Genome Signatures of *Escherichia coli* O157:H7 Isolates from the Bovine Host Reservoir

Mark Eppinger,1 Mark K. Mammel,2 Joseph E. LeClerc,2 Jacques Ravel,1* and Thomas A. Cebula3*

Institute for Genome Sciences (IGS), University of Maryland, School of Medicine, Baltimore, Maryland 212011; Division of Molecular Biology, Office of Applied Research and Safety Assessment, Center for Food Safety and Applied Nutrition, U.S. Food and Drug Administration, Laurel, Maryland 207082; and Johns Hopkins University, Department of Biology, Baltimore, Maryland 212183

Received 29 October 2010/Accepted 6 March 2011

Cattle comprise a main reservoir of Shiga toxin-producing *Escherichia coli* O157:H7 (STEC). The significant differences in host prevalence, transmissibility, and virulence phenotypes among strains from bovine and human sources are of major interest to the public health community and livestock industry. Genomic analysis revealed divergence into three lineages: lineage I and lineage I/II strains are commonly associated with human disease, while lineage II strains are overrepresented in the asymptomatic bovine host reservoir. Growing evidence suggests that genotypic differences between these lineages, such as polymorphisms in Shiga toxin subtypes and synergistically acting virulence factors, are correlated with phenotypic differences in virulence, host ecology, and epidemiology. To assess the genomic plasticity on a genome-wide scale, we have sequenced the whole genome of strain EC869, a bovine-associated *E. coli* O157:H7 isolate. Comparative phylogenomic analysis of this key isolate enabled us to place accurately bovine lineage II strains within the genetically homogenous *E. coli* O157:H7 clade. Identification of polymorphic loci that are anchored both in the chromosomal backbone and horizontally acquired regions allowed us to associate bovine genotypes with altered virulence phenotypes and host prevalence. This study catalogued numerous novel lineage II-specific genome signatures, some of which appear to be associated intimately with the altered pathogenic potential and niche adaptation within the bovine rumen. The presented extended list of polymorphic markers is valuable in the development of a robust typing system critical for forensic, diagnostic, and epidemiological studies of this emerging human pathogen.

Shiga toxin-producing, non-sorbitol-fermenting, and β-glucuronidase-negative *Escherichia coli* (STEC) O157:H7 has evolved from an O55:H7-like progenitor (24, 50, 65) into an emerging human pathogen, with cattle as the main asymptomatic reservoir (12, 54). *E. coli* O157:H7 is transmitted from cattle to human by means of contaminated food products, such as undercooked meat, unpasteurized milk, fruit and vegetables, or tainted water. As seasonal changes can influence the prevalence and load of *E. coli* O157:H7 in cattle and super shedders exist within the bovine population (30, 33, 39, 49, 67), physical contact of humans with cattle and their environment introduces added risk of *E. coli* infection (29). Human infection manifests in various ways, ranging from mild to more severe bloody diarrhea. In some cases, infection can lead to renal dysfunction, i.e., hemolytic uremic syndrome (HUS), and central nervous system (CNS) failure (10, 17, 73). Epidemiological data have demonstrated a high prevalence of *E. coli* O157:H7 in cattle and their environment but a comparatively low incidence of human infection. This supports the notion that a subset of STEC O157:H7 strains harbored in cattle causes the majority of human disease (18). Genetic heterogeneity among STEC O157:H7 strains has been established using a broad panel of typing methodologies, such as multilocus sequence tagging (2), octamer- and PCR-based genome scanning (41, 66), phage typing (4, 72), multiple-locus variable-number tandem repeat analysis (40, 64), microarrays (37), nucleotide polymorphism assays (56, 87) (M. Eppinger, M. K. Mammel, J. E. LeClerc, T. A. Cebula, and J. Ravel, unpublished data), pulsed-field gel electrophoresis (PFGE) (28), subtractive hybridization (79, 80), and optical mapping (43, 44). High-resolution genotyping studies on *E. coli* O157:H7 strains utilizing octamer-based genome scanning (OBGS) first demonstrated that the *E. coli* O157:H7 clonal complex has diverged into two lineages, designated lineages I and II, that were disproportionately represented among human and cattle isolates, respectively (41). Further analyses led to a refined classification system, termed the lineage-specific polymorphism assay (LSPA), that partitioned *E. coli* O157:H7 strains ultimately into three groupings—lineages I, I/II, and II—based on sequence length polymorphisms at six distinct loci within the O157:H7 genome (84). Whereas *E. coli* O157:H7 strains of lineages I and I/II were isolated from human clinical cases, lineage II strains were more commonly derived from the bovine reservoir (84). Lineage II isolates indeed are thought to be less pathogenic and possibly impaired in their transmissibility to humans (50). For example, branch-specific genotypes differ

---

* Corresponding author. Mailing address for Jacques Ravel: Institute for Genome Sciences (IGS), University of Maryland, School of Medicine, Baltimore, MD 21201. Phone: (410) 706-5674. Fax: (410) 706-1482. E-mail: jravel@som.umaryland.edu. Mailing address for Thomas A. Cebula: Johns Hopkins University, Department of Biology, Baltimore, MD 21218. Phone: (410) 516-7285. Fax: (410) 516-5213. E-mail: tcebu1@jhu.edu.
† Supplemental material for this article may be found at http://aem.asm.org.
‡ Published ahead of print on 18 March 2011.
substantially in the frequencies with which they are associated with isolates from clinical or bovine settings (8, 88), and testing of STEC infection in the gnotobiotic pig model suggests that the virulence of cattle-derived strains may differ from that of strains isolated from humans (7). The mohbiome is a major factor in the genome evolution and differentiation of these two lineages (11, 48, 88). Shiga toxin (stx)-converting phases of toxin subtypes stx2 and stx3 are more frequently carried in clinical STEC isolates, while stx1 is more frequent in O157 and non-O157 STEC isolates of bovine origin (13, 32) (Eppinger, Mammel, LeClerc, Cebula, and Ravel, unpublished). Lineages I, I/I, and II are further distinguished by polymorphisms in toxin production and expression of other synergistic key virulence factors (3, 10). STEC strains produce characteristic at-toxin production and expression of other synergistic key viru-

Mammel, LeClerc, Cebula, and Ravel, unpublished). Lineages I, I/I, and II are further distinguished by polymorphisms in toxin production and expression of other synergistic key virulence factors (3, 10). STEC strains produce characteristic at-toxin production and expression of other synergistic key viru-

86% of the coding capacity of the reference strain, EC4115. Phylogenomic analysis of the bovine isolates. (i) Comparative analysis of protein sequences. The close genetic relatedness within the E. coli O157:H7 clade is reflected in a high degree of overall protein sequence conservation, as evidenced by a BLAST score ratio (BSR) analysis that is evident for isolates derived from both humans and cattle (see Table S3 in the supplemental material) (71). Bioinformatic comparisons classified 748 divergent protein sequences (BSR \(> 0.8\)) and 4,567 conserved protein sequences (BSR \(> 0.8\)) which is about

\(0.8\), which is about
however, this data set is not sufficient to deduce a monomorphic population structure. Due to the genome dynamics driven by the highly homologous prophages, such comparative analysis must consider the overall genomic architecture and phage prevalence, which disrupts the *E. coli* O157:H7 genome synteny (Fig. 1).

(ii) SNP discovery and typing. Detection of genetic diversity between STEC strains is an important component of outbreak

FIG. 1. Genome organization in completed *E. coli* O157:H7 genomes. Genome synteny in strains EC4115, TW14359 (A), EDL933 (B), and Sakai (C) is disrupted by isolate-specific lateral acquired genomic regions introduced by prophages and mobile genetic elements, such as transposases and insertion sequence elements. Each protein of the reference genome EC4115 is plotted on the *x* axis and was queried for its presence in the *y* axis query genome using BLASTP. For a match, the N-terminal coordinates of both proteins were plotted as *x* and *y* coordinates. The color represents the level of similarity of the match expressed by the BLAST score ratio (71).
investigations. Nucleotide polymorphisms are key in determining relatedness of genetically homogenous pathogens and enable one to discriminate between and among environmental and outbreak strains (23) (Eppinger, Mammel, LeClerc, Cebula, and Ravel, unpublished). To achieve the necessary high phylogenetic resolution and discover branch-specific genome signatures, we applied a multilayer approach of SNP-derived genotyping coupled with an analysis of phage content and polymorphisms. SNPs have been identified previously in different subsets of STEC genomes (14, 50, 56, 87); however, the identified SNPs are biased toward strains originating from human outbreaks and thus may be limited in their ability to distinguish genetic STEC subtypes present in cattle (56). We applied a phylogenomic SNP discovery and SNP validation pipeline that was specifically developed for the high-resolution typing of E. coli O157:H7 strains. Genotyping identified 298 SNPs that separate strain EC869 from the remainder of the tested isolates. All 298 SNPs were confirmed in two other completed lineage II genomes, FRIK2000 and FRIK966 (20) (see Table S2 in the supplemental material). Thus, detected SNPs are considered to represent lineage II branch-specific SNPs. The SNP-derived high-resolution phylogenetic analysis places the bovine isolates on a distinct branch (Fig. 2) and identified the lineage I/II strains EC508 as a close phylogenetic relative.

Association of polymorphisms with altered physiology and virulence. Such detected fine polymorphisms are key in understanding the genetic makeup and altered physiology of the bovine lineage II isolates. The panel of SNPs is comprised of 81 synonymous SNPs (sSNPs), 153 nonsynonymous SNPs (nSNPs), and 64 intergenic SNPs (see Table S2 in the supplemental material). A 16-bp deletion within the fim (nsSNPs), and 64 intergenic SNPs (see Table S2 in the supplemental material). Such detected fine polymorphisms are key in understanding the genetic makeup and altered physiology of the bovine lineage II isolates. The panel of SNPs is comprised of 81 synonymous SNPs (sSNPs), 153 nonsynonymous SNPs (nSNPs), and 64 intergenic SNPs (see Table S2 in the supplemental material). A 16-bp deletion within the fim-switching region of fimA, shown to underlie the lack of expression of type I fimbriae in E. coli O157:H7 strains, was found in both human- and bovine-derived strains analyzed in the present studies (51, 75, 78). These data are consistent with the hypothesis that the fim switch mutation occurred early on in E. coli O157:H7 evolution, before the divergence into three sublineages (78). A C→A substitution in fimH results in the replacement of an asparagine residue by lysine in the mannos-binding pocket of the FimH protein (see Table S2); however, because of the fim mutation, type I fimbriae are not expressed. Although the N135K polymorphism has been previously noted (78), we show here that the C→A transition in the fimH gene is distributed nonrandomly within the three lineages analyzed. That is, while the C allele is found both in lineage I/II and lineage II strains, the A allele is found in each of the four lineage I strains, but only one of 18 lineage I/II strains, examined in this study (see Table S2).

A nonrandom distribution of a T255A base substitution within the intimin receptor has been reported, with isolates containing the A allele, to be 34 times more likely to be of bovine rather than human origin, although the physiological relevance of this base substitution that leads to an Asp→Glu substitution in LII strains is not known (14, 18). Indeed, the three analyzed bovine strains harbor the A allele, while the human lineage I/II isolate, strain EC508, contains the T allele. LII isolates are also distinguished by an 18-bp sequence (CA AAGGCGTTGGGAGGT), which is located within the N-terminal Tir receptor domain (11) (see Fig. S1 in the supplemental material). This sequence is perfectly duplicated to form an 18-bp tandem repeat in the three analyzed lineage II isolates. Though the underlying cause for this host restriction or its impact on pathogenicity is unknown, these novel alleles provide valuable genetic markers for the grouping of STEC isolates based on their likely human or bovine origin.

SNP discovery detected eight lineage II genes with branch-specific premature stop codons (see Table S2 in the supplemental material), and some of these are implicated in STEC pathogenicity and might play a role in the host adaption to the bovine environment. For example, one such stop occurs in luxR, a transcription regulator that governs expression of proteins of ETSS2, a type III secretion system (see Fig. S2 in the supplemental material). The genome of E. coli O157:H7 harbors two type III secretion systems. In strain EC4115, the loci are ETSS1, located within the LEE island (4691936 bp→4725720 bp), and ETSS2, localized between ECH74115_4114 and ECH74115_4154. In contrast to the ETSS1 system, the ETSS2 system is not directly involved in the injection of virulence factors. Recently, however, the expression of three genes within the ETSS2 locus has been shown to negatively impact the expression of ETSS1 genes
(86). In the absence of luxR, it is expected that overexpression of ETTS22 proteins would subvert injection of virulence factors by the ETTS1 system (59, 86).

Another premature stop occurs with the ethanolamine utilization (eut) operon. As recent studies have shown that E. coli O157:H7 incubated in bovine rumen under aerobic conditions can utilize ethanolamine present in the rumen as an efficient nitrogen source for growth (9), our finding of a lineage II-specific premature stop in eutA, encoding reactivating factor, is especially noteworthy (see Table S2 in the supplemental material). Although the eutA gene product is essential for ethanolamine utilization under aerobic conditions, as it rescues ethanolamine ammonia lyase from suicide inactivation by toxic by-products of exogenously added B12, only eutB and eutC gene products are necessary under anaerobic conditions for the utilization of ethanolamine as a nitrogen source (76). Thus, the mutation that we localized in the bovine lineage does not compromise its ability to grow and compete within the animal reservoir. Rather, as ethanolamine catabolism has been implicated in bacterial pathogenicity, i.e., linked to intestinal host colonization, impaired gut functioning, and immune evasion in a diverse range of enteric pathogens (27), we theorize that the nonsense mutation attenuates virulence in eutA strains. That is, host-pathogen interactions result in the upregulation of eut genes, which is accompanied, at least in Salmonella enterica serovar Typhimurium (47), by the activation of global virulence regulators. If the E. coli O157:H7 eutA mutation behaves as it does in S. Typhimurium, strains harboring this mutation, like the bovine isolates described here, may be impaired in their abilities for human transmission and pathogenicity.

STEC isolates produce cellulose as a component of their extracellular matrix that is involved in virulence, colonization, and biofilm formation (90). Bacterial cellulose biosynthesis (bcs) is conferred by the constitutively transcribed bcsABZC operon (ECH7EC869_1529, ECH7EC869_1534). This locus was initially described in Acetobacter xylinum and shares high protein conservation and syntenic arrangement with corresponding loci of E. coli and S. enterica (83, 90). Within E. coli O157:H7, we found four distinct genotypes, a 10-bp deletion (GTTCACAACAA), and three independent nsSNPs (see Table S2 in the supplemental material) within this operon (see Fig. S3 in the supplemental material). One of these SNPs, at position 1,674, leads to a truncated cellulose synthase C gene (ECH7EC869_1530, ECH7EC869_1531) (see Table S2). Previous research demonstrated that the four-gene synthase operon is essential for maximal cellulose synthesis in A. xylinum (83), and we therefore suspect reduced cellulose production in the genetic background of lineage II isolates. This novel polymorphism may underlie reported differences in cellulose production in STEC isolates (85). As cellulose is abundant in the bovine rumen, we speculate that the discovered lineage II cellulose phenotype is not disadvantageous within the bovine environment, but within the human milieu, bcs mutants may manifest with impaired transmissibility to humans and reduced pathogenicity (90).

The lsrACDBFgium operon is an integral part of the AI-2 quorum-sensing (QS) system monitoring cell density in biofilms (5, 52). We detected a lineage II-specific 1,339-bp deletion that leads to a C-terminal-truncated version of the S-adenosyl-L-methionine-dependent methyltransferase (tam) (ECH7EC869 4857, 88 amino acids [aa]) compared to those of the other STEC isolates, such as EC4115 (ECH74115_2132, 252 aa) (see Fig. S4 in the supplemental material). The lineage II genomes lack the neighboring conserved hypothetical protein (ECH74115_2133) with no assigned function. Tam catalyzes the methylesterification of trans-aconitate from the citric acid cycle intermediate cis-aconitate. The physiological role of aconitate conversion in E. coli cells is not clear. However, it is known that further conversion of trans-aconitate to tricarballylic acid by the rumen microbiome can induce hypomagnesemia and grass tetany in cattle (77). Thus, a tam mutant of E. coli O157:H7 might be a welcomed resident within the cattle host.

Prophage content and dynamics. E. coli O157:H7 has become increasingly more resistant to streptomycin, sulfisoxazole, and tetracycline, likely resulting from the antibiotic treatment of livestock (10) with noted differences in resistance phenotypes between the two major lineages (89). This finding is supported by the discovery of a novel strain-specific phage that underlies the multidrug resistance (MDR) phenotype in strain EC869. The human strain EDL933 and bovine strain EC869 carry phylogenetically unrelated P4-type prophages that, in both cases, are located within the cfpA locus (68) (Fig. 3A). The EC869 prophage is an entirely novel P4 prophage. This 56,859-bp phage introduces many loci that are potentially associated with the pathogenic potential and niche adaptation to the bovine rumen. The novel phage carries three predicted adhesins (ECH7EC869_5853, _5857, _5859). Phylogenetic analyses showed that one of the three predicted phage adhesins (ECH7EC869_5857) is related to adhesins found in the bovine E. coli strain RW1374 (see Fig. S5A in the supplemental material), which suggests a role in bovine niche adaptation (38). Delineated from its domain of unknown function (DUF638), this adhesin may act as hemagglutinin. We identified a Tn10-like transposon as an integral part of this phage. This transposon introduces resistance loci for streptomycin (strAB), tetracycline (tetDB4), sulfonamide (sul2), and cobalt-zinc-cadmium (czcAB) (53). Predictions made for the EC869 MDR genotype were validated and confirmed utilizing Biolog-derived phenotypic microarrays as well as antibiotic susceptibility assays (data not shown) compared to the P4 prophage-deficient strain EC508. The bovine tetDCBA and strB resistance loci are phage-borne, while analysis of the genome draft of EC4192 suggests a plasmid-borne origin of the tetRA and strB genes (82). The EC869 phage carries two heavy-metal resistance genes that mediate the efflux of cadmium, zinc, and cobalt (czcAB). This two-gene system is highly homologous and partly syntenic to the czcABCD resistance locus extensively studied in Alcaligenes eutrophus (63). The CzcA, CzcB, and CzcC transporters mediate Co2+, Zn2+, and Cd2+ efflux, while CzcC acts as a modifier protein required to change substrate specificity. To our knowledge, this is the first report of czc transporter loci in E. coli. The phylogenetic analysis of the cadmium/zinc antiporters is presented in Fig. S5B and C in the supplemental material.

Contact-dependent inhibition of growth (cdi), initially described in E. coli strain EC93, is advantageous to the microbe in competing for certain ecological niches (6). We discovered a P4 prophage-borne cdiAB locus that displays high homology to the corresponding locus in EC93, with the notable absence of cdiI (Fig. 3A; see also Fig. SSD and E in the supplemental

Downloaded from http://aem.asm.org/ on October 16, 2017 by guest
We speculate that this locus enables strain EC869 to compete successfully with phylogenetically diverse and abundant microorganisms that are known to colonize the bovine rumen.

**Polymorphisms in the LEE island.** Strain EC869 is distinguished by LEE polymorphisms, which are caused by the integration of an 11,098-bp P4 prophage remnant within the borders of the pathogenicity island. The integration site features a deviating G/C content neighboring a transposable element and is thus a likely target for lateral acquisition. Its insertion at the tRNA-Sec locus (ECH74115_5031) resulted in a 30-bp imperfect direct repeat (IDR) (Fig. 3B). We note that both the EC869-carried P4 prophage and smaller P4 prophage remnant are present in the draft genomes of lineage II strains FRIK2000 and FRIK966, which supports their SNP-derived phylogenetic placement (20).

**Mutator phage content and localization.** Both bovine genomes FRIK2000 and FRIK966 feature a mutator (MU) phage insertion that is not present in strain EC869. The mutator phages in strain Sakai and the 2006 Taco John outbreak isolates (EC4501, TW14588) are inserted within the ECs4942 and fbpC loci, respectively (68) (Eppinger, Mammel, LeClerc, Cebula, and Ravel, unpublished data), while this phage has yet another integration in FRIK2000, disrupting a predicted DEAD/DEAH box helicase domain gene (Escherichia coli O157_19017; introduced at the N-terminal end of a pseudo-gene, at position 5174) (see Fig. S6 in the supplemental material). We noted that all bovine isolates show a T deletion at position 1831, generating a premature stop (TAA) that results in a truncated variant of 389 aa compared to the full-length protein of 2014 aa (ECs5259). The Mu prophage is inserted at position 5174 of the helicase gene. The sequence data do not allow us to localize the Mu-like insertion for FRIK966 because of contig size and fragmentation, though from our sequence analyses, it is at a site distinct from those found for FRIK2000, Sakai, and Taco John. We detected length polymorphism in the mutator gene Mu-gp35 (ECH7EC4501_3951) due to variable numbers of a characteristic 6-bp perfect repeat (AGCCGA)\textsuperscript{5–15} (5–15 represents the range of repeat [AGCCGA] copy numbers in the analyzed Mu-gp35 mutator gene.). The analyzed mutator phage proteins (AE)\textsuperscript{5–15} within \textit{E. coli} O157:H7 range from 128 to 140 aa, while we note that this prophage is otherwise highly conserved and syntenically organized (60). The affected protein is annotated as conserved hypothetical with no assigned...
physiological function. However, this novel identified MU prophage polymorphism may serve as an additional genomic marker.

**Polymorphisms in the Enterobacteria P2 prophage.** Comprehensive analyses of an Enterobacteria P2-like prophage inserted at the yegQ locus within *E. coli* O157:H7 revealed another polymorphic region unique to the analyzed lineage II isolates (16, 44). It was noted that the deletion of 4,210 bp encompasses four genes that code hypothetical proteins with no assigned physiological function. This region may serve as an additional marker for lineage II (ECH74115_3055 to _3058), as this region is absent in the lineage I/II strain EC508 or the lineage I/II strains from the 2006 spinach (SO) and Taco Bell (TB) outbreaks.

**Association of Shiga toxin subtypes and bacterial habitat.** Carriage of both stx1 and stx2, rather than stx1 only and altered expression ratios of stx1/stx2 have been implicated in the greater virulence of human outbreak isolates and reduced virulence and impaired transmissibility in the lineage II clade (11, 22). Moreover, stx2c was identified as a key factor in HUS manifestation (56) (Eppinger, Mammel, LeClerc, Cebula, and Ravel, unpublished). An analysis of the prophage insertion sites demonstrates differences of the bovine STEC genome architecture (11). However, our findings do not support epidemiological data that show a nonrandom distribution of stx subtypes among bovine and human isolates, with stx2c typically biased toward human isolates (15, 18, 25, 26). The analyzed bovine lineage isolates contained stx1, and stx2c, much like a number of clinical isolates do, but lack the stx2 subtype. Draft genomes of the bovine isolates allowed us to locate the stx1 prophage at the yehV locus, as in clinical isolates, such as strain EDL933 (ECH7EC869_329, Escherichia coli O157 010100002259, Escherichia coli O157EcO 010100024977). Both the wrhA and argW loci that were previously identified as potential sites for stx2 prophage integrations are intact in the analyzed bovine isolates (20, 44) (Eppinger, Mammel, LeClerc, Cebula, and Ravel, unpublished). Unlike the stx1- and stx2-converting phages that are found integrated at more than one chromosomal loci in the *E. coli* O157:H7 lineage, the stx2c type preferentially targets the sbeB locus (ECH7EC869_3164). The overall stx2c prophage architecture of the bovine isolates is organized highly syntenic to other stx2c prophages compared to the SO strain EC4115 or P1717 (NC_011356, unpublished). Phage insertion resulted in 13-bp (TTTACGATTACG) perfect direct repeats that could be detected in all analyzed bovine isolates.

**Antiterminator Q polymorphisms may relate to toxicity.** Toxin production is regulated by the induction of the stx-converting phages resulting in multiplication of toxin gene copies and the transcription activator proteins Q and Q′ located upstream of the Shiga toxins (61). These antiterminator proteins control gene expression by recognizing control signals near the promoter and prevent transcriptional termination. In the stx-converting lambdoid prophages, stx expression has been genetically linked to the late antiterminator antQ and the prV promoter and associated quit site, with the tvV terminator located directly upstream of the stx coding sequences (42, 62). Two major subtypes of this regulator, antQ and antQ′, have been previously identified in stx1 and stx2c converting phages (46) (Eppinger, Mammel, LeClerc, Cebula, and Ravel, unpublished). We tried to locate the stx subtypes and investigate their association colocalization to stx-inducing antiterminator antQ among the bovine draft genomes (EC869, FRIK2000, FRIK966) and completed clinical genomes (EC4115, EDL933, and Sakai). Our comprehensive analysis found polymorphisms in the stx1 and stx2c antiterminator genes, in their respective prophage associations, and further in phage-carried insertion element insertions, all of which might be intimately associated with altered Shiga toxin production levels (21).

**Stx-converting prophage polymorphism in clinical and bovine isolates.** The stx2c-antQ′ terminator gene (ECH74115_2910, ECH7EC869_2226, Escherichia coli O157 010100022153, Escherichia coli O157EcO 010100023776) and the promoter terminator region between stx2c and antQ are genetically identical among the studied bovine lineage I and clinical lineage I/II isolates (Fig. 4). We further noted that the stx2c-converting prophage YYZ-2008 features antiterminator Q in the context of the stx2c prophage and not the typical stx2c-antQ′ combination (Fig. 4). Thus, these terminators are in general useful but not sufficient molecular markers to distinguish among Stx2-, Stx1-, and Stx2c-converting phages in the *E. coli* O157:H7 lineage. This finding is indicative of a potential antiterminator shuffling of this key regulator within the stx-converting prophage pool. Our comprehensive analysis of the toxin-producing prophages identified the transposable insertion sequence elements IS629 as a driver of stx1 and stx2c prophage microevolution in lineage II. The bovine and clinical isolates are distinguished by stx prophage-carried insertions of the IS629 element. The IS629 (ECH74115_3231, ECH74115_3232) insertion within the potentially stx1-converting phage at the yehV locus is absent in the analyzed bovine lineage II isolates (EC869, FRIK2000, FRIK966) and the closely related strain EC508 but present in all other analyzed human lineage I isolates, such as the 2006 SO isolates and the phylogenetically more distant human outbreak STEC isolates. Vice versa, the sbcB-occupying bovine stx2c phage variant carries an IS629 insertion element in strain EC869 (ECH7EC869_2207, ECH7EC869_2208), while this insertion is absent in a bovine lineage II isolate, such as in the SO- and TB-carried stx2c prophages. This insertion is present in the 62,147-bp Stx2c-converting prophage PP1717 (NC_011357, unpublished; Stx2-1717 gp55/g56) but is absent in the 54,896-bp phage YYZ2008 (NC_011356, unpublished). In strains FRIK2000 and FRIK966, the IS629 insertions are not traceable due to contig sizes and fragmentation. It was discovered that both stx phage insertion element polymorphisms may serve as additional genomic markers for lineage II isolates. We note that in both cases the IS629 insertion disrupts the prophage terminase and replication machinery, creating fragmented and truncated pseudogenes. Affected are the large prophage terminase subunit (stx2c, ECH7EC869_2206, 338 aa; ECH74115_2893, 553 aa) and replication gene repP (stx1, ECH7EC869_3230/3233), which may result in prophage immobilization. The detected polymorphisms might relate to altered stx production levels and asymptomatic manifestation (21). However, the underlying molecular mechanisms remain unclear from the current research and literature. As the alleged regulatory genes and sequences show no polymorphisms, we hypothesize that alterations in bacteriophage architectures and gene content may have altered the Stx2 toxin expression system in the bovine strains.
Understanding the genomic plasticity among the bovine and human lineages provides insights into differences in the pathogenic potential, physiology, and ecology. Here, we present multi-tier high-resolution typing approaches based on comparative genomic data that provide the basis for population-based epidemiological studies and surveillance of this genetically homogenous pathogen. Identifying genomic polymorphisms in virulence content and regulatory networks is key in understanding why cattle prevalence and incidence of human illness are of a nonlinear relationship, and likely only a subset of the STEC isolates residing in cattle may cause the majority of human disease (18). Key in studying the polymorphisms of these two STEC lineages is the association of lineage-specific genotypes with disease, strain prevalence, and source. The identified signatures primarily reside in the conserved chromosomal STEC backbone, biased by the applied SNP methodology, but also within mobile elements. Identified signatures can help to detect strain profiles that enable transmissibility from the bovine reservoir and human infectivity and monitor their prevalence in the bovine reservoir (49). The environmental conditions in bovine and human settings are also of major importance, such as the host-specific expression levels of the stx target receptors in cattle and human (69). This study could identify two parameters that might be intimately associated with Shiga toxin production in the E. coli O157:H7 lineage: first the prevalence, exchange, and polymorphism in the anti-terminator Q and Q/H11032 phage regulators and second the potential position effects due to insertion of the Stx2-converting phage at the argW and wrbA loci in clinical isolates. Next-generation sequencing will become increasingly important and enable us to include and compare in silico results with sequence data gathered from ongoing outbreaks. However, association studies in broader panels of isolates from cattle and other animal host species and environmental isolates are necessary to get insights into the genetically distinct STEC genotypes associated with differences in disease manifestation and host reservoir. The set of novel polymorphisms presented in this study complements current techniques used to classify strains and provides a basis for the phylogenomic analysis of this emerging pathogen.

ACKNOWLEDGMENT

This work was supported with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under NIAID contract N01 AI-30071.


