Recent O-antigen diversification masks highly pathogenic STEC O104:H4

Christina Lang[1], Angelika Fruth[1], Ian Winsten Campbell[2], Claire Jenkins[3], Peyton Smith[4], Nancy Strockbine[4], François-Xavier Weill[5], Ulrich Nübel[6, 7, 8], Yonatan H. Grad[9], Matthew K. Waldor[2,9,10], and Antje Flieger[1#]

[1] Robert Koch-Institut, Division of Enteropathogenic Bacteria and *Legionella*, National Reference Centre for *Salmonella* and other Enteric Bacterial Pathogens, Burgstr. 37, 38855 Wernigerode, Germany

[2] Brigham & Women's Hospital, Division of Infectious Diseases, Harvard Medical School, Department of Microbiology, Boston, MA 02115, United States of America

[3] Gastro and Food Safety (One Health) Division, Health Security Agency, London NW9 5HT, United Kingdom

[4] Centers for Disease Control and Prevention, Division of Foodborne, Waterborne and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, 1600 Clifton Road, Atlanta, GA, United States of America

[5] Institut Pasteur, Université Paris Cité, Unité des bactéries pathogènes entériques, Paris, F-75015, France

[6] Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

[7] German Center for Infection Research (DZIF), Partner Site Braunschweig-Hannover, Germany

[8] Braunschweig Integrated Center of Systems Biology (BRICS), Technical University, Braunschweig, Germany

[9] Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA 02115, United States of America

[10] Howard Hughes Medical Institute, Boston, MA 02115, United States of America

[#] corresponding author contact: email: fliegera@rki.de, Tel.: +49-30-18754-2522

Abstract

Background: Shiga toxin-producing *E. coli* (STEC) can give rise to a range of clinical outcomes from diarrhea to the life-threatening systemic condition, hemolytic uremic syndrome (HUS). A major outbreak of HUS occurred in 2011, and was caused by a rare serotype, STEC O104:H4. Prior to 2011 and since the outbreak, STEC O104:H4 were rarely associated with human infections.

Methods: From 2012 to 2020 intensified STEC surveillance was performed in Germany where subtyping of ~8,000 clinical isolates by molecular methods including whole genome sequencing was carried out. Virulence traits and phylogenetic context were investigated for a subset of strains.

Results: A rare STEC serotype O181:H4 associated with HUS was identified, belonging to sequence type (ST) 678, like the STEC O104:H4 outbreak strain. Virulence and genomic comparisons revealed that the two strains are phylogenetically related and differ principally in the gene cluster encoding their respective lipopolysaccharide O-antigens. In addition, five other serotypes belonging to ST678 from human clinical infection were identified from diverse locations worldwide.

Conclusion: Our data suggest the high virulence ensemble of STEC O104:H4 remains a global threat, but that horizontal exchange of O-antigen gene clusters has cloaked the pathogen with new O-antigens, confounding interpretation of their potential risk.

Keywords

Shiga toxin-producing *E. coli*, O104:H4, O-antigen diversification, phylogeny, risc profiling

STEC are food-borne pathogens responsible for a range of clinical syndromes from diarrhea to the life-threatening systemic condition, HUS, a triad of thrombotic microangiopathy, thrombocytopenia, and acute renal injury [1]. Historically, the classification of STEC strains into different serotypes has proven invaluable for epidemiology and risk profiling [2]. In *E. coli* the serotype is determined by a combination of O- and H-antigen types (see below) whereas the O group solely denotes the O-antigen. Globally, strains of STEC serotype O157:H7 are most frequently associated with HUS, but STEC belonging to O groups O26, O103, O111, and O145, have also been regularly linked to HUS development [2, 3]. In addition, a very rare serotype gave rise to a major HUS outbreak in early summer 2011, when an O104:H4 strain caused more than 3,000 cases of diarrhea and 800 cases of HUS including 54 fatalities, predominantly in Germany [4-6].

The O104:H4 outbreak strain encodes an exceptional set of virulence features [5-9]. Like other HUS-associated strains, the outbreak strain produces Shiga toxin (Stx), specifically the Stx2a variant. But unlike most *E. coli* strains causing HUS, this strain is an enteroaggregative *E. coli* (EAEC) and lacks a prime virulence factor, the type III secretion system encoded on a pathogenicity island designated as the locus of enterocyte effacement (LEE). The O104:H4 outbreak strain along with other O104:H4 strains, all belonging to multi-locus sequence type (MLST) ST678 and some possessing *stx*, formed a distinct clade among EAEC [5, 7, 10]. The outbreak strain harbors a plasmid (pAA) characteristic for EAEC encoding aggregative adhesion fimbriae of type I (AAF/I) [5, 7]. Furthermore, the outbreak strain encodes the virulence-linked serine-protease autotransporters (SPATE), SepA, SigA, and Pic [5, 10, 11]. Despite the extensive outbreak in May/June of 2011 and associated wide distribution of the strain in affected regions, intense molecular surveillance only uncovered relatively few O104:H4 cases in Germany after the outbreak dissipated by July 2011 [12].

In *E. coli*, serotypes are determined by the composition of the lipopolysaccharide (LPS) O-antigen and the flagellar H-antigen, both important surface features of microorganisms that

shape pathogen host interactions [13, 14]. LPS forms a major structural component of the Gram-negative cell's outer membrane and its most distal part is the O-antigen. The O-antigen is subject to strong selection pressure and is one of the most variable components of the bacterial cell [13]. Typically, in *E. coli* the O-antigen consists of chains of repeating oligosaccharide subunits, usually composed of two to seven sugars often with additional chemical modifications [15]; currently, 182 O groups and 53 flagellar antigen types have been described by phenotypic identification [14, 16]. The genes encoding O-antigen biosynthesis are organized in clusters that are flanked by colanic acid (*wca*) and histidine (*his*) biosynthesis genes [15]. O-antigen biosynthesis gene clusters typically have a lower GC content (often <40%) than that of the backbone of the *E. coli* chromosome, which is ~50% GC content [13, 17, 18]. These differences suggest that O-antigen biosynthesis gene clusters are exchanged by lateral gene transfer and are under diversifying selection, and therefore a hot spot of recombination [15, 19].

Although its wide distribution in the affected areas, the near disappearance of *E. coli* O104:H4 in Germany after the large outbreak in 2011 was unanticipated. Here, we show that the high virulence ensemble of the O104:H4 outbreak strain remains a threat, but that O-antigen gene exchange has cloaked the pathogen with several new O-antigens.

Methods

Ethical statement

Rabbit studies conducted according to protocols reviewed and approved by Brigham and Women's Hospital Committee on Animals (IACUC protocol 2016N000334) and Animal Welfare Assurance of Compliance (number A4752-01) in accordance with recommendations in the Guide for the Care and Use of Laboratory Animals of the National Institutes of Health and the Animal Welfare Act of the U.S. Department of Agriculture.

Study strains

In the context of intensified STEC surveillance in Germany, clinical isolates were collected at the National Reference Center for *Salmonella* and other Bacterial Enteric Pathogens and analyzed for serotype, *stx* and subtypes, *eaeA*, *hlyA* and *aatA* as described [20]. Genome sequencing was carried out on a subset of strains listed together with open source and study strain data in Table S1 and S2. Strains were grown on nutrient agar (Oxoid GmbH, Germany), Luria Bertani (LB) broth, or enterohemolysin agar (Sifin GmbH, Germany).

Whole genome sequencing (WGS)

Long read whole genome sequencing of O181:H4 strain 17-07187 was performed by GATC Biotech (Konstanz, Germany) using a PacBio RS II sequencer (Pacific Biosciences, USA) and short read genome sequencing of strains 12-04810, 14-03615, 14-01288, 16-00596, 16-01499, 16-05332, 17-01774, 17-00416, 17-07187 and 19-02696 (Table S1) on an Illumina MiSeq and HiSeq 1500 benchtop sequencer. Polishing of the assembled genome and plasmids was performed with Illumina short reads by Pilon (version 1.22) [21]. The sequences were uploaded to NCBI project: PRJNA833419.

Bioinformatics analyses

De novo assembly of the PacBio sequence data (103-fold average coverage) was performed by GATC utilizing HGAP3 (Pacific Biosciences, USA). Quality control and trimming of MiSeq raw reads with subsequent detection of serotype and virulence genes was performed as described [16]. Genomic comparisons were carried out using MAUVE (version: 1.1.1) and MAFFT (version 1.3.7) as plugin in Geneious (version: 11.1.5; Biomatters Ltd) [22, 23]. Ridom SeqSphere+ (version:7.2.0, Ridom GmbH, Germany) was used to determine MLST Warwick sequence types and to create minimal spanning trees based on 2513 allele targets from the *E. coli* cgMLST Enterobase with pairwise ignoring missing values from genome assemblies [24]. Phage prediction was carried out by analysing the genome sequences with PHASTER [25]. RAST was used for CDS annotation [26].

138

139 **SNP-based alignment and maximum likelihood based phylogenetic tree**

140 Mapping of sequencing reads, generation of consensus sequences, and alignment calculation
141 was performed using the BatchMap pipeline [27]. The genome sequence of FWSEC0009 was
142 used as reference and SNPs were filtered using Gubbins (version: 3.2.1) [28]. Alignment of
143 filtered SNPs by Gubbins was used to generate a maximum likelihood based phylogenetic tree
144 by PHYML 3.3.20180214 (Geneious plugin, substitution model: HKY85, 100 bootstraps)[29].

145

146 **Temporal signal and 'clocklikeness' of molecular phylogenies**

147 TempEst was used to analyze the RAxML tree generated by Gubbins in conjunction with
148 collection year data to validate the molecular-clock assumption [30]. Best-fitting root was
149 chosen for linear regression analyses.

150

151 **Cytotoxicity, adherence, and infection assays**

152 Viability of Vero cells after inoculation with diluted bacterial culture supernatants (1:200) was
153 examined using 3-[4,5-dimethylthiazole-2-yl]-2,5-diphenyltetrazolium bromide [5]. Bacteria
154 and Vero cells were prepared as described [20]. Adherence to HEp-2 cells was performed as
155 described [27]. For infant rabbit infection assays, litters of mixed gender 2-day-old New
156 Zealand White infant rabbits with the lactating doe were acquired from Charles River (Canada,
157 strain code 052). Infant rabbits were orogastrically inoculated on the day of arrival with 10^9
158 CFU of Streptomycin-resistant strains O104:H4 C227-11 and O181:H4 17-07187 suspended
159 in 500µl 2.5% sodium bicarbonate (pH9) using a size 4 French catheter as described
160 previously, except that no ranitidine was administered [11]. Infant rabbits were monitored for
161 signs of illness and euthanized three days post infection. Tissue samples taken from the
162 stomach, small intestine, cecum, and colon were homogenized and CFU determined by serial
163 dilution, and plating on LB media containing 200 µg/ml streptomycin [11].

164

165

Results

The HUS-associated STEC O181:H4 strain 17-07187 shares a close phylogenetic relationship and similar virulence traits with the O104:H4 outbreak strain

After the large STEC O104:H4 outbreak in 2011, the German National Reference Centre for *Salmonella* and other Bacterial Enteric Pathogens intensified STEC surveillance and analyzed ~ 8,000 clinical isolates primarily from diarrhea and HUS patients from 2012 to 2020. This strain collection included a stool sample isolate (17-07187) from a 6-year-old girl who had bloody diarrhea and HUS in December 2017. She had not travelled outside her home in Northwest Germany before becoming ill. Serotyping and whole genome sequencing revealed that the strain belonged to an unusual serotype, O181:H4, that had not been previously associated with HUS. The strain had *stx2a* but lacked the type III secretion system encoded on the LEE pathogenicity island (marker gene *eae*) found in STEC. Furthermore, the strain encoded characteristic EAEC markers including *aatA, aggR,* AAF/I genes and autotransporter protease genes *pic, sigA,* and *sepA* (Fig.1A).

MLST demonstrated that the 17-07187 isolate belonged to the same sequence type (ST678) as the O104:H4 outbreak strain (Table S2) [5, 7, 9]. Core genome MLST (cgMLST) [24] confirmed the close phylogenetic relationship of this isolate with a panel of O104:H4 strains (Fig.1B). There were only 30-34 allelic differences between this O181:H4 isolate and the O104:H4 outbreak strain from 2011 (e.g. strains FWSEC0009, C227-11, or 11-02027), and O104:H4 clinical isolates from the Republic of Georgia in 2009 (2009EL-2071) and France in 2012 (Ec12-0465). In contrast, there was considerable phylogenetic distance to STEC serotypes O181:H16 (ST6274), O181:H49 (ST173), O104:H21 (ST672), and O104:H7 (ST10075) isolated between 2012 and 2019 (allelic distance >1500) (Fig.1B). Comparison of the virulence gene repertoire of the O181:H4 and the O104:H4 outbreak strain also strongly suggested that they rely on very similar virulence mechanisms (Fig.1A). Indeed, both strains had comparable Stx-related cytotoxicity (Fig.1C), exhibited a characteristic enteroaggregative adherence pattern (Fig.1D), and colonized intestinal tissues, particularly the cecum and colon,

similarly during *in vivo* infant rabbit infections (Fig.1E). Thus, STEC O181:H4 and O104:H4 isolates share marked genomic similarity and virulence-associated genomic and phenotypic traits.

The genomes of the STEC O181:H4 strain 17-07187 and the O104:H4 outbreak strain mainly differ in their O-antigen gene clusters and mobile genetic elements

The chromosomes (without plasmids) of the O181:H4 strain 17-07187 and the O104:H4 outbreak strain FWSEC0009 were nearly identical (~94.5% nucleotide identity). The most striking difference between them were their respective O-antigen gene clusters (OAGC) (Fig.2A, 2B). Although these two clusters were both situated in the same location in the chromosome, between *galF* and *hisI* (Fig.2B), their gene contents and organization were very different. Furthermore, their respective GC contents, 36.8% for O181 and 37% for O104, differed from the chromosome GC content (~50.7%), highlighting the likely role of lateral gene transfer in driving OAGC exchange.

14 potential prophage regions are present in the O181:H4 strain and 16 in the O104:H4 outbreak strain (Fig.2A, Table S3). 11 of the 14 prophages exhibited substantial sequence identity (83-99.9%) with their O104:H4 counterparts and importantly, the *stx2*-encoding prophages were nearly identical (~99% identity) (Fig.2C, Table S4, Fig.2C). Both *stx* phages are inserted into the tryptophan repressor binding protein gene *wrbA.*

The genome of the O181:H4 strain included three plasmids of ~81kb, ~76kb, and ~63kb (Table S2). The largest O181:H4 plasmid (plasmid 1) was an incompatibility group I1 (IncI1) plasmid that showed only partial homology to the ESBL resistance plasmid of the O104:H4 outbreak strain (~57% nucleotide identity) but was very similar (> 90% nucleotide identity) to pHUSEC41-1 of STEC O104:H4 HUSEC41 from 2001 (92kb) (Fig.S1A) [3, 31]. Both of these plasmids encode the *pilI-V* genes for thin pili. Unlike pHUSEC041-1, the O181:H4 plasmid 1 did not contain antibiotic resistance genes (Fig.S1A, Tab.S5). O181:H4 plasmid 2 (76 kb) was nearly identical (95% identity, 100% coverage) to the O104:H4 outbreak strain pAA (pAA-

EA11) and harbored virulence associated loci including *aggA/B/C/D*, which encode the AAF/I fimbriae that promote bacterial adherence to host cells (Fig.1A, Fig.S1B, Tab.S6) [7, 10, 32]. O181:H4 plasmid 3 (63 kb) was not found in the O104:H4 outbreak strain; instead, it showed similarity to DHA plasmids of several enterobacteria coding for AmpC β-lactamase [33]. However, unlike the DHA plasmids, resistance determinants were not present in the O181:H4 strain (Fig.S1C, Tab.S7). Together these observations reveal the striking similarity of the chromosomes of the O181:H4 strain and the O104:H4 outbreak strain and that their chief differences are confined to hot spots of recombination, i.e. their O-antigen gene clusters, and to mobile genetic elements, particularly their plasmids.

Additional recent global isolates of serotype O181:H4 and five other O groups belong to ST678

Next, we identified 158 genomes belonging to ST678 in Enterobase [24], which contains ~202,200 *E. coli* genomes (as of April 11th, 2022). For a subset, serotype identification was not available, and for these cases, serotype was predicted based on the available genomic information[16]. 123 of the ST678 strains were O104:H4, however, 18 additional O181:H4 strains of ST678 were found. Furthermore, seven O127:H4, three O131:H4, and one each of O69:H4 and OX13:H4 were identified (Fig.3A, Tab.S1). We categorized these as non-O104:H4 ST678 strains. Additionally, based on a close phylogenetic relationship and single difference in MLST alleles, we added three non-ST678 strains to the non-O104:H4 ST678 category: two O181:H4 strains (1472912 and 1472968) and one strain of the new and provisionally assigned genoserotype OgN-RKI9:H4 (strain 608450) (Fig.3A, Tab.S1).

cgMLST confirmed the close phylogenetic relationship (allelic distance ~50) between AAF/I gene positive O104:H4 strains including the outbreak strain (FWSEC0009), and the 21 O181:H4 strains, the OX13:H4 strain, and three of the seven O127:H4 strains (Fig.3A, 3B). All of these strains showed AAF/I which were also found in the 2011 outbreak strain [5, 7]. Nine strains (four O127:H4, three O131:H4, one each of O69:H4 and OgN-RKI9:H4) had higher allelic distances up to 189. In contrast, these nine strains encoded AAF/III genes (Fig.3A, 3B).

The 25 AAF/I positive non-O104:H4 ST678 strains were all isolated in or after 2011 and were associated with diarrheal disease and a subset of six O181:H4 strains harbored *stx2a* (Fig.3A, 3B, Fig.S2A, and Tab.S1). Interestingly, four of these six strains shared a very similar *stx* phage with the O104:H4 outbreak strain (including strain 17-07187 from Germany), but two of the six prophages were more distinct (Fig.3A). AAF/I-positive strains shared regions with a higher relatedness than those positive for AAF/III with the pAA plasmid of the 2011 outbreak strain.

The virulence gene profiles of the 34 non-O104 ST678 strains were generally similar to the O104:H4 outbreak strain; however, there were a few differences. Specifically, *sepA* was exclusively found in AAF/I-positive strains and EAST1 was present in all AAF/III-positive strains and in only three of the AAF/I-positive isolates (Tab.S8).

The 34 non-O104:H4 ST678 strains were isolated in countries of Europe, Africa, and North and South America. In addition, several were from individuals with a travel history that might link these to East Asia (Fig.3A, S2B and C). Together, these observations show that ST678 *E. coli* strains are found among seven different *E. coli* serotypes that have been linked to diarrheal disease on several continents primarily during the past decade.

Phylogenomic analyses of ST678 *E. coli* suggest recency of O-antigen gene exchange

The O-antigen gene clusters encoding the six O groups in the 34 non-O104:H4 ST678 strains are found in the same chromosomal location as the 2011 O104:H4 outbreak strain but are composed of largely disparate genes (Fig.4A). These clusters also have a distinct GC content (36.8-42.1%) from the backbone genome (50.5-50.7%) (Tab.S9), suggesting that they were acquired by horizontal gene exchange. To explore the phylogenetic relationships among the 34 non-O104:H4 ST678 strains and a set of O104:H4 strains, their shared single nucleotide polymorphisms (SNPs) were analyzed using the O104:H4 outbreak strain FWSEC0009 as a reference. A positive correlation (R=0.81; R^2=0,66) was found between isolation time and genetic divergence (Fig.S3) which supports that mutations have accumulated in a clock-like fashion without notable outliers. Fig.4B shows that the AAF/III-positive strains (upper part),

which include nine non-O104:H4 strains of four different serotypes, are clearly separated from the AAF/I positive strains (lower part, referred to as clade I). Clade I is comprised of O104:H4 strains, such as from the 2011 outbreak, 21 O181:H4 strains, three O127:H4 strains, and an OX13:H4 strain (Fig.4B). Strains of clade I are much more closely related to one another than to the AAF/III positive strains. The structure of the phylogeny suggests that clade I strains are derived from an AAF/III-positive precursor. Within clade I, two subclades, Ia and Ib, contain non-O104:H4 strains. Clade Ia is composed of O104:H4 non-outbreak strains isolated from 2015-2021 in the United Kingdom and in Kenya and the 2018 OX13:H4 strain that was associated with travel to Ethiopia. Clade Ib contains the 21 O181:H4 strains and three O127:H4 strains isolated from 2011 to 2021 from diverse continents, but in contrast to the other branches in clade I, no O104:H4 strains belong to this subclade. In the phylogeny, the branch containing clade Ib arises from an older AAF/I positive O104:H4 dominated branch, suggesting these serotypes are derived from an O104:H4 precursor (Fig.4B). Similarly, the phylogeny suggests that the AAF/I positive O127:H4 strains of clade Ib arose from an O181:H4 precursor. Together these observations suggest that O-antigen gene cluster exchange has occurred repeatedly within ST678 strains. This is illustrated by the appearance of O127:H4 strains at multiple locations in the tree and suggests that O127 O-antigen gene donor strains share a niche with ST678 recipient strains.

Discussion

The *E. coli* O104:H4 outbreak in Germany in the early summer of 2011 was a public health emergency; however, this serotype was rarely isolated as a cause of HUS after the epidemic subsided. Nevertheless, our findings suggest that the unusual set of virulence factors that characterized this Shiga toxin 2-producing EAEC strain remains a threat to human health. Serotype conversion has cloaked this highly virulent genotype with several O-antigens, including O181, O127, and OX13, which are present both in *stx* positive and *stx* negative disease-linked isolates closely related to the O104:H4 outbreak strain. We found non-O104:H4 ST678 strains from a variety of countries in Europe, Africa, and the Americas and several of

the cases were associated with travel to Africa and Asia, suggesting these virulence-associated strains are globally distributed. Like the O104:H4 outbreak strain, the AAF/I positive ST678 strains had similar chromosomes, similar pAA-linked virulence genes, as well as virulence factors including the SPATEs, SigA, and Pic [5, 7, 8, 11]. The most salient difference of the chromosomes of these strains with that of the outbreak strain were their respective O-antigen gene clusters. Thus, horizontal exchange of these clusters appears to have been a critical step in the evolution of these new pathogenic serotypes, some of which were linked to HUS or bloody diarrhea.

The prime example uncovered here is the derivation of O181:H4 pathogens from an O104:H4 precursor via O-antigen gene exchange. Among the 21 O181:H4 strains, six harbored a *stx2a*-encoding prophage, including HUS-linked strain 17-07187. In four of these strains, the *stx2a* prophage was very similar to the *stx2a* prophage of the 2011 outbreak strain, suggesting that the O-antigen exchange was a more recent event in their evolution compared to the acquisition of the Shiga toxin-encoding prophage (Fig 3A). The absence of *stx* prophages in 15 of the O181:H4 isolates may be due to a lack of *stx* prophage acquisition or have resulted from loss of their *stx* prophages, which is well documented in STEC in other serotypes [34]. Thus, in addition to O-antigen exchange, on-going horizontal transmission of mobile genetic elements, such as *stx* phages, has contributed to diversification of diarrheagenic ST678.

O-antigen gene clusters of Gram-negative bacteria are hot spots for diversifying selection and recombination events [15, 19]. Serotype conversion by lateral exchange of OAGCs has played an important role in the evolution of enteric pathogens. For example, *Vibrio cholerae* serogroup O139, which arose by exchange of O1 and O139 OAGCs, transiently replaced the dominant O1 group as the cause of cholera in 1992-93 [13, 35]. Also, an interspecies exchange was described for O8 and O9 O-antigens of *E. coli*, which are identical to O5 and O3 *of Klebsiella pneumoniae*, respectively [36]. Our discovery of seven distinct O-groups that share the flagellar H4 antigen and a highly similar virulence-linked genetic makeup provides a compelling

example of the role of O-antigen exchange in the diversification of diarrheagenic *E. coli*. In addition, STEC O157:H7 is thought to have arisen from an enteropathogenic *E. coli* (EPEC)-like O55:H7 strain which initially acquired *stx* via phage transduction and subsequently acquired the O157 O group by exchange of the O55 with the O157 OAGC [37]. Also, an O182-O156 switch is thought to have occurred in STEC strains persistently infecting cattle [18]. Consequently, different OAGs may be found in highly related genotypes [38, 39].

We can only speculate about the conditions driving OAGC exchange among ST678 strains. Shiga toxin producing O104:H4 EAEC strains, such as the 2011 outbreak strain, are considered human-restricted pathogens and have not been isolated from animals, such as cattle [40]. It is possible that the OAGC exchange occurred in a human host, where the human intestinal microbiome may contain *E. coli* strains of O groups, such as O181 and O127, that could have donated their OAGC to an O104:H4 recipient by some means of horizontal exchange. Also, epidemiological investigations suggest that fenugreek sprouts were the food source that initiated the STEC O104:H4 epidemic; therefore plants colonized by microbiota may be another possible site for OAGC exchange [4].

In conclusion, our study highlights how lateral exchange of O-antigen gene clusters can lead to the rapid diversification of a globally important pathogen. Surveillance to uncover how highly virulent strains may reemerge and spread in new O-antigen outfits is warranted. Further, an important clinical implication of these findings is that serotype identification cannot be used as a simple proxy for strain virulence and needs to be complemented by comprehensive virulence gene analysis.

358 **Footnotes:**

359 **Funding**

360 The study sponsors had no role in study design; collection, analysis, and interpretation of data;
361 writing of the manuscript; and in the decision to submit it for publication.

362

363 **Acknowledgements**

364 We thank Ute Strutz, Thomas Garn, Tobias Größl, and Karsten Großhennig for excellent
365 technical assistance (Robert Koch Institute, Wernigerode, Germany). We also acknowledge
366 the MF1 bioinformatics and MF2 genome sequencing unit of the Robert Koch Institute (Berlin)
367 for the support of bioinformatic analysis and Illumina MiSeq and MinION sequencing. We thank
368 members of the Waldor and Flieger groups for helpful comments and discussions during the
369 study. AFl received funding from the German Ministry of Health (IGS-ZOO grant number
370 2518FSB706) and AFl and AFr from the intensified molecular surveillance initiative of the
371 Robert Koch- Institute. IWC received funding from the National Institute of Diabetes and
372 Digestive and Kidney Diseases of the National Institutes of Health (Award Number 5
373 T32DK007477-37).

374

375 **Conflict of interests**

376 We declare no conflict of interests.

377

378 **Data availability**

379 All data generated or analyzed during this study are available in this Article and the
380 appendices. The sequences were uploaded to NCBI project: PRJNA833419.

381

382 **Corresponding author contact information:**

383 Antje Flieger
384 email: fliegera@rki.de
385 Tel.: +49-30-18754-2522

References

1. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev **2013**; 26:822-80.

2. Caprioli A, Scavia G, Morabito S. Public Health Microbiology of Shiga Toxin-Producing *Escherichia coli*. Microbiol Spectr **2014**; 2:EHEC-0014-2013.

3. Mellmann A, Bielaszewska M, Kock R, et al. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic *Escherichia coli*. Emerg Infect Dis **2008**; 14:1287-90.

4. Werber D, Krause G, Frank C, et al. Outbreaks of virulent diarrheagenic *Escherichia coli*--are we in control? BMC Med **2012**; 10:11.

5. Bielaszewska M, Mellmann A, Zhang W, et al. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lancet Infect Dis **2011**; 11:671-6.

6. Frank C, Werber D, Cramer JP, et al. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. The New England journal of medicine **2011**; 365:1771-80.

7. Rasko DA, Webster DR, Sahl JW, et al. Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany. The New England journal of medicine **2011**; 365:709-17.

8. Rohde H, Qin J, Cui Y, et al. Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. The New England journal of medicine **2011**; 365:718-24.

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

10. Boisen N, Melton-Celsa AR, Scheutz F, O'Brien AD, Nataro JP. Shiga toxin 2a and Enteroaggregative *Escherichia coli*--a deadly combination. Gut Microbes **2015**; 6:272-8.

412 11. Munera D, Ritchie JM, Hatzios SK, et al. Autotransporters but not pAA are critical for
413 rabbit colonization by Shiga toxin-producing *Escherichia coli* O104:H4. Nat Commun **2014**;
414 5:3080.
415 12. Coipan CE, Friesema IH, van den Beld MJC, et al. Sporadic Occurrence of
416 Enteroaggregative Shiga Toxin-Producing *Escherichia coli* O104:H4 Similar to 2011
417 Outbreak Strain. Emerg Infect Dis **2022**; 28:1890-4.
418 13. Lerouge I, Vanderleyden J. O-antigen structural variation: mechanisms and possible
419 roles in animal/plant-microbe interactions. FEMS Microbiol Rev **2002**; 26:17-47.
420 14. Orskov F, Orskov I. *Escherichia coli* serotyping and disease in man and animals. Can J
421 Microbiol **1992**; 38:699-704.
422 15. Iguchi A, Iyoda S, Kikuchi T, et al. A complete view of the genetic diversity of the
423 *Escherichia coli* O-antigen biosynthesis gene cluster. DNA Res **2015**; 22:101-7.
424 16. Lang C, Hiller M, Konrad R, Fruth A, Flieger A. Whole-Genome-Based Public Health
425 Surveillance of Less Common Shiga Toxin-Producing *Escherichia coli* Serovars and
426 Untypeable Strains Identifies Four Novel O Genotypes. J Clin Microbiol **2019**; 57.
427 17. Samuel G, Hogbin JP, Wang L, Reeves PR. Relationships of the *Escherichia coli* O157,
428 O111, and O55 O-antigen gene clusters with those of *Salmonella enterica* and *Citrobacter*
429 *freundii,* which express identical O antigens. J Bacteriol **2004**; 186:6536-43.
430 18. Geue L, Menge C, Eichhorn I, et al. Evidence for Contemporary Switching of the O-
431 Antigen Gene Cluster between Shiga Toxin-Producing *Escherichia coli* Strains Colonizing
432 Cattle. Front Microbiol **2017**; 8:424.
433 19. Milkman R, Jaeger E, McBride RD. Molecular evolution of the *Escherichia coli*
434 chromosome. VI. Two regions of high effective recombination. Genetics **2003**; 163:475-83.
435 20. Prager R, Lang C, Aurass P, Fruth A, Tietze E, Flieger A. Two novel EHEC/EAEC hybrid
436 strains isolated from human infections. PLoS One **2014**; 9:e95379.
437 21. Walker BJ, Abeel T, Shea T, et al. Pilon: an integrated tool for comprehensive microbial
438 variant detection and genome assembly improvement. PLoS One **2014**; 9:e112963.

22. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One **2010**; 5:e11147.

23. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol **2013**; 30:772-80.

24. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Agama Study G, Achtman M. The EnteroBase user's guide, with case studies on Salmonella transmissions, *Yersinia pestis* phylogeny, and Escherichia core genomic diversity. Genome Res **2020**; 30:138-52.

25. Arndt D, Grant JR, Marcu A, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Res **2016**; 44:W16-21.

26. Aziz RK, Bartels D, Best AA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics **2008**; 9:75.

27. Lang C, Fruth A, Holland G, et al. Novel type of pilus associated with a Shiga-toxigenic *E. coli* hybrid pathovar conveys aggregative adherence and bacterial virulence. Emerg Microbes Infect **2018**; 7:203.

28. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res **2015**; 43:e15.

29. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol **2010**; 59:307-21.

30. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol **2016**; 2:vew007.

31. Kunne C, Billion A, Mshana SE, et al. Complete sequences of plasmids from the hemolytic-uremic syndrome-associated *Escherichia coli* strain HUSEC41. J Bacteriol **2012**; 194:532-3.

32. Schiller P, Knodler M, Berger P, et al. The Superior Adherence Phenotype of *E. coli* O104:H4 is Directly Mediated by the Aggregative Adherence Fimbriae Type I. Virulence **2021**; 12:346-59.

33. Gaillot O, Clement C, Simonet M, Philippon A. Novel transferable beta-lactam resistance with cephalosporinase characteristics in *Salmonella enteritidis*. J Antimicrob Chemother **1997**; 39:85-7.

34. Bielaszewska M, Prager R, Kock R, et al. Shiga toxin gene loss and transfer in vitro and in vivo during enterohemorrhagic *Escherichia coli* O26 infection in humans. Applied and environmental microbiology **2007**; 73:3144-50.

35. Faruque SM, Sack DA, Sack RB, Colwell RR, Takeda Y, Nair GB. Emergence and evolution of *Vibrio cholerae* O139. Proc Natl Acad Sci U S A **2003**; 100:1304-9.

36. Sugiyama T, Kido N, Kato Y, Koide N, Yoshida T, Yokochi T. Evolutionary relationship among rfb gene clusters synthesizing mannose homopolymer as O-specific polysaccharides in *Escherichia coli* and Klebsiella. Gene **1997**; 198:111-3.

37. Feng P, Lempel KA, Karch H, Whittam TS. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. J Infect Dis **1998**; 177:1750-3.

38. Ingle DJ, Valcanis M, Kuzevski A, et al. In silico serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. Microb Genom **2016**; 2:e000064.

39. Ochman H, Selander RK. Evidence for clonal population structure in *Escherichia coli*. Proc Natl Acad Sci U S A **1984**; 81:198-201.

40. Wieler LH, Semmler T, Eichhorn I, et al. No evidence of the Shiga toxin-producing *E. coli* O104:H4 outbreak strain or enteroaggregative *E. coli* (EAEC) found in cattle faeces in northern Germany, the hotspot of the 2011 HUS outbreak area. Gut Pathog **2011**; 3:17.

Figures

A

		EHEC virulence genes											EAEC virulence genes										
		chromosome						plasmid					chromosome				pAA						
serotype	strain	*stx*	*eae*	*lpf*$_{O26}$	*lpf*$_{O113}$	*terA*	*irp2*	*fyuA*	*ehxA*	*subAB*	*espP*	*iha*	*saa*	*set1*	*pic*	*sigA*	*aaiC*	*aatA,B,C,D,P*	AAFI	*aap*	*aggR*	EAST1	*iucA*	*sepA*
O104:H4	outbreak 2011	2a		+	+	+	+	+				+		+	+	+		+	+	+	+		+	+
O181:H4	17-07187	2a		+	+	+	+	+				+		+	+	+		+	+	+	+		+	+

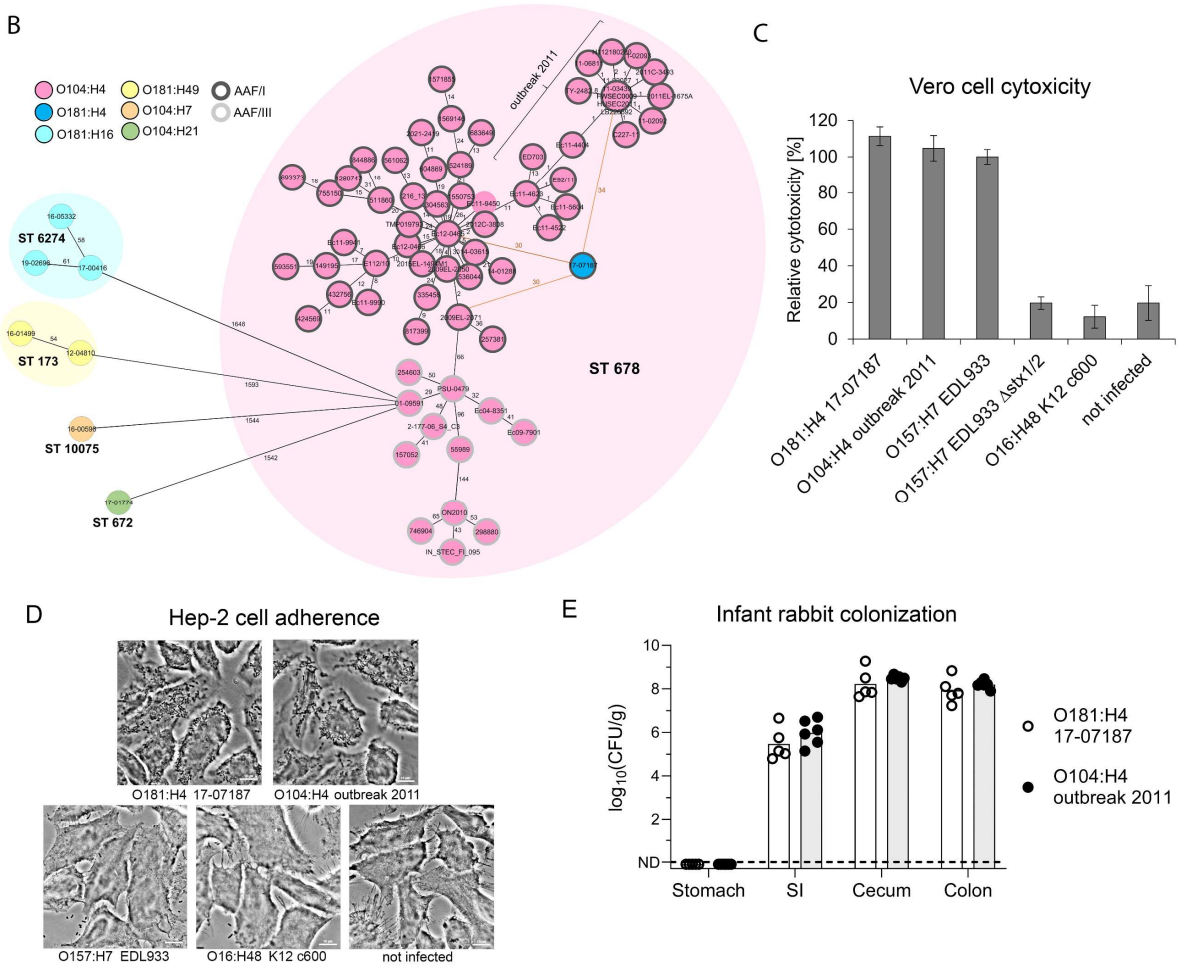


Figure 1: STEC O181:H4/17-07187 shares close phylogenetic relationship and similar virulence traits with O104:H4 outbreak strain. (A) Overlapping virulence gene profiles of the O104:H4 outbreak strain (FWSEC0009) and the O181:H4 strain 17-07187. (B) Minimal spanning tree of O181:H4 17-07187 and selected O104:H4 strains (representing phylogenetic diversity of O104:H4 strains with respect to isolation time from 1998 to 2022 and location) based on cgMLST involving 2513 alleles confirms their close phylogenetic relationship. O181:H4 17-07187 and O104:H4 strains 2009EL-2071, Ec12-0465, and outbreak strains from 2011 (including strains FWSEC0009, C227-11, and 11-02027) differ by only 30-34 alleles. Serotypes and AAF type are indicated in the key. (C) Cytotoxicity of O181:H4 17-07187

towards Vero cells and (D) adherence pattern to HEp-2 cells are comparable to the O104:H4 outbreak strain 11-02027. Results are representative of at least two additional experiments. EHEC O157:H7 EDL933 producing Stx1 and Stx2 served as reference in the cytotoxicity assay and was set to 100%. The cytotoxicity results represent the means and standard deviations of triplicate samples (n=3) and are representative of at least two additional experiments. E) Intestinal colonization in infant rabbits inoculated with O104:H4 outbreak strain C227-11 (n=6) or O181:H4 strain 17-07187 (n=5). CFU/g (colony forming units per gram of tissue) denotes concentration of bacteria recovered three days post inoculation from homogenated intestinal tissues. Data points represent individual rabbits from two independent litters split between the two strains. Bars show the geometric mean, SI is small intestine, ND is not detected.

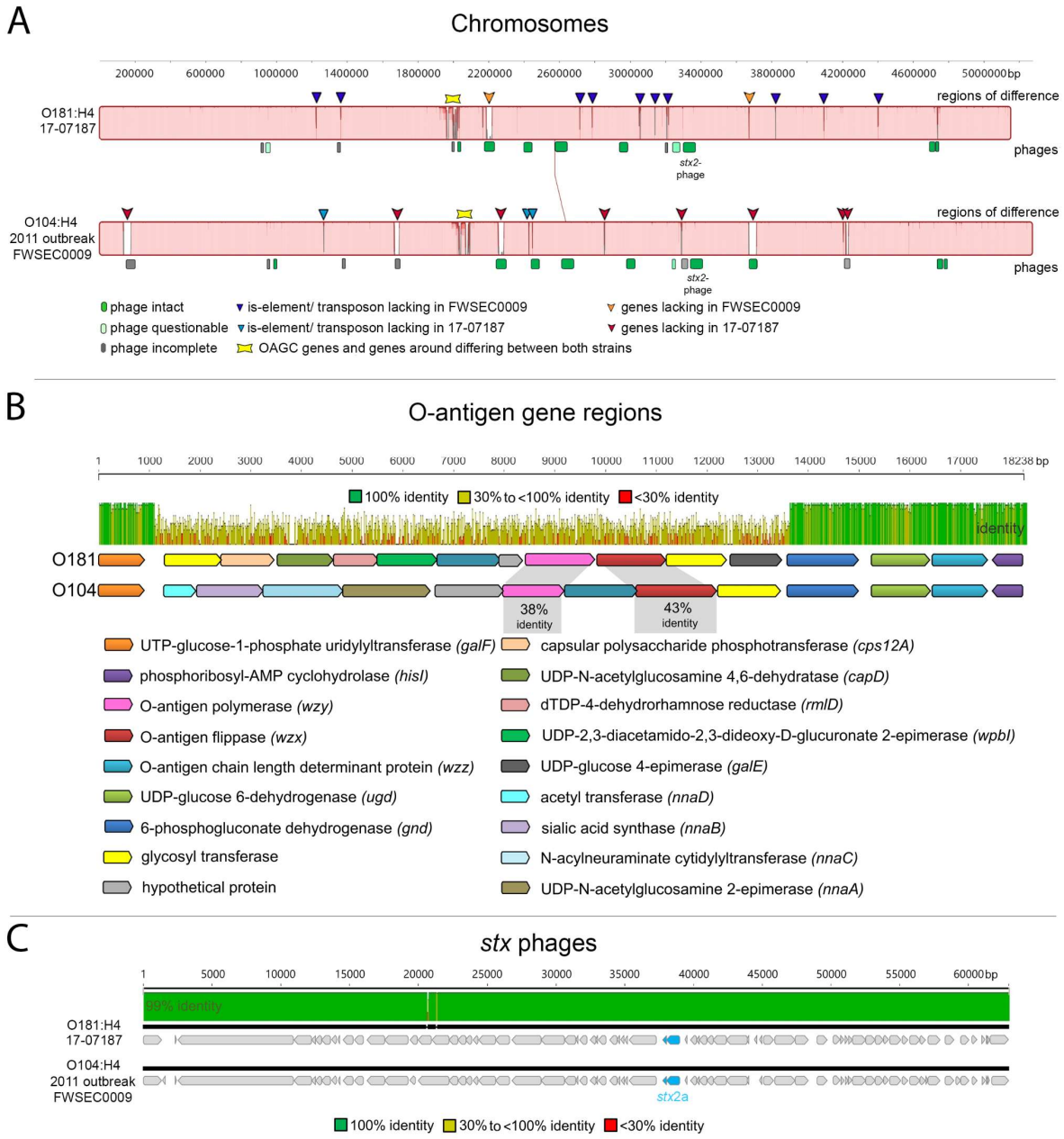


Figure 2: O181:H4 17-07187 and O104:H4 outbreak strain genomes mainly differ in OAGCs and mobile genetic elements. (A) Whole chromosome MAUVE alignment of O181:H4 17-07187 and the O104:H4 outbreak strain FWSEC0009 highlighting mobile genetic elements and differences in phage regions, IS-/ transposon elements, and OAGC regions. (B) OAGCs of O181:H4 17-07187 and the O104:H4 FWSEC0009 outbreak strain are flanked by homologous upstream (*galF*) and downstream (*gnd* to *hisI*) regions. MAFFT alignment shows that the regions in between *galF* and *gnd* are very different. (C) MAFFT alignment of the *stx2a*-encoding phages in O181:H4 17-07187 and the outbreak strain O104:H4 outbreak strain FWSEC0009 shows that they are very similar (99% nucleotide identity).

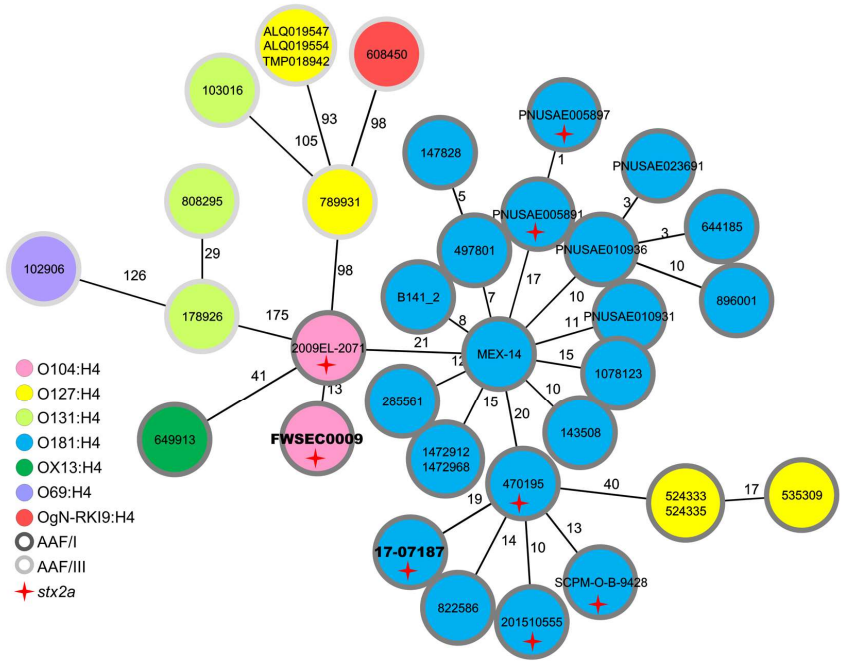
A

strain	serotype	allele distance to strain FWSEC0009	country of isolation	travel association	year of isolation	FWSEC0009 O104:H4			
						pI1	pI2 pAA	pI3	stx2a-phage
FWSEC0009	O104:H4	0	Germany		2011	100	100	100	100
MEX-14	O181:H4	26	Mexico		2011	5.7	100	100	
497801	O181:H4	27	United Kingdom	Morocco	2013	1.7	100	100	
147828	O181:H4	28	United Kingdom	Morocco	2015	90.7	100	100	
B141_2	O181:H4	28	Ecuador		2015	2.9	95.5	100	
143508	O181:H4	30	United Kingdom		2015	17.5	100	100	
PNUSAE010936	O181:H4	30	United States		2017	1.1	100	100	
644185	O181:H4	31	United Kingdom	Mexico	2018	8.4	100	100	
PNUSAE023691	O181:H4	31	United States		2019	1.1	100	99.9	
285561	O181:H4	32	United Kingdom	Egypt	2016	32.1	100	100	
470195	O181:H4	32	United Kingdom	India	2018	16.1	100	20.3	100
PNUSAE010931	O181:H4	32	United States		2017	1.6	100	100	
201510555	O181:H4	34	France		2015	9	100	0	100
1078123	O181:H4	35	United Kingdom		2021	1.9	97.1	100	
1472912	O181:H4	35	United Kingdom		2021	4.7	100	100	
17-07187	O181:H4	35	Germany		2017	37.3	100	0	100
1472968	O181:H4	36	United Kingdom		2021	4.3	100	100	
SCPM-O-B-9428	O181:H4	36	Russia		2018	36.8	100	87.8	98.9
822586	O181:H4	38	United Kingdom	Morocco	2019	3.3	99.7	22.1	
896001	O181:H4	38	United Kingdom		2020	32.7	99.6	100	
PNUSAE005891	O181:H4	38	United States		2016	35.6	100	100	58.1
PNUSAE005897	O181:H4	39	United States		2017	35	99.9	100	58.4
649913	OX13:H4	46	United Kingdom	Ethiopia	2018	5.4	99.4	100	
524333	O127:H4	53	United Kingdom	India	2018	3.6	100	18.9	
524335	O127:H4	53	United Kingdom	India	2018	3	100	53.6	
535309	O127:H4	54	United Kingdom		2018	14.8	100	19.4	
789931	O127:H4	106	United Kingdom	Kenya	2019	17.5	70.8	13	
608450	RKI9:H4	112	United Kingdom		2018	7.3	62.1	30.5	
ALQ019547	O127:H4	113	Kenya		2015	16.2	69.6	0	
ALQ019554	O127:H4	113	Kenya		2015	16.4	69.3	0	
TMP018942	O127:H4	113	Kenya		2015	16.1	69.5	0	
103016	O131:H4	121	Gambia		2015	3.5	60.4	5.9	
178926	O131:H4	182	United Kingdom	India	2019	31.2	62.9	13	
102906	O69:H4	187	Gambia		2009	15	70.2	5.7	
808295	O131:H4	189	United Kingdom	India	2019	4.8	60.9	52.6	

AAF/I
AAF/III

0 — 100% coverage of reference

B



524 **Figure 3: Recent global isolates of serotype O181:H4 and five other O groups belong to**
525 **ST678.** (A) Summary of serotypes, isolation dates, along with relatedness of genomes
526 (cgMLST-based allelic distances) and mobile genetic elements (plasmids and *stx* phage) of
527 the 34 non-O104:H4 ST678 (ST12598, ST12610) clinical strains (found on Enterobase)
528 compared to the O104:H4 outbreak strain FSWEC0009. (B) Serotypes of the non-O104:H4
529 ST678 strains and the O104:H4 strains 2009EL-207 and the outbreak strain FSWEC0009 in a
530 minimal spanning tree based on cgMLST. Although, phylogenetically very close to ST678
531 strains, O181:H4 strains 1472912 and 1472968 in fact belong to ST12610 due to a point
532 mutation in *icd* and OgN-RKI9:H4 strain 608450 to ST12598 due to a point mutation in *recA*.
533

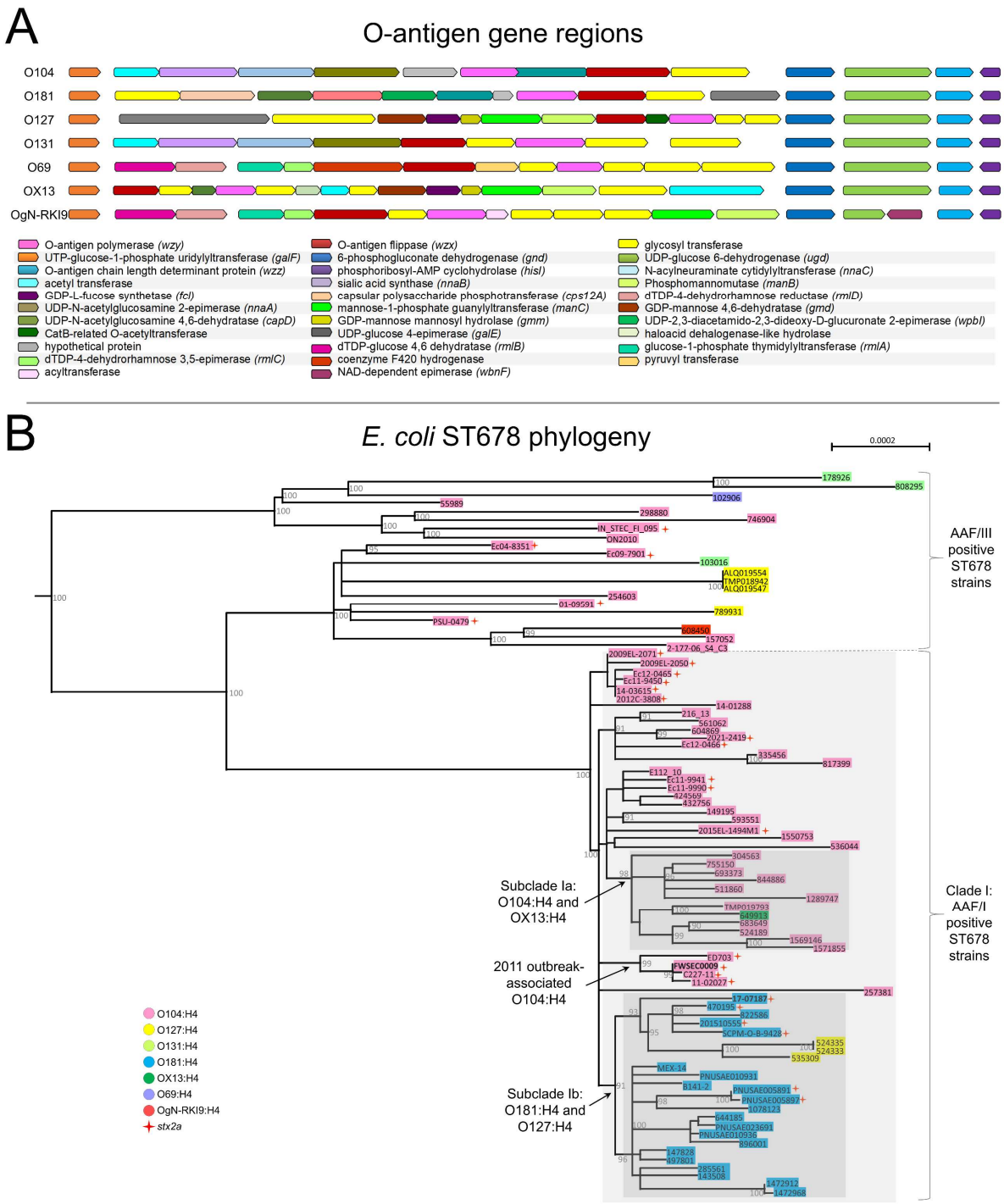


Figure 4: O-antigen gene clusters and phylogeny within *E. coli* ST678. (A) O-antigen gene clusters in ST678 strains differ in their gene contents and organization, yet they are flanked by homologous regions starting upstream by *galF* and downstream by *gnd* to *hisI*, respectively. (B) Maximum-likelihood phylogenetic tree based on a recombination-corrected alignment of genome-wide polymorphic sites. Strains selected for representation in Fig.1B and Fig.3 were combined and O104:H4 2011 outbreak strain FWSEC0009 served as reference. The tree was

541 rooted with an outgroup consisting of *E. coli* strains K12 c600, EcO42, EDL933, O157:H7 strain
542 Sakai, O26:H11 strain 11368, and O103:H2 strain 12009. Bootstrap values >90% are
543 indicated. Serotypes are as indicated in the key and presence of *stx* is indicated by a red star.
544 Scale bar refers to a phylogenetic distance of 0.0002 nucleotide substitutions per site.
545

546 **Supplementary Data:**

547 **Figure S1:** The O181:H4 STEC strain 17-07187 and the O104:H4 outbreak strain harbor a
548 partially distinct set of plasmids with overlapping pAA. MAUVE alignments highlight that (A)
549 STEC O181:H4 17-07187 plasmid 1 is most similar to pHUSEC41-1 of STEC O104:H4
550 HUSEC41 from 2001 but lacks antibiotic resistance genes, (B) O181:H4 17-07187 plasmid 2
551 is very similar to pAA of the O104:H4 outbreak strain and (C) O181:H4 17-07187 plasmid 3 is
552 similar to DHA-1 plasmids of enterobacteria but lacks antibiotic resistance genes.

553

554 **Figure S2:** Minimal spanning tree based on cgMLST of the 34 non-O104:H4 ST678 strains
555 and the O104:H4 strains 2009EL-2071 and FWSEC0009 colored by year of isolation (A),
556 country of isolation (B), and travel association (C).

557

558 **Figure S3:** TempEst revealed a positive correlation (R=0.81; R^2=0,66) between isolation time
559 and genetic divergence represented by the root to tip regression analysis using RAxML tree
560 generated by Gubbins based on a recombination-corrected alignment of genome wide
561 polymorphic sites with O104:H4 2011 outbreak strain FWSEC0009 as reference.

562

563 **Table S1:** Overview of all strains used in the study and information on their serotypes,
564 virulence genes, country of isolation, travel association, clinical data, and genome
565 accessions. (provided as excel file)

566

567 **Table S2:** Genome characteristics of O181:H4 17-07187 and the 2011 O104:H4 outbreak
568 strain FWSEC0009. (provided as word file)

569

570 **Table S3:** Comparison of phage-regions detected in STEC O181:H4 17-07187 and in STEC
571 O104:H4 strains FWSEC00009 from the outbreak 2011 and 2009EL-2071 from 2009.
572 (provided as excel file)

573

574 **Table S4:** Comparison of the *stx2a* phages detected in STEC O181:H4 17-07187 and in the
575 O104:H4 outbreak strain FWSEC00009. (provided as excel file)
576
577 **Table S5:** Comparison of O181:H4 17-07187 plasmid 1 to pHUSEC41-1 of STEC O104:H4
578 HUSEC41 from 2001. (provided as excel file)
579
580 **Table S6:** Comparison of O181:H4 17-07187 plasmid 2 to pAA of the O104:H4 outbreak
581 strain. (provided as excel file)
582
583 **Table S7:** Comparison of O181:H4 17-07187 plasmid 3 to pDHA1 of enterobacteria.
584 (provided as excel file)
585 **Table S8:** Overview of the virulence gene profiles of the 34 non-O104:H4 ST678 strains and
586 the 2011 O104:H4 outbreak strain FWSEC0009. (provided as excel file)
587
588 **Table S9:** GC contents of *E. coli* chromosomes and their respective O-antigen gene clusters.
589 (provided as excel file)