Evolution of Enterohemorrhagic *Escherichia coli* O26 Based on Single-Nucleotide Polymorphisms

Stefan Bletz[1], Martina Bielaszewska[1], Shana R. Leopold[1], Robin Köck[1], Anika Witten[2], Jörg Schuldes[3], Wenlan Zhang[1], Helge Karch[1], and Alexander Mellmann[1],*

[1]Institute of Hygiene, University of Münster, Germany

[2]Leibniz Institute for Arteriosclerosis, University of Münster, Germany

[3]Institute of Microbiology and Genetics, Genomic and Applied Microbiology & Göttingen Genomics Laboratory, Georg-August University Göttingen, Germany

*Corresponding author: E-mail: mellmann@uni-muenster.de.

Abstract

Enterohemorrhagic *Escherichia coli* (EHEC) O26:H11/H⁻ is the predominant non-O157 EHEC serotype among patients with diarrhea, bloody diarrhea, and hemolytic uremic syndrome (HUS) worldwide. To elucidate their phylogeny and association between their phylogenetic background and clinical outcome of the infection, we investigated 120 EHEC O26:H11/H⁻ strains isolated between 1965 and 2012 from asymptomatic carriers and patients with diarrhea or HUS. Whole-genome shotgun sequencing (WGS) was applied to ten representative EHEC O26 isolates to determine single nucleotide polymorphism (SNP) localizations within a predefined set of core genes. A multiplex SNP assay, comprising a randomly distributed subset of 48 SNPs, was established to detect SNPs in 110 additional EHEC O26 strains. Within approximately 1 Mb of core genes, WGS resulted in 476 high-quality bi-allelic SNP localizations. Forty-eight of these were subsequently investigated in 110 EHEC O26 and four different SNP clonal complexes (SNP-CC) were identified. SNP-CC2 was significantly associated with the development of HUS. Within the subsequently established evolutionary model of EHEC O26, we dated the emergence of human EHEC O26 to approximately 19,700 years ago and demonstrated a recent evolution within humans into the 4 SNP-CCs over the past 1,650 years. WGS and subsequent SNP typing enabled us to gain new insights into the evolution of EHEC O26 suggesting a common theme in this EHEC group with analogies to EHEC O157. In addition, the SNP-CC analysis may help to assess a risk in infected individuals for the progression to HUS and to implement more specific infection control measures.

Key words: enterohemorrhagic *E. coli*, EHEC O26, whole-genome shotgun sequencing, SNP typing, evolution.

Introduction

Enterohemorrhagic *Escherichia coli* (EHEC) are a highly pathogenic subgroup of Shiga toxin (Stx)-producing *E. coli*. In humans, EHEC infections cause watery and bloody diarrhea, hemolytic uremic syndrome (HUS) (Tarr et al. 2005; Mellmann et al. 2009), and is the most common cause of acute renal failure in children (Brandt et al. 1994; Kaplan 1998; Tarr et al. 2005). Although EHEC O157:H7 is the serotype most commonly associated with HUS worldwide (Banatvala et al. 2001; Robert-Koch-Institut 2008; Centers for Disease Control and Prevention [CDC] 2011), the large O104:H4 outbreak in Germany in spring 2011 (Bielaszewska et al. 2011; Mellmann et al. 2011) and several outbreaks caused by other non-O157 EHEC such as O26 (Bradley et al. 2012;

Brown et al. 2012; L'Abée-Lund et al. 2012; Wahl et al. 2011) attest to the potential menace of non-O157 EHEC. Among these, EHEC O26:H11/H⁻ (nonmotile) are the serotypes that are most frequently associated with severe human diseases in Europe (Gerber et al. 2002; Tozzi et al. 2003; Ethelberg et al. 2004; Espié et al. 2008; Mellmann et al. 2008; Zimmerhackl et al. 2010; Käppeli et al. 2011; Buvens et al. 2012) and the United States (Jelacic et al. 2003; Brooks et al. 2005; Hedican et al. 2009). Furthermore, EHEC O26 has also been increasingly detected in South American (Rivas et al. 2006), Asian (Hiroi et al. 2012), and Australian (Vally et al. 2012) patients, demonstrating the global dissemination of this EHEC serogroup. Moreover, EHEC O26 infections can be comparable with EHEC O157 infections in the severity of

the acute HUS and long-term sequelae (Gerber et al. 2002; Pollock et al. 2011; Zieg et al. 2012; Rosales et al. 2012).

The evolution of EHEC O157 was thoroughly characterized in step-wise evolutionary models, where EHEC O157 emerged from *E. coli* O55:H7 by loss and acquisition of virulence and phenotypic traits (Feng et al. 2007). This scenario was built on multilocus enzyme electrophoresis and multilocus sequence typing (MLST) data (Feng et al. 1998, 2007). Later, analyses based on multilocus variable number of tandem repeat analysis (MLVA) and single nucleotide polymorphisms (SNPs) enabled a precise reconstruction of this model and further improved branching into different and evolutionary conserved subtypes (Leopold et al. 2009; Jenke et al. 2010, 2012). This approach allowed for assigning different rates of HUS to different subtypes (Alpers et al. 2009; Jenke et al. 2010). In contrast, little is known about the evolution of EHEC O26. Recently, we identified a newly emerging, highly virulent clone within EHEC O26 based on MLST and specific virulence determinants (Bielaszewska et al. 2013); however, neither its evolutionary origin nor its reservoir is currently known. We therefore applied whole-genome shotgun sequencing (WGS) of representative EHEC O26 isolates from human diseases to develop an evolutionary model of this important pathogen and subsequently investigated the molecular epidemiology of a diverse European collection of EHEC O26 using a SNP-based assay. In addition, we examined whether the detected genotypes also reflect the presence of highly pathogenic clones within the population of EHEC O26 ultimately enabling a risk assessment in EHEC O26 infections to their progression to HUS.

Materials and Methods

Bacterial Isolates

In total, we investigated 120 EHEC O26:H11/H$^-$ isolates (supplementary table S1, Supplementary Material online). All isolates were intimin (*eae*) positive and harbored either the Stx1a-encoding gene (stx_{1a}), Stx2a-encoding gene (stx_{2a}), or both. Ten of the 120 isolates representing phylogenetic breadth based on isolation year and country (if applicable) were subjected to WGS for subsequent development of the evolutionary model and the SNP assay. For evaluation of the model and investigation of the clinical association of genotypes to certain clinical outcomes, we included a representative subset of well-characterized clinical isolates from the previously published study (Bielaszewska et al. 2013) and four isolates from asymptomatic carriers. Into this otherwise randomly selected subset of isolates, all isolates from countries other than Germany ($n = 60$) and the rare MLST sequence types (ST) STs396, 591, 1565, 1566, and 1705 were included (for details see table S1, Supplementary Material online). The chromosomal sequence of O26:H11 strain 11368 (Ogura et al. 2009) (NCBI accession number NC_013361.1) served as reference.

Whole-Genome Shotgun Sequencing, Sequence Analysis, SNP Discovery

For WGS of the ten EHEC O26 isolates, sequencing libraries were prepared using the Nextera XT chemistry (Illumina Inc., San Diego, CA, USA) for a 100 bp paired-end sequencing run on an Illumina HiScanSQ sequencer in accordance to the manufacturer's recommendations (Illumina Inc.). Sequencing reads were assembled using the CLC bio Genomic Workbench reference assembler (CLC bio, Denmark) using the chromosomal sequence of the O26:H11 strain 11368 (Ogura et al. 2009) as reference. For creation of a robust phylogeny, we extracted the core genome open reading frames (ORF) sequences starting from the previously published list of core ORFs (Mellmann et al. 2011) and included all ORFs that were present in all O26 isolates ($n = 1,130$). As an outgroup for phylogenetic analysis, the chromosomal sequence of EHEC O111:H$^-$ strain 11128, NCBI acc. no. NC_013364.1 (Ogura et al. 2009), was used. SNPs were discovered by mapping the consensus sequence of the respective isolate against the O26 reference sequence using the Ridom SeqSphere software version 0.99 beta (Ridom GmbH, Münster, Germany).

Phylogenetic Analysis

For inferring the evolutionary model of EHEC O26 based on core genome ORF sequences of the ten shotgun genome sequenced isolates and the reference strain 11368, a neighbor-joining tree was initially constructed using the MEGA5 software with default parameters (Tamura et al. 2011). We concatenated the ORFs present in all isolates and calculated the Ks values according to the modified version of the Yang-Nielsen algorithm (MYN) by using the KaKs calculator 2.0 (Zhang et al. 2006; Wang et al. 2010) for deeper analysis of the evolutionary history. This, together with an estimated synonymous substitution rate of 1.44×10^{-10} per base pair per generation (Lenski et al. 2003) and 300 generations per year for *E. coli* (in vivo) (Guttman and Dykhuizen 1994), we determined the age intervals between two isolates: Ks/ ($[1.44 \times 10^{-10}$/bp]/generation $\times$ 300 generation/year). The concatenated sequences of the intermediate isolates (postulated ancestors) were defined by the sequences of the precursor isolate within the neighbor-joining tree. To portray the SNP data from all 120 isolates, we generated a minimum spanning tree using the Ridom SeqSphere software version 0.99 beta (Ridom GmbH).

Bead-Based Multiplex Assay for SNP Detection

For robustness, the discovered SNP localizations were divided into four multiplex sets (supplementary table S4, Supplementary Material online) for subsequent detection using MagPlex-TAG microspheres on the Luminex MAGPIX platform (Luminex Corp., Austin, TX). All 96 multiplex polymerase chain reaction (PCR)-primers and 96 multiplex allele specific primer extension (ASPE) primers with the appending

of appropriate TAG sequence were designed for each set of SNPs (wild type and variant) with PrimerPlex 2.6 (PREMIER Biosoft International, Palo Alto, CA). For each SNP locus, one ASPE primer for the reference (wild type) and one ASPE primer for the allelic variant (variant) were designed. This double provision, in principle, ensures that in the SNP-screening of samples the allele calls once positively and once negatively. In this case, possible tri-allelic or even tetra-allelic polymorphisms would be found during the measurements. The multiplex PCR primers with the respective amplicon sizes and the multiplex ASPE-primers with capture sequences (TAG sequence) and corresponding bead number are shown in supplementary table S4, Supplementary Material online. The following procedure was performed with minor modifications (discussed later) in accordance with the manufacturer's recommendations (Song et al. 2010). Briefly, the multiplex PCR reaction for amplification of 12 loci per set (supplementary table S4, Supplementary Material online) was performed in 12.5 μl containing 6.25 μl REDTaq ReadyMix (Sigma-Aldrich, Munich, Germany), 1 μl of each forward and reverse primer mix (5 μM), 3.25 μl PCR water and 1 μl template DNA extracted from a single fresh colony. The PCR cycling parameters consisted of an initial step at 80 °C for 5 min, 30 cycles at 94 °C for 45 s, 60 °C for 45 s and 72 °C for 60 s, and a final step at 72 °C for 10 min. Subsequently, 5 μl of the PCR product were purified using 1.5 U Exonuclease I (E. coli) and 1.5 U Shrimp Alkaline Phosphatase (Exo/SAP) by incubation at 37 °C for 45 min followed by an inactivation step at 80 °C for 15 min. For the ASPE reaction, a 20 μl reaction contained 2 μl 10 × buffer, 1 μl MgCl$_2$ (25 mM), 1 μl dNTP mix (100 μM dTTP, dGTP, dATP; Sigma-Aldrich), 0.25 μl biotin-dCTP (400 μM, Invitrogen, Darmstadt, Germany), 1 μl ASPE primer mix (0.5 μM), 0.75 μl AmpliTaq (5 U/μl, Applied Biosystems, Foster City, CA), 9 μl PCR-water, and 5 μl purified PCR product. The cycling conditions were 96 °C for 2 min, followed by 30 cycles at 94 °C for 30 s, 60 °C for 60 s and 74 °C for 2 min. For the final hybridization step, the appropriate MagPlex TAG microspheres for each set of the 24 corresponding bead types (1,250 beads of each per reaction) were used. The hybridization mix was subjected to two washing steps and incubated in 1× Tm Hybridization buffer containing 4 μg/ml streptavidin-R-phycoerythrin conjugate (SAPE) (Invitrogen, Darmstadt, Germany) at 37 °C for 15 min. Finally, the fluorescence was measured in 50 counts within 1 minute using the xPonent 4.1 software (Luminex Corp.). Based on the median fluorescence intensity (MFI) and the net MFI, a SNP call was evaluated only if the following quality criteria were met: detection of ≥50 beads per bead type, an MFI > 300, and a ratio of MFI$_{called\ allele}$/(MFI$_{wild\ type\ allele}$ + MFI$_{variant\ allele}$) > 0.9 (Song et al. 2010). Measurements that did not fulfill these criteria were Sanger sequenced to validate the putative SNPs. For initial validation of this SNP assay, the putative SNP localizations of two samples (1226/65 and 3271/00) that differed considerably in their SNP profile were Sanger

sequenced. Moreover, reproducibility of the assay was evaluated by analyzing one O26 isolate (126814/98) starting from the extracted DNA in five independent replicates (Song et al. 2010).

Data and Statistical Analysis

The Luminex MAGPIX analysis calculates the following values with the xPonent software: MFI, net MFI, Count, Allelic Call, and Allelic Ratio. The data were exported and further processed in Microsoft Excel. We tabulated the clinical picture (HUS, bloody diarrhea, diarrhea, asymptomatic, and unknown disease) to related SNP allelic profiles to evaluate whether certain SNP genotypes were associated with a specific disease. For statistical analysis, we used Fisher's exact test of Epi Info 7 (Centers for Disease Control and Prevention, Atlanta, GA). Data were evaluated as statistically significant with P values < 0.05.

Results

Whole-Genome Shotgun Sequencing of Ten Representatives EHEC O26 Isolates for SNP Discovery

The SNP discovery was performed using WGS of ten EHEC O26 isolates listed in supplementary table S1, Supplementary Material online. After assembly, we queried the assemblies for the 1,144 ORF sequences of previously published E. coli core ORF definition (Mellmann et al. 2011). In total, 1,130 of these 1,144 ORFs were present in all O26 isolates and extracted for further analysis (supplementary table S2, Supplementary Material online). Within these 1,130 ORFs representing 952,632 bp of the chromosome, we identified 476 SNPs (in 298 ORFs) (supplementary table S3, Supplementary Material online). Of these, we selected in total 48 SNP localizations manually at random, 12 per quarter of the chromosome (fig. 1), to develop a multiplex SNP assay. All 48 SNPs were bi-allelic, synaptomorphic polymorphisms; of these 30 were nonsynonymous (ns) and 18 were synonymous (s), when compared with the reference sequence (table 1).

SNP Typing of 120 O26 Isolates Using the Multiplex Assay

To achieve the desired robustness of the assay, we divided the multilocus genotyping assay into four sets comprising 12 SNP localizations each (supplementary table S4, Supplementary Material online). To test the assay accuracy, two isolates (1226/65 and 3271/00) were initially SNP genotyped and all alleles were Sanger sequenced. In all localizations, the SNP assay was concordant with the sequencing results. To test reproducibility, one strain (126814/98) was tested in five independent repeats. Supplementary figures S1 and S2, Supplementary Material online, show the data of the average MFI minus background correction (net MFI) with standard deviation of the different SNP alleles of the 48 investigated SNPs.
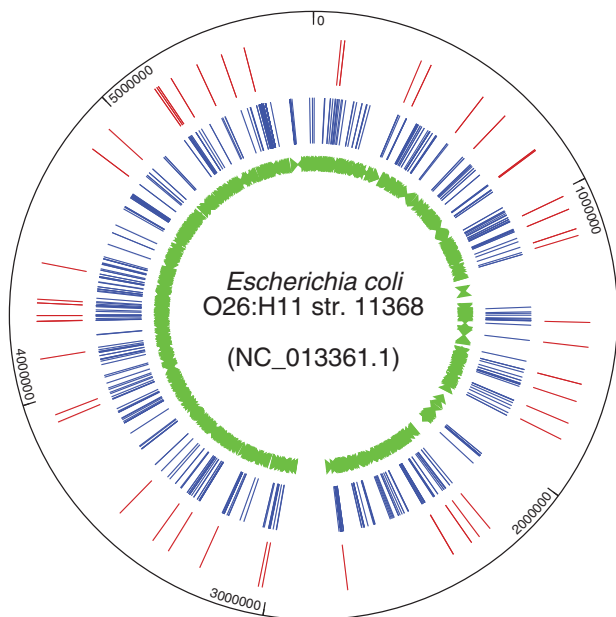
Fig. 1.—Distribution of the investigated 1,130 core genome ORFs (in green), of the discovered 476 SNPs (in blue), and the 48 SNPs of the multiplex assay (in red) illustrated in a circular map of reference genome of O26:H11 strain 11368.

In all cases, the SNP call was unambiguous and the MFI of called alleles was always at least 13-fold greater than the MFI of corresponding uncalled alleles. After this validation, a total of 5,760 SNPs were called in the 120 EHEC O26 strains. Only 2.1% (122 SNPs) had to be Sanger sequenced for confirmation, because the values were ambiguous. In all cases, however, the failure was due to mutations in the binding region of ASPE primer and sequencing confirmed the missing SNPs as known alleles (table 1).

Overall, SNP genotyping of the 120 EHEC O26 isolates resulted in ten unique SNP profiles. Their phylogenetic relationships are displayed in a minimum spanning tree (MST) in figure 2. Clustering of SNP genotypes enabled us to assign SNP clonal complexes (SNP-CCs) as phylogenetically conserved groups, which is analogous to MLST, where MLST clonal complexes are phylogenetically informative groups (Feil and Spratt 2001). Isolates sharing ≥90% of the 48 SNPs (i.e., ≥44 SNPs) were grouped, resulting in four different SNP-CCs (SNP-CC1 to SNP-CC4) (fig. 2). Further details of the 48 investigated SNP localizations, for example, their ability to serve as a canonical SNP for a certain SNP-CC, are given in table 1. Of the determined SNP-CCs, SNP-CC2 and SNP-CC3 encompassed most isolates (60 [50.0%] and 39 [32.5%] isolates, respectively). The remaining SNP-CCs contained 13 (10.1%, SNP-CC4) and 8 (6.7%, SNP-CC1) isolates. Comparison of SNP data with MLST corroborated this separation as all isolates of SNP-CC2 and SNP-CC4 were exclusively MLST ST29 and ST21, respectively. Moreover, nearly all

(36 of 39) SNP-CC3 isolates were ST21 and the majority of SNP-CC1 (6 of 8 isolates) were ST29 with few single locus variants (slv) of either ST21 (ST591, ST1565, and ST1705 in SNP-CC3) or ST29 (ST396 and ST1566 in SNP-CC1). Taken together, SNP genotyping subdivided the EHEC O26 population into four different SNP-CCs, one of which (SNP-CC2) was the recently described highly pathogenic "new clone" that separated from the remaining O26 population (Bielaszewska et al. 2013) and three further SNP-CCs.

Evolutionary Model of EHEC O26

The phylogenetic topology of the ten representative EHEC O26 isolates (supplementary table S1, Supplementary Material online), the O26:H11 reference strain 11368 and, as an outgroup, the next closely related EHEC serotype O111:H⁻ strain 11128 (NC_013364.1) (Ogura et al. 2009; Ju et al. 2012) are shown in a neighbor-joining tree (fig. 3). The branching within this tree was concordant to the separation based on the SNP assay (fig. 2) underlining the unbiased representativeness of the selected 48 SNP localizations. We propose an evolutionary model of EHEC O26 with a subdivision into four SNP-CCs. Using the Ks values (number of synonymous substitutions per synonymous site) of strains 11128 (EHEC O111) and A10, along with hypothetical intermediate isolates A01 to A10 and a common ancestor of E. coli O26 and O111 as previously proposed (Whittam et al. 1988), we postulate that E. coli O26 and O111 separated 19,700 years ago (fig. 4). Since then, EHEC O26 likely developed sequentially from SNP-CC1 to SNP-CC4. The evolution of these clonal clusters occurred within 1,650 years of this bifurcation (fig. 4). During this evolution, there was a parallel evolution of the core genome and stx as the most important virulence marker as both were almost exclusively associated in a fixed combination within the different SNP-CCs (fig. 4).

Clinical Implications of EHEC O26 Separation into Four SNP-CCs

Finally, we investigated whether the separation of EHEC O26 into SNP-CCs is associated with human diseases by analyzing the association of the SNP genotypes with clinical outcomes of the infection. Indeed, isolates of SNP-CC2, which comprised 50.0% of all strains and were responsible for 61.8% all of (47 of 76 strains) HUS cases in this study (table 2), showed a highly significant association (odds ratio 3.86, 95% confidence interval 1.63–9.30, P < 0.01) with the development of HUS. In contrast, none of the other SNP-CCs exhibited a statistically significant association with HUS (table 2).

Discussion

To elucidate the evolutionary history of EHEC O26 and to analyze more precisely the differentiation of this globally emerging pathogen into groups with different genotypic

Table 1

List of the 48 Synaptomorphic SNPs in 47 Loci in This Study Based on the Genome Sequence of O26:H11 Strain 11368 (GenBank accession number NC_013361.1)

Locus Tag	Gene (No. of SNPs)	Absolute SNP Position	SNP	SNP Effect	SNP Allele Frequency[a] (%)
ECO26_0083	*fruR* (1)	90659	T → G[b]	Synonymous	78.33
ECO26_0094	*murC* (1)	102806	T → G	Nonsynonymous (Tyr → Asp)	5.83
ECO26_0341	*ykgF* (1)	363280	T → G[c]	Synonymous	56.67
ECO26_0370	*prpC* (1)	398342	T → C[c]	Nonsynonymous (Tyr → His)	56.67
ECO26_0554	*ybcF* (1)	598509	A → G	Nonsynonymous (Thr → Ala)	5.83
ECO26_0653	*ybdK* (1)	693357	A → C[b]	Nonsynonymous (Lys → Thr)	89.17
ECO26_0785	*sdhB* (1)	841192	C → G[d]	Nonsynonymous (Asp → Glu)	6.67
ECO26_0787	*sucB* (1)	844779	T → G	Synonymous	10.00
ECO26_0968	*ybjG* (1)	1020736	T → A[d]	Synonymous	6.67
ECO26_1012	*cydC* (1)	1066493	G → C[c]	Nonsynonymous (Arg → Thr)	56.67
ECO26_1062	*ssuD* (1)	1135005	T → G[b]	Nonsynonymous (Ile → Ser)	89.17
ECO26_1083	*ycbG* (1)	1157959	A → G[b]	Nonsynonymous (Asn → Asp)	89.17
ECO26_1434	*ptsG* (1)	1447293	A → C	Nonsynonymous (Asp → Ala)	100.00
ECO26_1531	*hyaE* (1)	1528765	C → T[b]	Nonsynonymous (Ala → Val)	89.17
ECO26_1687	*minD* (1)	1652262	A → G	Synonymous	100.00
ECO26_1741	*narH* (1)	1711223	T → C[b]	Nonsynonymous (Val → Ala)	89.17
ECO26_1835	*yciK* (1)	1787173	T → C	Nonsynonymous (Met → Thr)	100.00
ECO26_1890	*ycjF* (1)	1844594	G → A[b]	Nonsynonymous (Gly → Asp)	89.17
ECO26_2286	*speG* (1)	2221328	A → T[c]	Nonsynonymous (Glu → Val)	56.67
ECO26_2339	*fumC* (1)	2265373	T → C[b]	Nonsynonymous (Leu → Pro)	89.17
ECO26_2367	*pdxH* (1)	2297263	C → G[c]	Synonymous	56.67
ECO26_2432	*ydiA* (1)	2368018	C → A[d]	Nonsynonymous (Leu → Ile)	6.67
ECO26_2433	*aroH* (1)	2368309	C → A	Nonsynonymous (Pro → Thr)	99.17
ECO26_2838	*rcsA* (1)	2735850	A → G[b]	Nonsynonymous (His → Arg)	89.17
ECO26_3081	*fruB* (1)	3018001	T → C[b]	Nonsynonymous (Val → Ala)	89.17
ECO26_3092	*yejE* (1)	3030971	A → T[b]	Nonsynonymous (Ser → Cys)	89.17
ECO26_3306	*truA* (1)	3232688	G → T[c]	Nonsynonymous (Arg → Leu)	56.67
ECO26_3433	*oxc* (1)	3349414	G → A[d]	Synonymous	6.67
ECO26_3489	*yfeG* (1)	3410529	T → G[d]	Nonsynonymous (Cys → Gly)	6.67
ECO26_3612	*recO* (1)	3551044	T → A[b]	Synonymous	89.17
ECO26_3961	*ygeY* (1)	3914494	A → G	Synonymous	88.33
ECO26_3979	*lysS* (1)	3939467	G → A[d]	Nonsynonymous (Val → Ile)	6.67
ECO26_4164	*ttdB* (1)	4129239	A → G	Synonymous	100.00
ECO26_4280#1	*glmM* (2)	4253102	T → C	Synonymous	98.33
ECO26_4280#2		4254024	T → G[c]	Synonymous	56.67
ECO26_4302	*kdsC* (1)	4272578	T → C[c]	Nonsynonymous (Val → Ala)	56.67
ECO26_4343	*tldD* (1)	4314205	T → C[c]	Nonsynonymous (Cys → Arg)	56.67
ECO26_4351	*yhdH* (1)	4326496	G → T[b]	Synonymous	89.17
ECO26_4486	*yrfF* (1)	4446501	A → C[d]	Synonymous	6.67
ECO26_4837	*gidA* (1)	4858567	T → C	Nonsynonymous (Val → Ala)	90.00
ECO26_4917	*uhpT* (1)	4943021	A → C[b]	Synonymous	89.17
ECO26_5089	*secE* (1)	5143059	C → A[b]	Nonsynonymous (Asn → Lys)	89.17
ECO26_5096	*rpoC* (1)	5151319	T → C	Synonymous	100.00
ECO26_5099	*thiG* (1)	5157488	C → T	Nonsynonymous (Leu → Phe)	90.00
ECO26_5139	*lysC* (1)	5206902	T → C[b]	Nonsynonymous (Val → Ala)	89.17
ECO26_5227	*adiC* (1)	5302486	A → G[b]	Nonsynonymous (Tyr → Cys)	89.17
ECO26_5302	*dipZ* (1)	5389233	C → G[b]	Synonymous	89.17
ECO26_5384	*cysQ* (1)	5467060	T → G[b]	Synonymous	89.17

[a]SNP allele frequency of 120 isolates used in this study.
[b]canSNP for SNP-CC4.
[c]SNP differentiating SNP-CC2 and SNP-CC3.
[d]Canonical SNP (canSNP) for SNP-CC1.
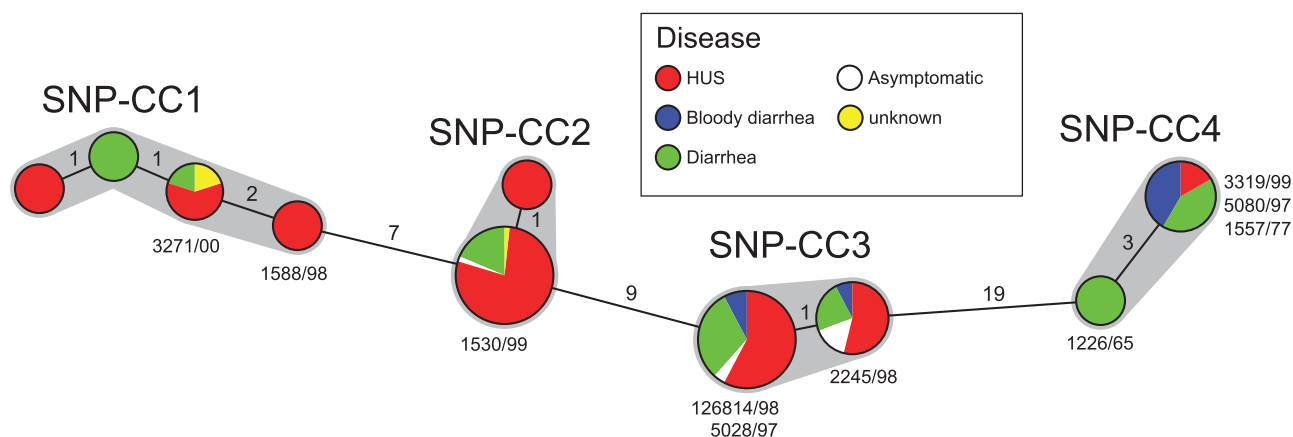
FIG. 2.—The minimum spanning tree (MST) shows the molecular phylogeny of 120 EHEC O26 isolates. The different colors represent the symptoms of the infected patients. Each node represents a unique SNP profile. SNP clonal clusters (SNP-CCs) are numbered (SNP-CC1 to SNP-CC4). The node size reflects the number of isolates. Small numbers on connecting lines display the distance (number of differing SNPs) between two nodes.
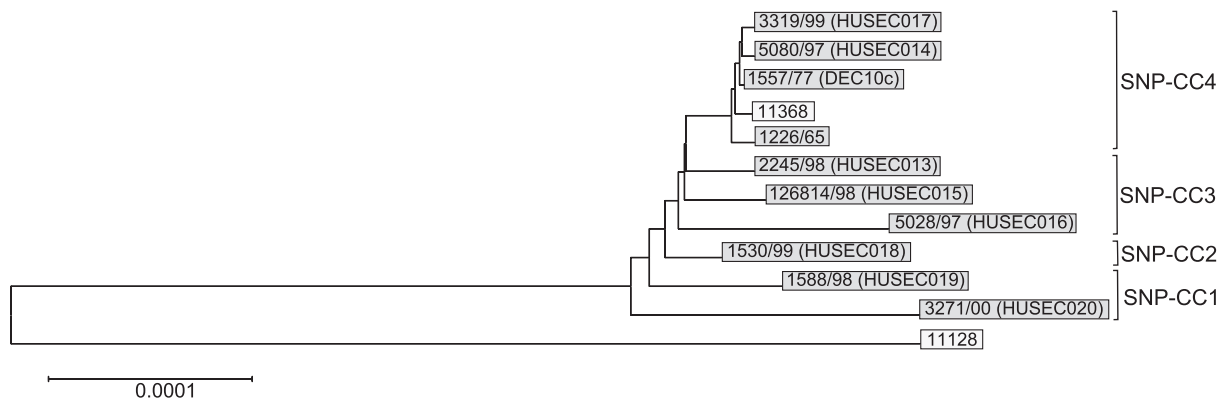

FIG. 3.—Neighbor-joining tree based on 1,130 concatenated ORFs of ten representative isolates (gray) and reference strains O26:H11 (11368) and O111:H⁻ (11128) (white). The SNP clonal clusters (SNP-CCs) are marked and demonstrate the quartering of the phylogenetic tree. Phylogenetic analysis generated by MEGA5 (Tamura et al. 2011).

and clinical characteristics, we applied WGS and SNP typing of a diverse collection of ten EHEC O26. Analysis of SNPs within 1,130 core genome genes enabled us not only to develop a multiplex assay with a reduced number of SNP localizations for high-throughput grouping of EHEC O26 into distinct and phylogenetically and clinically meaningful SNP-CCs but also to establish an evolutionary model of EHEC O26.

We were surprised by the clear-cut separation of EHEC O26 into four distinct SNP-CCs based on SNP data of the 120 strains as our previous investigations based on MLST and virulence profiling distinguished only a single highly pathogenic clone that was distinct from the remaining EHEC O26 population in Europe (Bielaszewska et al. 2013). However, this separation corroborates with studies of EHEC O157 (Manning et al. 2008; Leopold et al. 2009; Eppinger et al. 2011) and EHEC O104 (Brzuszkiewicz et al. 2011; Mellmann et al. 2011; Rasko et al. 2011), where genome sequencing information precisely refined and thereby proved evolutionary models.

From an evolutionary perspective, different scenarios could explain the separation of EHEC O26 into distinct SNP-CCs. First, an evolutionary bottleneck could have led to a reduction of the EHEC O26 population into four major genotypes. As this must have occurred in a highly specific manner favoring at least four genotypes during the emergence of EHEC O26, this scenario is unlikely. Another explanation for the emergence of different EHEC O26 SNP-CCs is the evolutionary concept of an "epidemic" population structure (Smith et al. 2000). In this model, highly adaptive and frequently pathogenic clones arise from a recombining background population for a certain time period before they disappear again in the background population because of diversification predominantly driven by recombination and secondarily by point mutations (Smith et al. 2000). In addition to the fact that *E. coli* in general exhibits frequent recombination (Wirth et al. 2006), especially between closely related members of this species (Leopold et al. 2011), it is also known that diversity and recombination
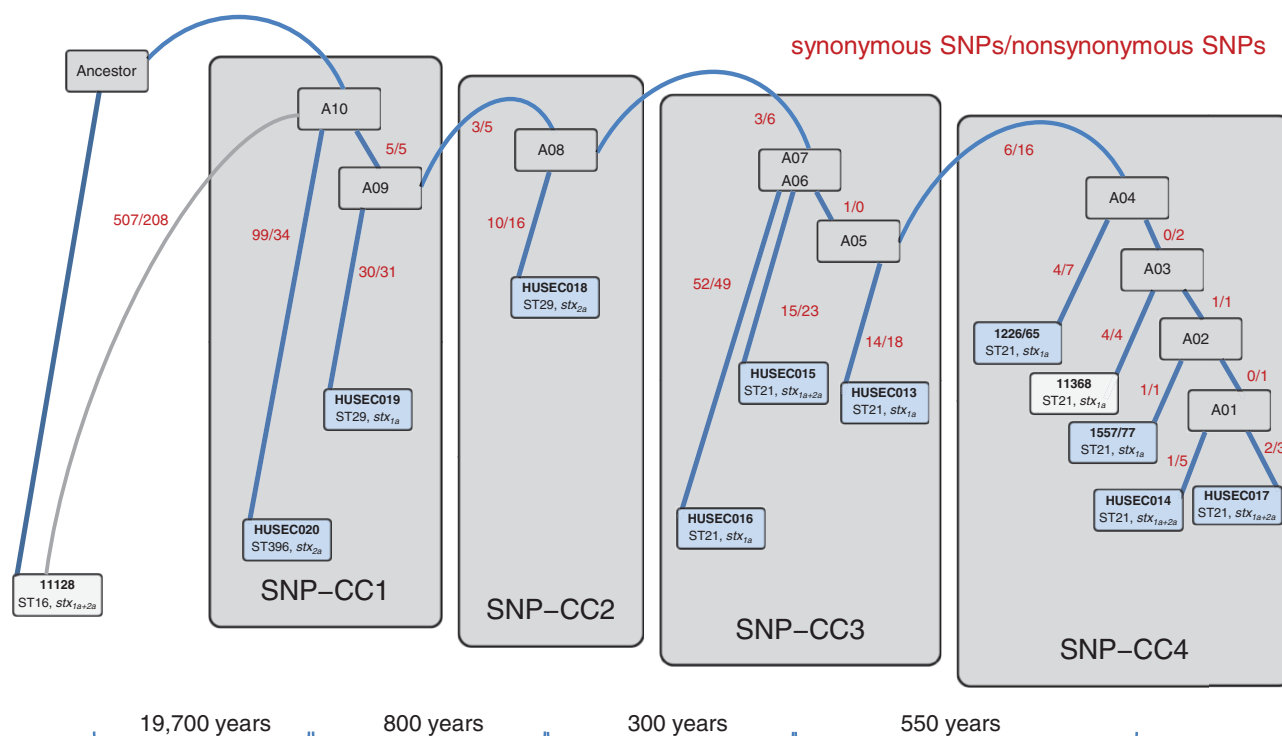
Fig. 4.—Evolutionary model and calculated age distances for EHEC O26 pathogens based on the neighbor-joining tree (fig. 3) and inserted in the SNP clonal clusters (SNP-CC1 to SNP-CC4). Blue boxes are the EHEC O26 isolates with shotgun genome sequencing data. In gray, hypothetical founders of O26 isolates are shown (A01 to A10). The ancestry is calculated based on the phylogeny displayed in figure 3. White boxes show the two EHEC O26:H11 and EHEC O111:H⁻ reference strains (strains 11368 and 11128, respectively) that are fully sequenced; EHEC O111:H⁻ is assumed to be the closest relative of serogroup O26 (Whittam et al. 1988). Blue lines connect the isolates and the hypothetical ancestors and red numbers show the synonymous/nonsynonymous SNPs between these genotypes. The gray line connects the O111:H⁻ reference strain and the first O26 ancestor A10 as the common O26/O111 ancestor is not known. Distances are not drawn to scale.

Table 2

Distribution of Diseases over Four Different SNP-CC of 120 EHEC O26 Strains Isolated from Patients and Associations of EHEC O26 SNP-CCs with HUS

SNP-CC (Total No. Isolates)	Disease (HUS/BD/D/A/U)	OR (95% CI) (HUS)	P Value
SNP-CC1 (8)	5/0/2/0/1	0.96 (0.19–5.41)	0.96
SNP-CC2 (60)	47/0/11/1/1	3.86 (1.63–9.30)	<0.01
SNP-CC3 (39)	22/3/11/3/0	0.65 (0.27–1.52)	0.28
SNP-CC4 (13)	2/5/6/0/0	0.08 (0.01–0.42)	<0.01

NOTE.—BD, bloody diarrhea; D, diarrhea; A, asymptomatic; U, unknown; OR, odds ratio; CI, confidence interval.

within tightly constrained clones such as highly pathogenic EHEC is very limited (Noller et al. 2003; Wirth et al. 2006; Manning et al. 2008; Leopold et al. 2009), thus not favoring this model. The most likely model to explain this scenario could be the model of source–sink evolution dynamics that was introduced for bacterial pathogens by Sokurenko et al. (2006) and has been described for EHEC O157 (Leopold et al. 2009) and uropathogenic *E. coli* (Chattopadhyay et al. 2007). This model postulates that a diverse population of EHEC O26

has already circulated over a longer period of time in an evolutionary stable niche (source) and only few strains were able to adapt during the transfer into a new niche (sink) with positive and purifying selection (Chattopadhyay et al. 2007). Calculation of the Ka/Ks value further corroborated this hypothesis by indicating significant purifying selection with a Ka/Ks value of 0.19 between the O26/O111 ancestor and the first strain (A10) of SNP-CC1 during the emergence of O26, that is, the transfer into the sink.

This model immediately raises the question of the natural reservoir of EHEC O26. Although cattle are the known major reservoir of EHEC O26 (Blanco et al. 2004; Geue et al. 2009; Chase-Topping et al. 2012), highly pathogenic EHEC O26 of the new clone carrying solely stx_{2a} (Bielaszewska et al. 2013) have only rarely been isolated outside humans (Allerberger et al. 2003; Blanco et al. 2004; Chase-Topping et al. 2012). Future studies applying the described SNP multiplex assay on nonhuman samples are necessary to elucidate potential reservoirs and to further understand the evolutionary dynamics between source and sink.

By calculating the Ks value, we were also able to develop a timeline of the EHEC O26 evolution (fig. 4). Using EHEC O111

as the closest relative to EHEC O26 (Whittam et al. 1988; Ogura et al. 2009), we dated the separation from a common ancestor at approximately 19,700 years ago. Estimating 200 generations/year as done for EHEC O157:H7 (Leopold et al. 2009), we calculate that EHEC O26 separated from the common EHEC O26/O111 ancestor 29,600 years ago. Interestingly, like EHEC O157 where two different groups emerged (sorbitol-fermenting O157:H⁻ [subgroup B] and nonsorbitol-fermenting EHEC O157:H7 [subgroup C]) and of which the latter further differentiated over at least 2,500 years, EHEC O26 also diverged into the four extant SNP-CCs over the course of 2,400 years.

We further calculated the association of certain EHEC disease entities (HUS vs. diarrhea without HUS and asymptomatic cases) with SNP-CCs to determine whether there is a clinical impact of the clustering into 4 SNP-CCs (table 2). Indeed, only SNP-CC2 was significantly associated with what the treating physicians termed HUS, underlining the clinical importance of the EHEC O26 grouping. Whether the infected individual finally develops a severe EHEC disease is of course also influenced by yet unknown host factors. Interestingly, the presence of certain genotypes with an increased virulence is again similar to EHEC O157 (Manning et al. 2008; Jenke et al. 2010, 2012). Altogether, these observations suggest a common theme in the EHEC evolution, which is driven by the transfer into new hosts, that is, the humans, and rapid selection processes.

One limitation of our study might be a potential sampling bias toward isolates from severely ill patients. However, SNP data of the diverse collection of strains spanning several decades and different countries still reflected this separation into four SNP-CCs. Moreover, also the four isolates from asymptomatic carriers shared the identical SNP genotype with isolates from severely ill patients further corroborating the separation into the four SNP-CCs. A second limitation might be the limited geographical distribution of the strains used to determine the evolutionary model as, with the exception of one strain, all were isolated from patients in Germany. An inclusion of additional strains from different continents may provide additional information; however, grouping of the SNP genotypes of the 110 isolates from seven European countries into the four SNP-CCs also approved our model at least for Europe. Another limitation might be the inclusion of only 1,130 ORFs; we were, however, not aiming for separation of closely related strains during outbreak investigations but used these genes that are conserved within *E. coli* solely for generation of a robust phylogenetic signal.

In summary, based on WGS and subsequent multiplexed SNP calling, we established an evolutionary model of the emergence and further diversification of EHEC O26 into four phylogenetically and clinically meaningful SNP-CCs. These data broadened our knowledge about the evolution of this important human pathogen and suggest a common theme in the EHEC evolution. Moreover, information about the SNP-CC may help implement more specific infection control measures and may enable a risk assessment for each detected EHEC O26 isolate. Future studies should also focus on EHEC O26 in nonhuman environments to understand their behavior and evolutionary processes in their likely reservoirs.

Supplementary Material

Supplementary tables S1–S4 and figures S1 and S2 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

Literature Cited

Allerberger F, et al. 2003. Hemolytic-uremic syndrome associated with enterohemorrhagic *Escherichia coli* O26:H infection and consumption of unpasteurized cow's milk. Int J Infect Dis. 7:42–45.

Alpers K, et al. 2009. Sorbitol-fermenting enterohaemorrhagic *Escherichia coli* O157:H⁻ causes another outbreak of haemolytic uraemic syndrome in children. Epidemiol Infect. 137:389–395.

Banatvala N, et al. 2001. The United States National Prospective Hemolytic Uremic Syndrome Study: microbiologic, serologic, clinical, and epidemiologic findings. J Infect Dis. 183:1063–1070.

Bielaszewska M, et al. 2011. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect Dis. 11: 671–676.

Bielaszewska M, et al. 2013. Enterohemorrhagic *Escherichia coli* O26:H11/H⁻: a new virulent clone emerges in Europe. Clin Infect Dis. 56:1373–1381.

Blanco M, et al. 2004. Serotypes, virulence genes, and intimin types of Shiga toxin (verotoxin)-producing *Escherichia coli* isolates from cattle in Spain and identification of a new intimin variant gene (*eae*-ξ). J Clin Microbiol. 42:645–651.

Bradley KK, et al. 2012. Epidemiology of a large restaurant-associated outbreak of Shiga toxin-producing *Escherichia coli* O111:NM. Epidemiol Infect. 140:1644–1654.

Brandt JR, et al. 1994. *Escherichia coli* O157:H7⁻ associated hemolytic-uremic syndrome after ingestion of contaminated hamburgers. J Pediatr. 125:519–526.

Brooks JT, et al. 2005. Non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States, 1983–2002. J Infect Dis. 192: 1422–1429.

Brown JA, et al. 2012. Outbreak of Shiga toxin-producing *Escherichia coli* serotype O26: H11 infection at a child care center in Colorado. Pediatr Infect Dis J. 31:379–383.

Brzuszkiewicz E, et al. 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. 193:883–891.

Buvens G, et al. 2012. Incidence and virulence determinants of verocytotoxin-producing *Escherichia coli* infections in the Brussels-Capital Region, Belgium, in 2008–2010. J Clin Microbiol. 50:1336–1345.

Centers for Disease Control and Prevention (CDC). 2011. Vital signs: incidence and trends of infection with pathogens transmitted commonly through food—foodborne diseases active surveillance network, 10 U.S. sites, 1996–2010. MMWR Morb Mortal Wkly Rep. 60: 749–755.

Chase-Topping ME, et al. 2012. Pathogenic potential to humans of bovine *Escherichia coli* O26, Scotland. Emerg Infect Dis. 18:439–448.

Chattopadhyay S, et al. 2007. Haplotype diversity in 'source-sink' dynamics of *Escherichia coli* urovirulence. J Mol Evol. 64:204–214.

Eppinger M, Mammel MK, Leclerc JE, Ravel J, Cebula TA. 2011. Genomic anatomy of *Escherichia coli* O157:H7 outbreaks. Proc Natl Acad Sci U S A. 108:20142–20147.

Espié E, et al. 2008. Surveillance of hemolytic uremic syndrome in children less than 15 years of age, a system to monitor O157 and non-O157 Shiga toxin-producing *Escherichia coli* infections in France, 1996-2006. Pediatr Infect Dis J. 27:595–601.

Ethelberg S, et al. 2004. Virulence factors for hemolytic uremic syndrome, Denmark. Emerg Infect Dis. 10:842–847.

Feil EJ, Spratt BG. 2001. Recombination and the population structures of bacterial pathogens. Annu Rev Microbiol. 55:561–590.

Feng P, Lampel KA, Karch H, Whittam TS. 1998. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. J Infect Dis. 177:1750–1753.

Feng PCH, et al. 2007. Genetic diversity among clonal lineages within *Escherichia coli* O157:H7 stepwise evolutionary model. Emerg Infect Dis. 13:1701–1706.

Gerber A, Karch H, Allerberger F, Verweyen HM, Zimmerhackl LB. 2002. Clinical course and the role of Shiga toxin–producing *Escherichia coli* infection in the hemolytic-uremic syndrome in pediatric patients, 1997–2000, in Germany and Austria: a prospective study. J Infect Dis. 186:493–500.

Geue L, et al. 2009. Analysis of the clonal relationship of serotype O26:H11 enterohemorrhagic *Escherichia coli* isolates from cattle. Appl Environ Microbiol. 75:6947–6953.

Guttman DS, Dykhuizen DE. 1994. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science 266:1380–1383.

Hedican EB, et al. 2009. Characteristics of O157 versus non-O157 Shiga toxin-producing *Escherichia coli* infections in Minnesota, 2000-2006. Clin Infect Dis. 49:358–364.

Hiroi M, et al. 2012. Serotype, Shiga toxin (Stx) type, and antimicrobial resistance of Stx-producing *Escherichia coli* isolated from humans in Shizuoka Prefecture, Japan (2003-2007). Jpn J Infect Dis. 65: 198–202.

Jelacic JK, et al. 2003. Shiga toxin-producing *Escherichia coli* in Montana: bacterial genotypes and clinical profiles. J Infect Dis. 188:719–729.

Jenke C, et al. 2010. Phylogenetic analysis of enterohemorrhagic *Escherichia coli* O157, Germany, 1987-2008. Emerg Infect Dis. 16: 610–616.

Jenke C, et al. 2012. Identification of intermediate in evolutionary model of enterohemorrhagic *Escherichia coli* O157. Emerg Infect Dis. 18: 582–588.

Ju W, et al. 2012. Phylogenetic analysis of non-O157 Shiga toxin-producing *Escherichia coli* strains by whole-genome sequencing. J Clin Microbiol. 50:4123–4127.

Kaplan BS. 1998. Shiga toxin-induced tubular injury in hemolytic uremic syndrome. Kidney Int. 54:648–649.

Käppeli U, Hächler H, Giezendanner N, Beutin L, Stephan R. 2011. Human infections with non-O157 Shiga toxin-producing *Escherichia coli*, Switzerland, 2000-2009. Emerg Infect Dis. 17:180–185.

L'Abée-Lund TM, et al. 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.

Lenski RE, Winkworth CL, Riley MA. 2003. Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. J Mol Evol. 56:498–508.

Leopold SR, Sawyer SA, Whittam TS, Tarr PI. 2011. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. BMC Evol Biol. 11:183.

Leopold SR, et al. 2009. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc Natl Acad Sci U S A. 106:8713–8718.

Manning SD, et al. 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. Proc Natl Acad Sci U S A. 105:4868–4873.

Mellmann A, Bielaszewska M, Karch H. 2009. Intrahost genome alterations in enterohemorrhagic *Escherichia coli*. Gastroenterology 136: 1925–1938.

Mellmann A, et al. 2008. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic *Escherichia coli*. Emerg Infect Dis. 14:1287–1290.

Mellmann A, et al. 2011. Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. PLoS One 6:e22751.

Noller AC, et al. 2003. Multilocus sequence typing reveals a lack of diversity among *Escherichia coli* O157:H7 isolates that are distinct by pulsed-field gel electrophoresis. J Clin Microbiol. 41:675–679.

Ogura Y, et al. 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. 106:17939–17944.

Pollock KGJ, et al. 2011. Highly virulent *Escherichia coli* O26, Scotland. Emerg Infect Dis. 17:1777–1779.

Rasko DA, et al. 2011. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. N Engl J Med. 365: 709–717.

Rivas M, et al. 2006. Characterization and epidemiologic subtyping of Shiga toxin-producing *Escherichia coli* strains isolated from hemolytic uremic syndrome and diarrhea cases in Argentina. Foodborne Pathog Dis. 3:88–96.

Robert-Koch-Institut. 2008. Erkrankungen durch Enterohämorrhagische *Escherichia coli* (EHEC). Epid Bull. 2:11–15.

Rosales A, et al. 2012. Need for long-term follow-up in enterohemorrhagic *Escherichia coli*-associated hemolytic uremic syndrome due to late-emerging sequelae. Clin Infect Dis. 54:1413–1421.

Smith JM, Feil EJ, Smith NH. 2000. Population structure and evolutionary dynamics of pathogenic bacteria. Bioessays 22:1115–1122.

Sokurenko EV, Gomulkiewicz R, Dykhuizen DE. 2006. Source-sink dynamics of virulence evolution. Nat Rev Microbiol. 4:548–555.

Song Y, et al. 2010. A multiplex single nucleotide polymorphism typing assay for detecting mutations that result in decreased fluoroquinolone

susceptibility in *Salmonella enterica* serovars Typhi and Paratyphi A. J Antimicrob Chemother. 65:1631–1641.

Tamura K, et al. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. 28:2731–2739.

Tarr PI, Gordon CA, Chandler WL. 2005. Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome. Lancet 365: 1073–1086.

Tozzi AE, et al. 2003. Shiga toxin-producing *Escherichia coli* infections associated with hemolytic uremic syndrome, Italy, 1988–2000. Emerg Infect Dis. 9:106–108.

Vally H, et al. 2012. Epidemiology of Shiga toxin producing *Escherichia coli* in Australia, 2000–2010. BMC Public Health. 12:63.

Wahl E, Vold L, Lindstedt BA, Bruheim T, Afset JE. 2011. Investigation of an *Escherichia coli* O145 outbreak in a child day-care centre—extensive sampling and characterization of eae- and stx1-positive *E. coli* yields epidemiological and socioeconomic insight. BMC Infect Dis. 11:238.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 8:77–80.

Whittam TS, Wachsmuth IK, Wilson RA. 1988. Genetic evidence of clonal descent of *Escherichia coli* O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome. J Infect Dis. 157:1124–1133.

Wirth T, et al. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 60:1136–1151.

Zhang Z, Li J, Yu J. 2006. Computing Ka and Ks with a consideration of unequal transitional substitutions. BMC Evol Biol. 6:44.

Zieg J, et al. 2012. Fatal case of diarrhea-associated hemolytic uremic syndrome with severe neurologic involvement. Pediatr Int. 54: 166–167.

Zimmerhackl L-B, et al. 2010. Enterohemorrhagic *Escherichia coli* O26:H11⁻ associated hemolytic uremic syndrome: bacteriology and clinical presentation. Semin Thromb Hemost. 36:586–593.

Associate editor: John McCutcheon