cattle water	O103	H2	*stx1a*	--	+	Epsilon
50	106	WSU	bovine feces	O26	H11	*stx1a*	--	+	Beta
51	106	WSU	bovine feces	O26	H11	*stx1a*	--	+	Beta

[1]Sequence Type according to EcMLST 7 allele scheme (Weihong et al., 2004)
[2]n47 and n43 have identical flagella types (but unknown)

Table 2 CRISPR spacer sequences conserved within each lineage

Sequence type	Spacer ID	Spacer sequence
ST106A	5	GCGTATCGTCTCGTTATTGCGCCGCCCCAACT
	6	GGCGTTTTTGACTGTACGAATCCCTGCGCCGC
	7	GGATCTGCAGGCGATGAATTACCGCTTGACTA
	8	TCTACGTGAAGAATATTTGCAACACCCGCAAGAA
ST106B	5	GCGTATCGTCTCGTTATTGCGCCGCCCCAACT
	7	GGATCTGCAGGCGATGAATTACCGCTTGACTA
	10	ACAATCGTGTGTAAATTCGCGGCGGCTCCACTGG
	12	ACACACTATCCGGGCGGTATTACGCCCAAATATC
ST118	5	GCGTATCGTCTCGTTATTGCGCCGCCCCAACT
	6	GGCGTTTTTGACTGTACGAATCCCTGCGCCGC
	13	ACCTGCCGGGTGAAACCACTCGCGGCAGATCTTG
ST89	5	GCGTATCGTCTCGTTATTGCGCCGCCCCAACT
	21	ACACAATCGTGTGTAAATTCGCGGCGGCTCCACTGG
	22	ACACTGGTCGAAATATAGACAGCATGTTCCGTACCA
	25	ACACACTACTGTCGGTAGCTGGGAGGATGAGGAGAT
	31	ACACTGGTCGAAATATAGACAGCATGTTCCGTACCAC
	32	ACACACACTATCCGGGCGGTATTACGCCCAAATATCC
	33	ACACCTGCCGGGTGAAACCACTCGCGGCAGATCTTGC
	35	ACTCCAACCTTCCATGAGATACGCGCATTAGCGG
	36	ACCGTGACCGCCTGTACACGCTGTAATGGCTCAC
	37	AACGGAGCTCTACGTGAAGAATATTTGCAACACCCGCAAGAA
	38	CGTGACCGCCTGTACACGCTGTAATGGCTCAC
	39	CGTGACCGCCTGTACACGCTGTAATGGCTCAC
	40	CTGCCGGGTGAAACCACTCGCGGCAGATCTTG
	4	AGGGCCGCCGCTACCCCAGAAGTGCCACTCCC
	42	TAATCACGTTTTAGCGCGCCCTCGTCCGGTTT
	43	ATCACGATAACGCTGCTGTGATTCGTCCCCGT
ST234	89	CGTTCCTCGCATTATTTCCCTTTCTTCTTCGAC
	90	GTCGGCAGAGAATCGTTCGATTGCCCCTACATC
	91	ATCAACGTTATCGATTACAACTGACAGGGAGCC
ST461	21	ACACAATCGTGTGTAAATTCGCGGCGGCTCCACTGG
	25	ACACACTACTGTCGGTAGCTGGGAGGATGAGGAGAT
	44	AAGACGACGTGATCCGCAAAGTCGAAGGCACG
	72	CAAACAGGTCGACATGTTTGGCTAACAGCTAA
	95	CCGGCGTTGAGCGCCAGATGACTGAGAAAGAGC
	97	CCGTTCATATTCGTTTCCTCGTGGCGCGATCTA
	101	ATCATCTCCGCTGAATAGCGTAAATTATCAGGC
	102	ATTAAATCGTCAGAAAATAGCGGTAATCAAGTC
	103	ATTAAATCGTCAGAAAATAGCGGTAATCAAGTC
ST379	21	ACACAATCGTGTGTAAATTCGCGGCGGCTCCACTGG
	44	AAGACGACGTGATCCGCAAAGTCGAAGGCACG
	112	TGGCAAAACAAACATCGGGGTACGCGTGGTGC
	113	AAAATTCATATTGATAAACACCGCGTTTGTAT

Table 3 Genomic region conserved across STEC

Conserved	O26 Start	O 26 Stop	O26 Length	O26 Locus Tags	O157 EDL933 Locus Tags	Description
*	21257	22142	885	ECO26_0022	Z0025	Partial match to "T3SS effector-like protein EspX-homolog". (O-island 1)
*	**248034**	**258211**	**10177**	**ECO26_0220-ECO26_0230**	**Z0250-Z0258**	**Macrophage toxin and Type VI secretion system (O-island 7)**
	312255	312976	721	ECO26_0291-ECO26_0292	Absent	'Conserved predicted protein' between *yafP* and *pepD*. Absent in EDL933.
	640740	645802	5062	ECO26_0611-ECO26_0619	Z1888-Z1896	Structural phage related proteins (O-island 52)
	658907	659790	883	ECO26_0630	Z1930	Putative hydrolase, phage related (O-island 52)
	669640	670733	1093	ECO26_0636	Z0700	Putative receptor and insertion sequence:IS677 (O-island 30)
	675294	677991	2697	ECO26_0639-ECO26_0641	Z0705-Z0707	Unknown (Rhs related) and VgrG protein (O-island 30)
*	695271	695387	116	ECO26_0656	Z0722	HokC – small toxic membrane peptide
*	1052404	1054915	2511	ECO26_1001-ECO26_1003	Z1108-Z1110	Conserved putative proteins and aquaporin AqpZ, (O-island 42)
*	1624945	1625425	480	ECO26_1653	Z1844	Hypothetical protein (Phage related) Cryptic prophage CP-933C
	1740868	1741751	883	ECO26_1769-ECO26_1771	Z2036-Z2037	Integrase, excisionase and exonuclease, Phage related (O-island 57)
	1746722	1747229	507	ECO26_1779-ECO26_1780	Z2046-Z2047	Phage repressor and anti-repressor (O-island 57)
*	2003329	2003974	645	ECO26_2057	Z2262	VgrE gene – Rhs related (O-island 65)
	2052881	2053623	742	ECO26_2092-ECO26_2093	Z2213-Z2214	Predicted peptidase and porin protein
	2487136	2487833	697	ECO26_2563	Z3664	putative IS609 transposase TnpB (O-island 103)
	2488392	2488884	492	ECO26_2564	Z3665	putative IS609 transposase TnpA (O-island 103)
	3445476	3446301	825	ECO26_3520-ECO26_3521	Absent	Between *ypfJ* and *purC* : predicted post-segregational-killing toxin/anti-toxin
*	3746864	3747995	1131	ECO26_3806-ECO26_3807	Z4045-Z4046	putative 4-hydroxybenzoate decarboxylase (O-island 110)
*	**3761305**	**3770325**	**9020**	**ECO26_3823-ECO26_3832**	**Z4062-Z4071**	**CRISPR 2 – Cas genes**
	3885013	**3892930**	**7917**	**ECO26_3931-ECO26_3944**	**Z4180-Z4190**	***E. coli* Type Three Secretion System 2 (ETT2) Cryptic Type III secretion system (O-island 115)**
	4167893	4173254	5361	ECO26_4197-ECO26_4200	Absent	Predicted pillin, usher and fimbrial protein
	5024670	5026291	1621	ECO26_4995	Z5029	Putative adhesin (O-island 144)

Table 4 Distribution of CRISPR1 *cas* genes across *E. coli*

Strain Name	Pathotype	CRISPR type	Accession no.	Reference
***E. coli* B7A**	ETEC	STEC	AAJT02000001	(Rasko et al., 2008)
***E. coli* F11**	UPEC	Neither/Absent	AAJU02000001	(Rasko et al., 2008)
***E. coli* E22**	EPEC	STEC	AAJV02000001	(Rasko et al., 2008)
***E. coli* E110019**	aEPEC	STEC	AAJW02000001	(Rasko et al., 2008)
***E. coli* B171**	EPEC	STEC	AAJX02000001	(Rasko et al., 2008)
***E. coli* 53638**	EIEC	K12-like	AAKB02000001	(Turner et al., 2006)
***E. coli* 101-1**	aEAEC	Neither/Absent	AAMK02000001	(Rasko et al., 2008)
***E. coli* O157:H7 str. EC4206**	STEC	STEC	ABHK02000001	(Kotewicz et al., 2008)
***E. coli* O157:H7 str. EC4045**	STEC	STEC	ABHL02000001	(Kotewicz et al., 2008)
***E. coli* O157:H7 str. EC4042**	STEC	STEC	ABHM02000001	(Kotewicz et al., 2008)
***E. coli* O157:H7 str. EC4113**	STEC	STEC	ABHP01000001	(Lukjancenko et al., 2010)
***E. coli* O157:H7 str. EC4076**	STEC	STEC	ABHQ01000001	(Lukjancenko et al., 2010)
***E. coli* O157:H7 str. EC4401**	STEC	STEC	ABHR01000001	(Lukjancenko et al., 2010)

***E. coli* O157:H7 str. EC4486**	STEC	STEC	ABHS01000001	(Lukjancenko et al., 2010)
***E. coli* O157:H7 str. EC4501**	STEC	STEC	ABHT01000001	(Lukjancenko et al., 2010)
***E. coli* O157:H7 str. EC869**	STEC	STEC	ABHU01000001	(Lukjancenko et al., 2010)
***E. coli* O157:H7 str. EC508**	STEC	STEC	ABHW01000001	(Lukjancenko et al., 2010)
***E. coli* O157:H7 str. FRIK966**	STEC	STEC	ACXN01000001	(Dowd et al., 2010)
***E. coli* O157:H7 str. FRIK2000**	STEC	STEC	ACXO01000001	(Dowd et al., 2010)
***E. coli* O157:H7 str. EC4009**	STEC	STEC	ADMX01000001	(Xiong et al., 2012)
***E. coli* O157:H7 EDL933**	STEC	STEC	AE005174	(Perna et al., 2001)
***E. coli* CFT073**	UPEC	Neither/Absent	AE014075	(Welch et al., 2002)
***E. coli* W**	Lab-adapted	STEC	AEDF01000001	(Archer et al., 2011)
***E. coli* str. K-12 substr. MG1655star**	Lab-adapted	K12-like	AEFE01000001	(Fabich et al., 2011)
***E. coli* TW10598**	ETEC	K12-like	AELA01000001	(Sahl et al., 2011)
***E. coli* TW10722**	ETEC	STEC	AELB01000001	(Sahl et al., 2011)
***E. coli* TW10828**	ETEC	STEC	AELC01000001	(Sahl et al., 2011)
***E. coli* TW11681**	ETEC	K12-like	AELD01000001	(Sahl et al., 2011)
***E. coli* TW14425**	ETEC	STEC	AELE01000001	(Sahl et al., 2011)
***E. coli* O157:H7 str. G5101**	STEC	STEC	AETX01000001	(Rump et al., 2011)
***E. coli* O157:H- str. 493-89**	STEC	STEC	AETY01000001	(Rump et al., 2011)
***E. coli* O157:H- str. H 2687**	STEC	STEC	AETZ01000001	(Rump et al., 2011)
***E. coli* O55:H7 str. 3256-97**	EPEC	STEC	AEUA01000001	(Rump et al., 2011)
***E. coli* O55:H7 str. USDA 5905**	EPEC	STEC	AEUB01000001	(Rump et al., 2011)
***E. coli* O157:H7 str. LSU-61**	STEC	STEC	AEUC01000001	(Rump et al., 2011)
***E. coli* STEC_7v**	STEC	K12-like	AEXD01000001	(Steyert et al., 2012)
***E. coli* cloneA_i1**	UPEC	Neither/Absent	AEYT01000001	(Reeves et al., 2011)
***E. coli*PCN033**	ExPEC	Neither/Absent	AFAT01000001	(Tan et al., 2011)
***E. coli* STEC_B2F1**	?	STEC	AFDQ01000001	(Steyert et al., 2012)
***E. coli* STEC_C165-02**	STEC	STEC	AFDR01000001	(Steyert et al., 2012)
***E. coli* STEC_94C**	STEC	STEC	AFDU01000001	(Tan et al., 2011)
***E. coli* STEC_DG131-3**	STEC	STEC	AFDV01000001	(Steyert et al., 2012)
***E. coli* STEC_EH250**	STEC	STEC	AFDW01000001	(Steyert et al., 2012)
***E. coli* STEC_H.1.8**	STEC	STEC	AFDY01000001	(Steyert et al., 2012)
***E. coli* STEC_MHI813**	STEC	STEC	AFDZ01000001	(Steyert et al., 2012)
***E. coli* STEC_S1191**	STEC	STEC	AFEA01000001	(Steyert et al., 2012)
***E. coli* AA86 53_1**	Commensal	Neither/Absent	AFET01000001	(Yi et al., 2011)
***E. coli* O104:H4 str. LB226692**	STEC	STEC	AFOB02000001	(Mellmann et al., 2011)
***E. coli* O104:H4 str. C227-11**	STEC	STEC	AFRH01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. C236-11**	STEC	STEC	AFRI01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 09-7901**	STEC	STEC	AFRK01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 04-8351**	STEC	STEC	AFRL01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-3677**	STEC	STEC	AFRM01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. C227-11**	STEC	STEC	AFST01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4404**	STEC	STEC	AFUX01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4522**	STEC	STEC	AFUY01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4623**	STEC	STEC	AFUZ01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4632 C1**	STEC	STEC	AFVA01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4632 C2**	STEC	STEC	AFVB01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4632 C3**	STEC	STEC	AFVC01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4632 C4**	STEC	STEC	AFVD01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. 11-4632 C5**	STEC	STEC	AFVE01000001	(Grad et al., 2012)

***E. coli* O104:H4 str. Ec11-5538**	STEC	STEC	AFVF01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. Ec11-5537**	STEC	STEC	AFVG01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. Ec11-5536**	STEC	STEC	AFVH01000001	(Grad et al., 2012)
***E. coli* O104:H4 str. TY-2482**	STEC	STEC	AFVR01000001	(Rohde et al., 2011)
***E. coli* O104:H4 str. TY-2482**	STEC	STEC	AFVS01000001	(Rohde et al., 2011)
***E. coli* XH140A**	Lab-adapted	K12-like	AFVX01000001	(Yang et al., 2011b)
***E. coli* O104:H4 str. GOS1**	STEC	STEC	AFWO001000001	(Brzuszkiewicz et al., 2011)
***E. coli* O104:H4 str. GOS2**	STEC	STEC	AFWP01000001	(Brzuszkiewicz et al., 2011)
***E. coli* XH001**	Lab-adapted	K12-like	AFYG01000001	(Yang et al., 2011a)
***E. coli* O103:H25 str. NIPH-11060424**	STEC	STEC	AGSG01000001	(L'Abee-Lund et al., 2012)
***E. coli* UMNF18**	STEC	STEC	AGTD01000001	(Shepard et al., 2012)
***E. coli* O113:H21 str. CL-3**	STEC	STEC	AGTH01000001	(Shen et al., 2004)
***E. coli* SE11**	Commensal	STEC	AP009240	(Oshima et al., 2008)
***E. coli* O26:H11 str. 11368**	STEC	STEC	AP010953	(Ogura et al., 2009)
***E. coli* O103:H2 str. 12009**	STEC	STEC	AP010958	(Ogura et al., 2009)
***E. coli* O111:H- str. 11128**	STEC	STEC	AP010960	(Ogura et al., 2009)
***E. coli* O157:H7 str. Sakai**	STEC	STEC	BA000007	(Hayashi et al., 2001, Perna et al., 2001)
***E. coli* HM605**	AIEC	Neither/Absent	CADZ01000001	(Clarke et al., 2011)
***E. coli* O25b:H4-ST131 str. EC958**	UPEC	Neither/Absent	CAFL01000001	(Totsika et al., 2011)
S. dysenteriae Sd197	Shigella	Neither/Absent	CP000034	(Yang et al., 2005)
S. boydii Sb227	Shigella	STEC	CP000036	(Yang et al., 2005)
S. sonnei Ss046	Shigella	STEC	CP000038	(Yang et al., 2005)
***E. coli* UTI89**	UPEC	Neither/Absent	CP000243	(Chen et al., 2006)
***E. coli* 536**	?	Neither/Absent	CP000247	(Hochhut et al., 2006)
S. flexneri 5b str. 8401	Shigella	Neither/Absent	CP000266	(Nie et al., 2006)
***E. coli* APEC O1**	?	Neither/Absent	CP000468	(Blattner et al., 1997)
***E. coli* E24377A**	ETEC	STEC	CP000800	(Rasko et al., 2008)
***E. coli* HS**	Commensal	K12-like	CP000802	(Rasko et al., 2008)
***E. coli* B str. REL606**	Commensal	Neither/Absent	CP000819	(Jeong et al., 2009)
***E. coli* ATCC 8739**	Lab-adapted	K12-like	CP000946	
***E. coli* str. K12 substr. DH10B**	Lab-adapted	K12-like	CP000948	(Durfee et al., 2008)
***E. coli* SMS-3-5**	?	STEC	CP000970	(Fricke et al., 2008)
S. boydii CDC 3083-94	Shigella	Neither/Absent	CP001063	No citation
***E. coli* O157:H7 str. EC4115**	STEC	STEC	CP001164	(Eppinger et al., 2011)
***E. coli* O157:H7 str. TW14359**	STEC	STEC	CP001368	(Kulasekara et al., 2009)
***E. coli* BW2952**	Commensal	K12-like	CP001396	(Ferenci et al., 2009)
***E. coli*BL21(DE3)**	Commensal	Neither/Absent	CP001509	(Jeong et al., 2009)
***E. coli* O55:H7 str. CB9615**	EPEC	STEC	CP001846	(Zhou et al., 2010)
***E. coli* 55989**	EAEC	STEC	CU928145	(Touchon et al., 2009)
E. fergusonii ATCC 35469	?	K12-like	CU928158	(Touchon et al., 2009)
***E. coli* IAI1**	Commensal	STEC	CU928160	(Touchon et al., 2009)
***E. coli* S88**	?	Neither/Absent	CU928161	(Touchon et al., 2009)
***E. coli* ED1a**	?	Neither/Absent	CU928162	(Touchon et al., 2009)
***E. coli* UMN026**	UPEC	STEC	CU928163	(Touchon et al., 2009)
***E. coli*0127:H6 E2348/69**	EPEC	Neither/Absent	FM180568	(Iguchi et al., 2009)
***E. coli* str. K-12 substr. MG1655**	Lab-adapted	K12-like	U00096	(Blattner et al.,

				1997)

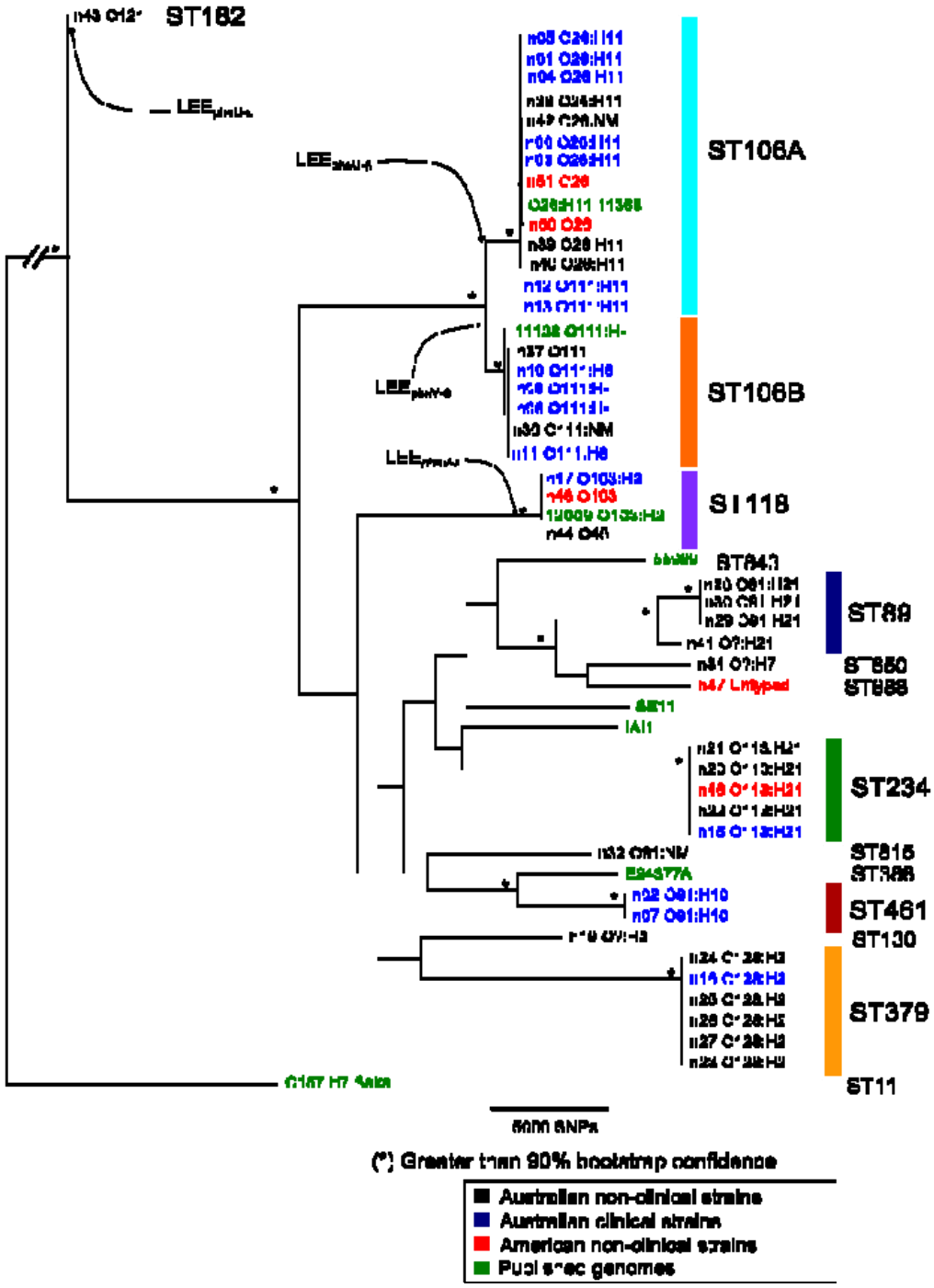

Figure 1 Phylogenetic relationship of non-O157 STEC with *E. coli* B1 Phylogroup

Maximum Likelihood (ML) phylogram with asterisks indicating bootstrap support greater than 90% from 400 replicates. The tree was rooted using *E. coli*O157:H7 Sakai (Accession no. BA000007). The phylogram includes forty-four Shiga toxin positive *E. coli* from this study and seven other *E. coli* strains from the B1 phylogroup including 55989, SE11, IAI1, E24377A, and previously sequenced non-O157 strains; 11368, 11128 and 12009. Accession numbers of these strains are listed in Table 4. Genomes have been annotated

and highlighted according to lineage and named according to the EcMLST seven allele schema. Isolate sources are indicated in the key. The phylogram was built from 48,912 nucleotide SNPs from 2,153 *E. coli* genes, which are the number of genes conserved across the B1 Phylogroup, using PhyML (v20120412)(Guindon et al., 2010) with the HKY85 substitution model. SplitsTree4(Huson, 1998) was used to generate the final consensus tree. The final figure was prepared in FigTree (v1.4) (http://tree.bio.ed.ac.uk/software/figtree/). Labels indicate independent acquisition of the Locus for Enterocyte Effacement (LEE), site of insertion (*pheV*, *pheU*) and Intimin type (epsilon, theta and beta).

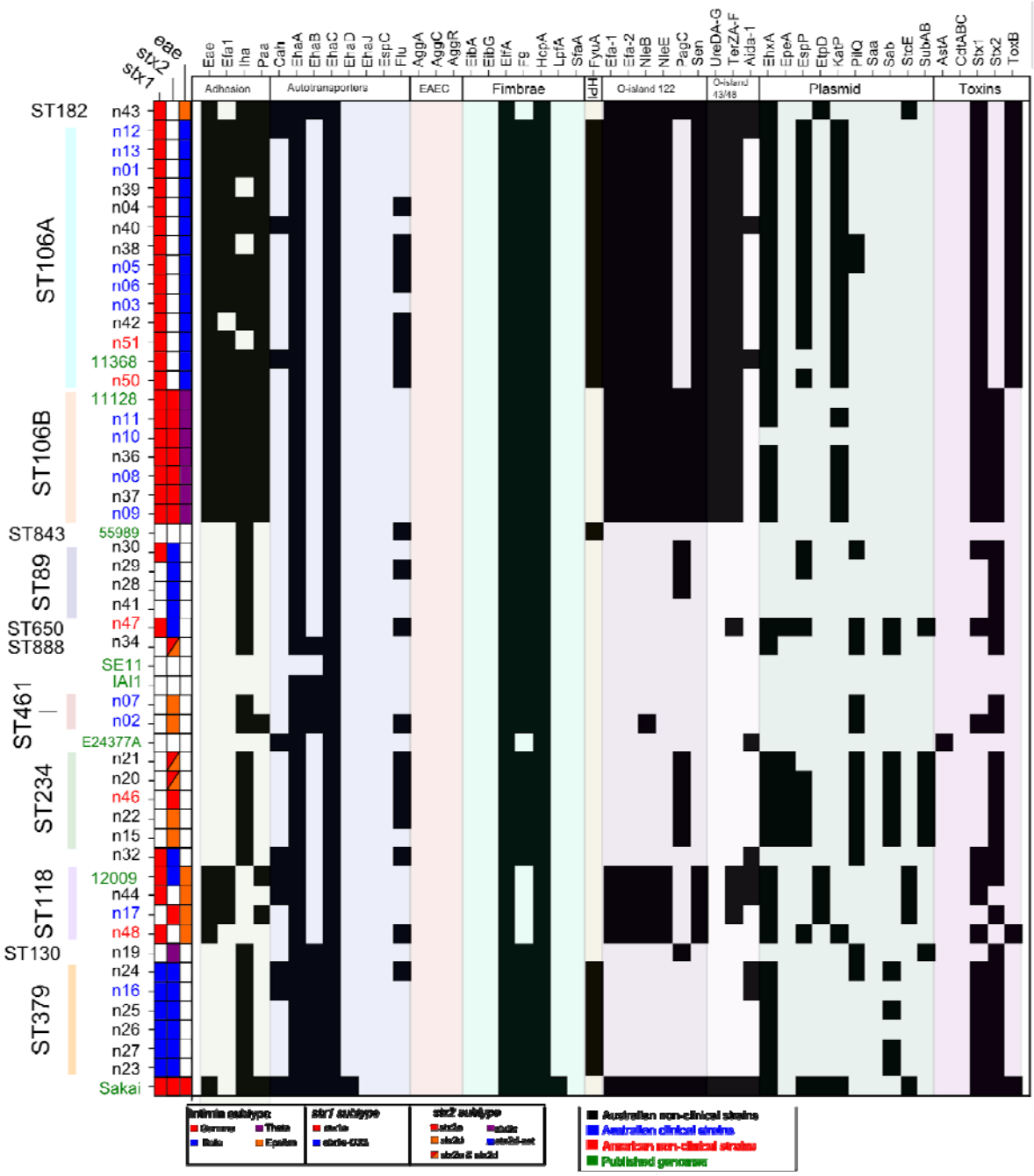

Figure 2.Virulence profile of non-O157 STEC and *E. coli* within B1 Phylogroup.

Presence/absence matrix of a panel of STEC virulence factors. Virulence factors are shown along the x-axis with strains along the y-axis, listed in the order presented in the whole genome phylogeny (Figure 1). Genes are considered present (black) with greater than 80% average translated nucleotide identity, calculated using BLASTx (BLAST+ v2.2.26 (Camacho et al., 2009)),across the total reference gene length. Figure was prepared using SeqFindr (http://github.com/mscook/seqfindr).

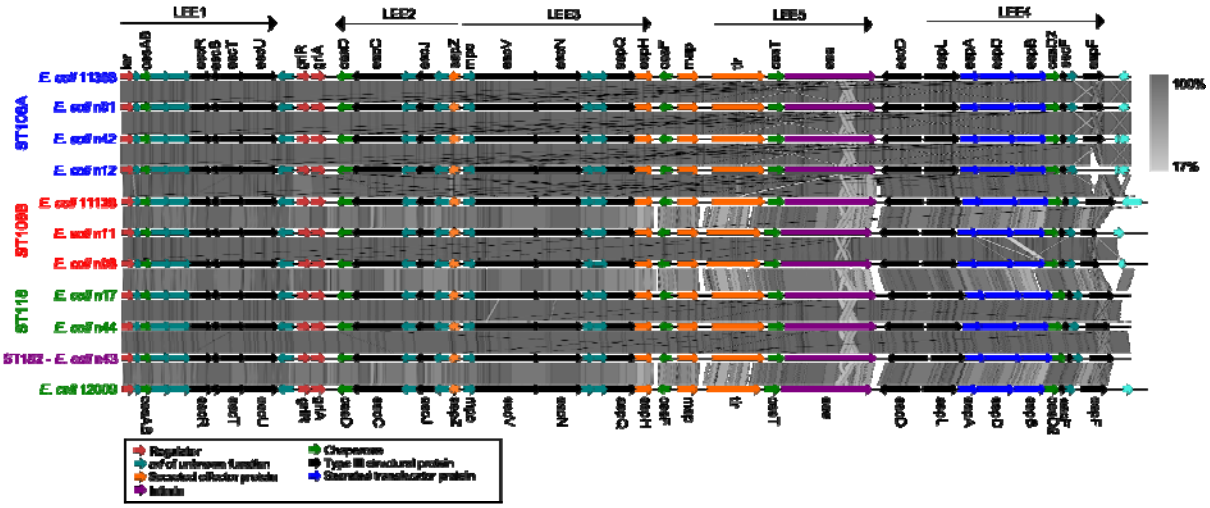

Figure 3 Sequence comparison of LEE region from non-O157 STEC

Translated nucleotide comparison (BLAST 2.2.27+ tBLASTx) (Camacho et al., 2009) of part of the Locus of Enterocyte Effacement (between *ler* and *espF*) in representative non-O157 STEC from different serogroups and sequence types. Strain labels have been color coded according to lineages defined in Figure 1. tBLASTx alignment identity score is indicated by scale gradient. Figure was prepared using EasyFig(Sullivan et al., 2011). CDS were color coded according to function outlined in (Garmendia et al., 2005).

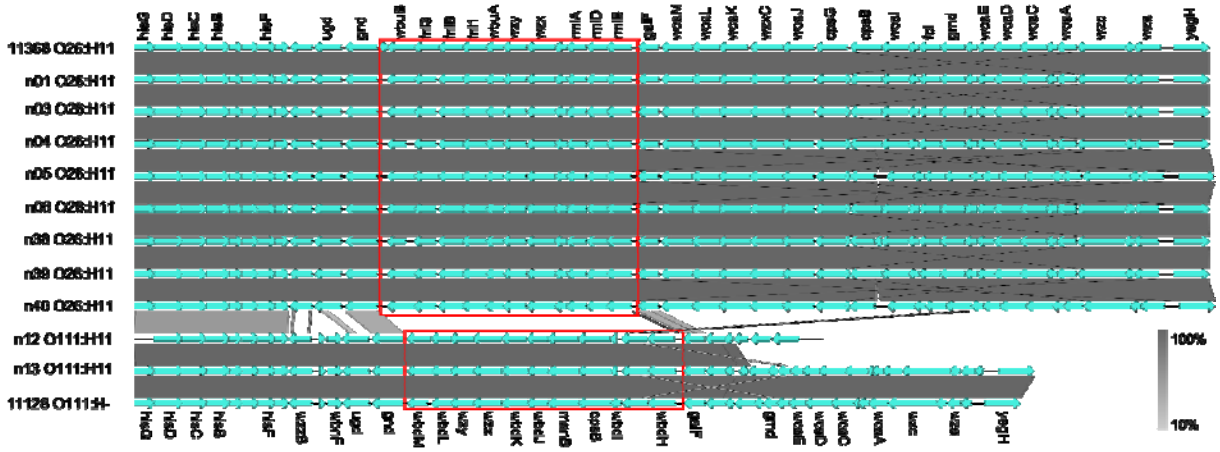

Figure 4 Comparison of O-antigen synthesis region in ST106A and ST106B strains

Nucleotide to nucleotide BLAST (BLASTn) comparison of O-antigen synthesis region (boxed) between *hisG* and *yegH* from non-O157 STEC strains from ST106A and ST106B. O-antigen region from ST106A strains, n12 and n13, matched O-antigen region from published O111:H- ST106B strain 11128. BLASTn alignment identity score is indicated by scale gradient. Figure was prepared using EasyFig(Sullivan et al., 2011).

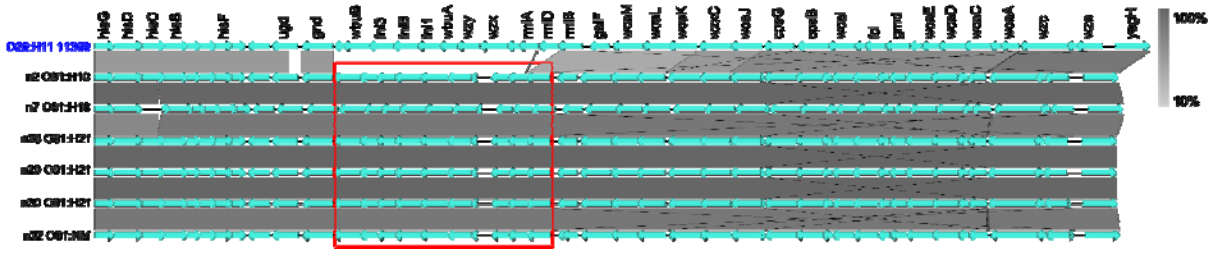

Figure 5 Comparison of O-antigen synthesis region in O91 strains

Nucleotide to nucleotide BLAST (BLASTn) comparison of O-antigen synthesis region (boxed) between *hisG* and *yegH* from O91 STEC strains. O-antigen regions for O91 strains were identical, while a comparison to O26:H11 strain 11368 (blue) showed no detectable nucleotide similarity. BLASTn alignment identity score is indicated by scale gradient. Figure was prepared using EasyFig (Sullivan et al., 2011).

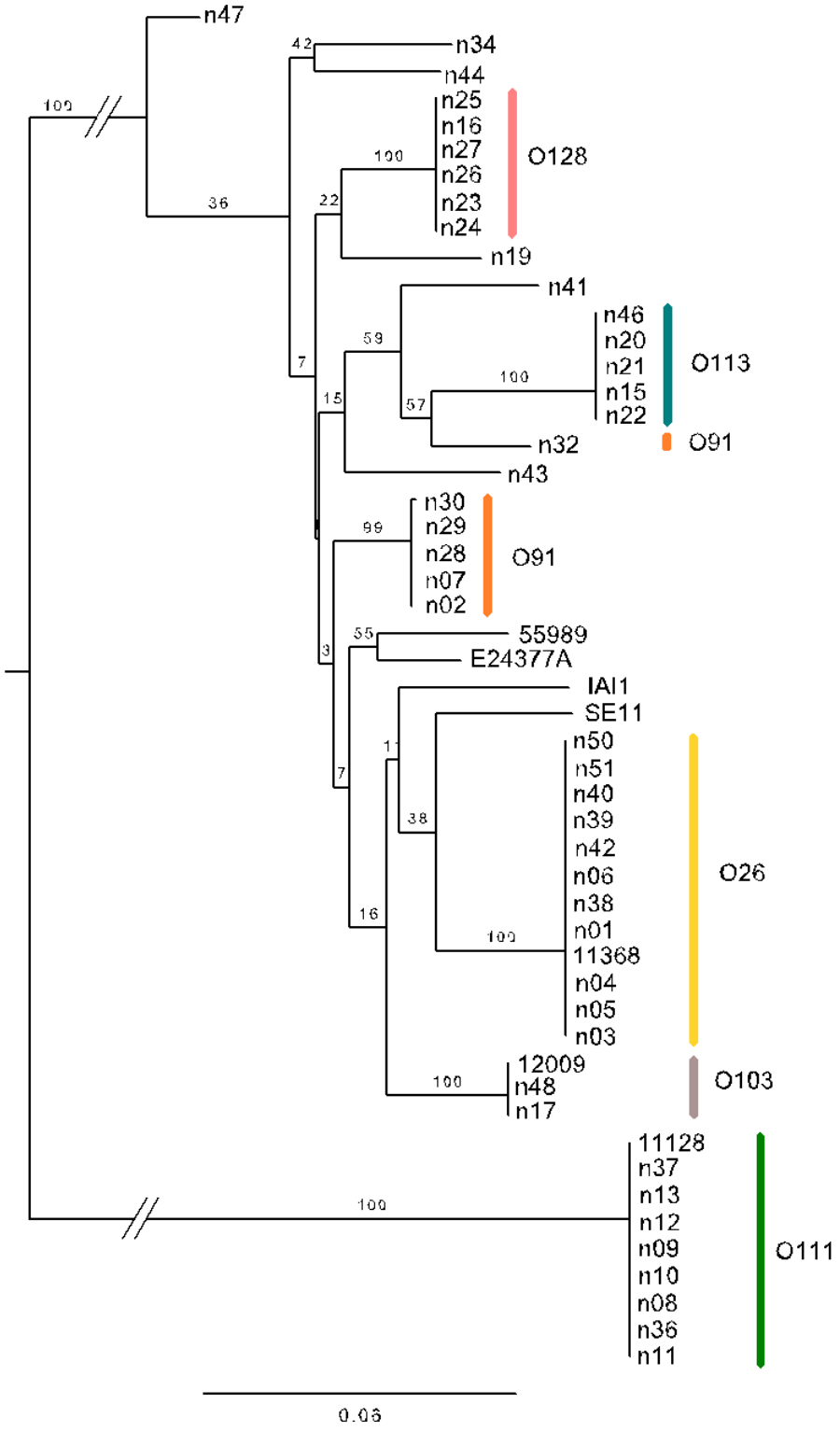

Figure 6 Phylogenetic relationship of *gnd*

Maximum Likelihood (ML) phylogram, based on *gnd* gene sequence, from 1000 replicates. The phylogram included forty-four STEC from this study and seven other *E. coli* strains from the B1

phylogroup including EAEC 55989, SE11, IAI1, ETEC E24377A, and previously sequenced non-O157 strains; 11368, 11128 and 12009. Accession numbers of these strains are listed in Table 4. Genomes highlighted according to O-antigen type. The final figure was prepared with FigTree(v1.4) (http://tree.bio.ed.ac.uk/software/figtree/). The phylogram was built using RaXML(Stamatakis, 2014) with the GTR substitution model.

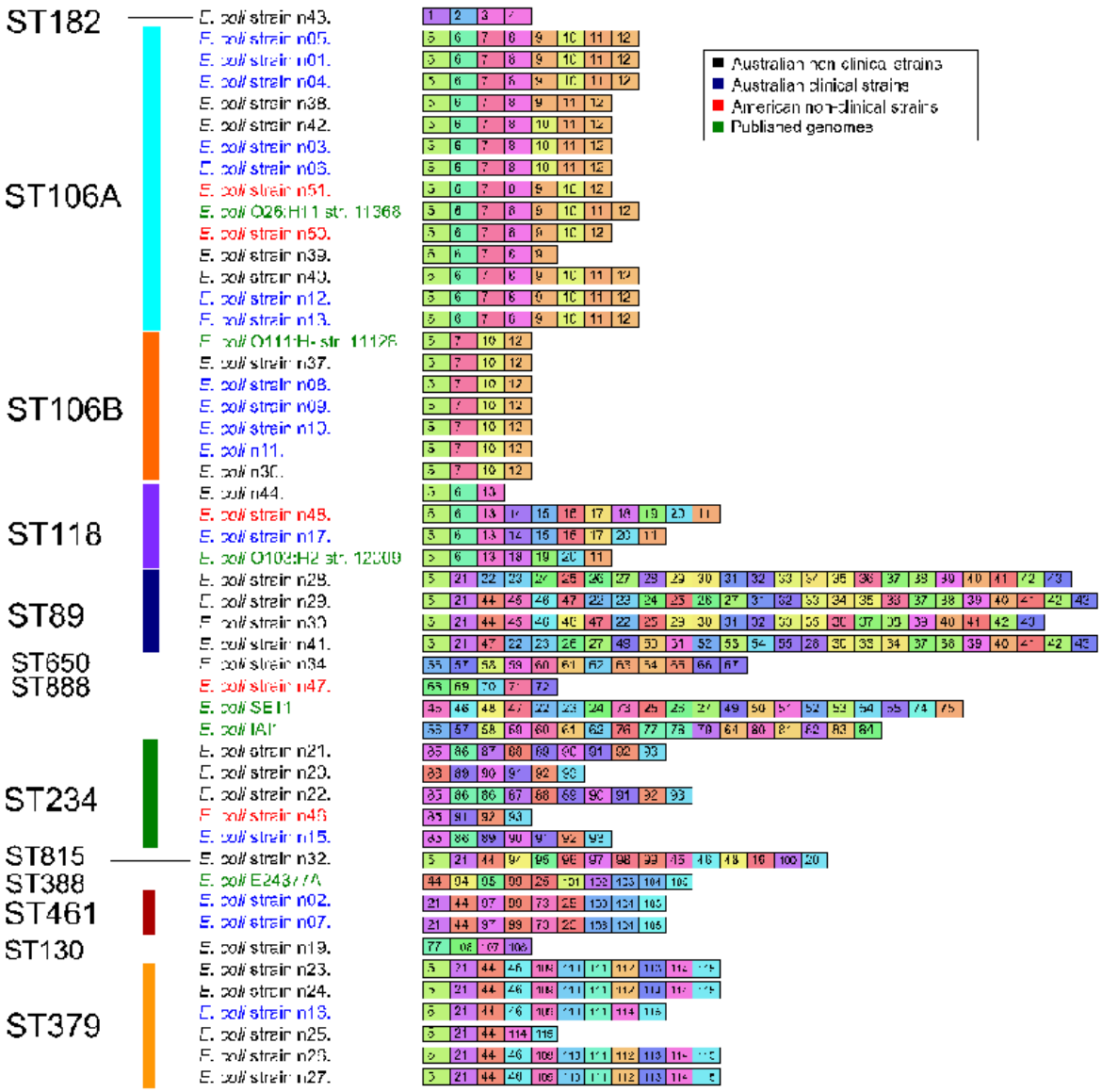

Figure 7 Conservation of CRISPR1 spacers in non-O157 STEC and B1 phylogroup *E. coli*

Graphic representation of spacer content from CRISPR1 for B1 phylogroup *E. coli* including strains

from this study and other B1 phylogroup *E. coli* that share spacer sequences including SE11, E24377A and IAI1. A uniquely colored box and symbol combination designates each spacer sequence. Sequences are listed (left to right) from furthest to nearest the CRISPR leader sequence. Strains and lineages are listed in order and colored according to the scheme used for the phylogram in Figure 1.

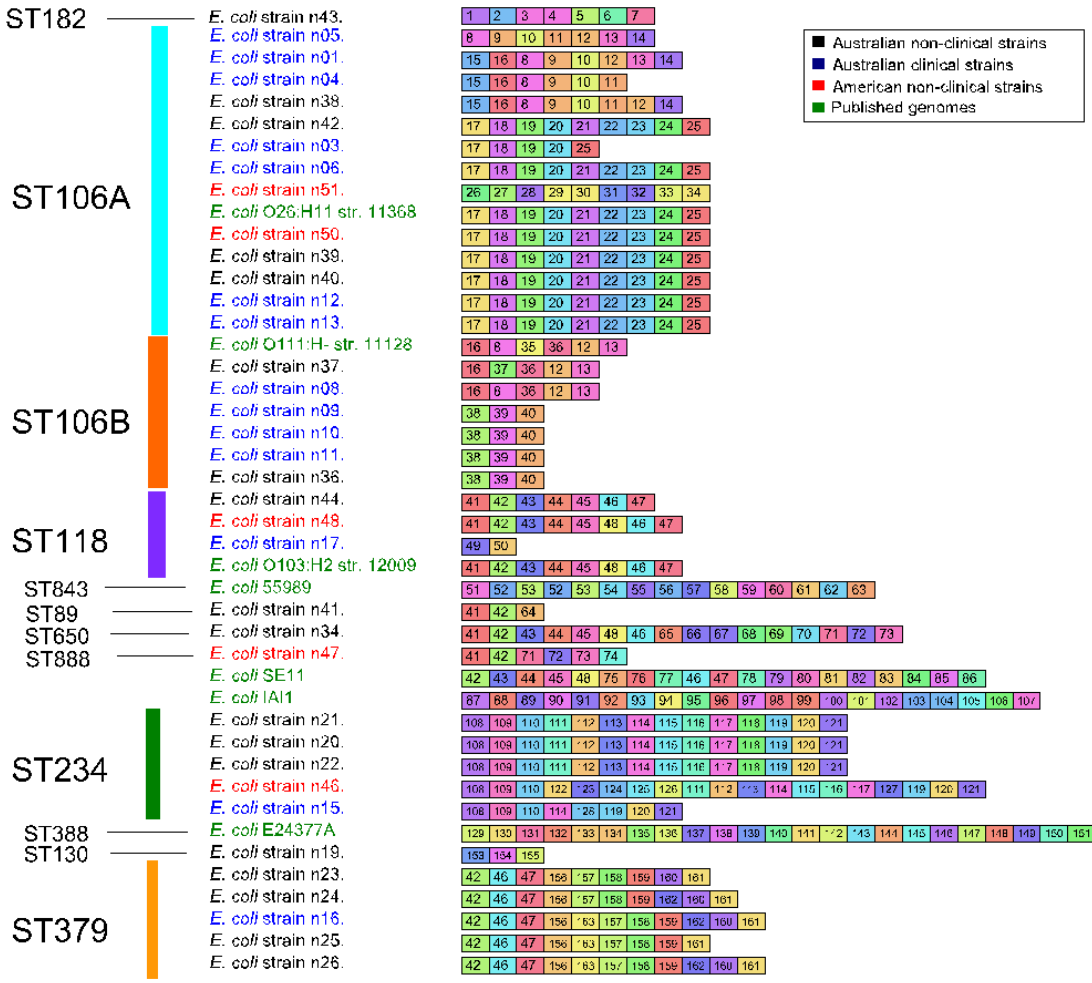

Figure 8 Conservation of CRISPR2 spacers in non-O157 STEC and B1 phylogroup *E. coli*

Graphic representation of spacer content from CRISPR2 for non-O157 STEC strains and B1 phylogroup *E. coli* that share space sequences. A uniquely colored box and symbol combination designates each spacer sequence. The colors and the numbers were assigned arbitrarily and are different from those in Figure 7. Sequences are listed (left to right) from farthest to nearest the

CRISPR leader sequence. Strains and lineages are listed in order and colored according to the scheme used for the phylogram in Figure 1.