Whole genome-based public health surveillance of less common STEC serovars and untypable strains identifies four novel O genotypes

Christina Lang,[a] Miriam Hiller,[a] Regina Konrad,[b] Angelika Fruth,[a] Antje Flieger,[a]#

[a]Division of Enteropathogenic Bacteria and *Legionella*, National Reference Centre for Salmonella and other Enteric Bacterial Pathogens, Robert Koch-Institute, Wernigerode, Germany

[b]Department of Infectiology, Bavarian Health and Food Safety Authority, Oberschleissheim, Germany

Running Head: Whole genome-based surveillance of STEC

#Address correspondence to Antje Flieger, mail to fliegera@rki.de

Abstract

Shiga toxin-producing *E. coli* (STEC) and the subgroup of enterohemorrhagic *E. coli* cause intestinal infections with symptoms ranging from watery diarrhea to hemolytic uremic syndrome (HUS). A key tool for epidemiological differentiation of STEC is serotyping. The serotype in combination with the main virulence determinants gives an important insight into the virulence potential of a strain. However, a larger fraction of STEC found in human disease, including strains causing HUS, belongs to less frequently detected STEC serovars or their O/H antigens are unknown or even untypable. Recent implementation of whole genome sequence (WGS) analysis in principal allows the deduction of serovar and virulence gene information. Therefore we here compared classical serovar and PCR-based virulence marker detection with WGS-based methods for 232 STEC strains focusing on less frequently detected STEC serovars and non-typable strains. We found that WGS-based extraction showed a very high degree of overlap with the more classical methods. Specifically, concordance was 97% for O antigens (OAGs) and 99% for H antigens (HAGs) of typable strains and > 99% concerning *stx1/2* or *eaeA* for all strains. 98% of non-typable OAGs and 100% of non-typable HAGs were defined by WGS analysis. In addition, the novel methods enabled a more complete analysis of strains causing severe clinical symptoms and the description of four novel STEC OAG loci. In conclusion, WGS is a promising tool for gaining serovar and virulence gene information especially from a public health perspective.

Introduction

Shiga toxin producing *E. coli* (STEC), including the subgroup of enterohemorrhagic *E. coli* (EHEC), cause intestinal infections ranging from sporadic disease to large outbreaks worldwide (1). In Germany, about 2.000 cases of STEC associated diarrhea/bloody diarrhea and about 70 cases of severe hemolytic uremic syndrome (HUS) are annually reported since 2015. Of note, the number has been steadily increasing during the last years, a tendency which is observed throughout Europe (2, 3).

The most important virulence determinant of STEC/EHEC is Shigatoxin (Stx). Stx is responsible for severe pathologies like HUS and is divided in two different types (4). Stx1 has three subtypes, namely a, c, and d whereas the more toxic Stx2 is represented by eight different subtypes, designated a-h. Subtyped *stx* genes are important epidemiological markers. Additionally, disease outcome has been attributed to specific Stx types. HUS-associated strains (e.g. strains from the HUSEC collection) often carry genes coding for *stx2a*, *stx2d*, *stx2c* and *stx1a* alone or in combination with other types (5, 6). *Stx* gene subtyping and detection of other virulence determinants therefore may permit risk profiling of such pathogens (5). Further virulence factors/genes are present in so-called classical STEC but are absent in a variety of other often less characterized STEC. Examples are a type III protein secretion system coded on a pathogenicity island, namely the locus of enterocyte effacement (LEE), and the enterohemolysin HlyA, encoded by the gene *ehxA*. LEE induces intimate attachment of the bacteria to the intestinal epithelia and HlyA is a pore-forming toxin (4, 7).

A key tool for differentiation of STEC is serotyping. Classical STEC serotyping has been performed for more than 50 years routinely and assignment of a serovar is important for surveillance and cluster detection. Typically used for sub differentiation are the O and H surface antigens, specifically lipopolysaccharide and flagellin of the bacteria, respectively (8). So far 182 O serogroups (O1-O188 except O31, O47, O67, O72, O93 and O94) and 53 associated H forms (H1-56 except H13, H22, H50) are described (9). Interestingly, only

75 strains of a few O antigens (OAGs) often combined with specific H antigens (HAGs) cause
76 more than 50% of STEC infections, such as O91, O103, O146, O157, O26, O113, O128, O76,
77 and O145 (1, 9). Of these more frequently found serogroups, O157 is principally associated
78 with development of severe disease (1, 10).
79 However, it is important to note that a larger fraction (about 30%) of the HUSEC collection
80 strains does not belong to strains of frequently found STEC OAGs (6). In addition, the 2011
81 HUS outbreak in Germany caused by an STEC of the rare serovar O104:H4 illustrates the
82 high potential of these more unusual strains to cause severe disease. Worldwide it was the
83 largest outbreak of bloody diarrhea / HUS so far and involved 53 deaths, 833 HUS cases, and
84 about 3,000 cases of gastroenteritis (11-13).
85 Implementation of whole genome sequencing (WGS) techniques into public health
86 microbiology now permits genome-based typing for pathogen surveillance and cluster
87 analysis. The new method also enables deduction of serovar information (14-19). This is
88 especially important for previously non-serotypable strains, namely for rough, non-motile and
89 Ont/Hnt strains. Joensen et al. created a FASTA database of specific O-antigen processing
90 systems and flagellin genes for O and H typing, respectively. This resource is a component of
91 the publicly available web tool hosted by the center of genomic epidemiology (CGE, DTU;
92 Denmark) (http://www.genomicepidemiology.org). They analyzed ~500-600 *E. coli* WGS
93 data with serotype information with the SerotypeFinder CGE tool. In 560 of 569 cases and
94 504 of 508 cases, respectively, the O and H types were predicted consistently with classical
95 serotyping. The authors therefore concluded that *E. coli* serotyping can be done solely from
96 WGS data and provides a superior alternative to conventional serotyping (16). Further,
97 Chattaway et al. evaluated the use of WGS for routine public health surveillance of non-O157
98 STEC by comparing this approach to phenotypic serotyping. Of the 102 isolates, 98 had
99 concordant results. The most common non-O157 STEC serogroups detected were O146 and

O26. 38 isolates could not be phenotypically serotyped. Only one of these was not successfully serotyped using the WGS data (19).

In the here presented study, we compared classical serovar analysis with WGS-based genoserotyping in the STEC routine analysis setting of the German National Reference Center for *Salmonella* and other Enteric Bacterial Pathogens (NRC). Whereas previous studies mostly concentrated on strains with more common OAGs, we focused on less frequently detected STEC serovars and non-typable strains. In addition, we compared PCR-based virulence gene analysis with WGS-based data. As a conclusion, we found a very high degree of overlap with classical or PCR-based methods. In addition the novel methods enabled further analysis of strains causing severe clinical symptoms and the description of four novel STEC OAG loci.

Material & Methods

Strains

The strains used in the study are listed in Tab. S1. All strains were human isolates, except for seven food isolates. Strains were grown on nutrient agar (Oxoid GmbH, Germany) or in tryptic soy broth (TSB) (BD-BBL, Germany), if not stated otherwise. Testing of enterohemolysin production was performed on enterohemolysin agar (Sifin GmbH, Germany).

E. coli serotyping

Serotyping was performed using antisera against *E. coli* O-antigens 1–188 and *E. coli* H-antigens 1–56 by use of a microtitre agglutination method as described elsewhere (20).

Antibiotic susceptibility testing

All of the strains were tested for antibiotic susceptibility according to EUCAST recommendation for *E.coli* by broth microdilution assay against 16 antibiotics: (http://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Disk_test_documents/2019_manuals/Reading_guide_BMD_v_1.0_2019.pdf).

PCR-based virulence gene analysis

All *stx* genotypes, presence of *eae* (encoding adhesin intimin), and *ehxA* gene were first determined using polymerase chain reaction (PCR) (5, 21).

Whole genome sequencing (WGS)

Whole-genome sequencing was accomplished using short-read paired-end sequencing provided by MiSeq (2 × 300 bp) and HiSeq 1500 (2 × 250bp) instruments (Illumina, San Diego, CA). For this, DNA from *E. coli* strains was isolated with the Qiagen DNeasy Blood

& Tissue Kit (Qiagen) according to manufacturer's instructions and 1 ng of the extracted DNA was used to generate libraries by using the Nextera XT DNA Library according to the manufacturer's instructions (Illumina, San Diego, CA). Requirements for the sequence raw data were: sequence yield >600 000 reads/sample, mean sequence quality score (Phred score) >25 and genome coverage >30fold. On average, the sequence yield was about 2,6 million reads/ sample and the genome coverage was 120fold. The raw FASTQ sequences were uploaded to the European Nucleotide Archive (ENA) in study Acc. No PRJEB32361.

Bioinformatics analyses

Raw reads were subjected to quality control and trimming via the QCumber pipeline (version 2.1.1; https://gitlab.com/RKIBioinformaticsPipelines/QCumber) utilizing FastQC (version: 0.11.5; https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), Trimmomatic (version: 0.36; (22), and Kraken (version: 0.10.6; (23)). Trimmomatic was used with default parameters (Phred score 33). On average, 80% of reads remained after trimming. To identify the serotype and virulence genes trimmed reads were mapped by means of the standard Geneious assembler (settings: medium-sensitivity, none-iterations; Geneious (version: R10.0.5, Biomatters Ltd) against the respective reference sequence. Requirements for positive matches were 100% coverage of the reference sequence, >90% identity with the reference sequence and >90% bases in sequence that are of high quality.

Reference sequences for *the wzm*, *wzy*, *wzm*, *wzt* and *fliC* genes for serotype determination and for virulence marker genes were downloaded from Center for Genomic Epidemiology (CGE, DTU; Denmark; https://cge.cbs.dtu.dk/services/data.php; SerotypeFinder, VirulenceFinder). Further reference sequences for serotyping were obtained from NCBI (Tab. S2).

Ridom Seqsphere+ (version:5.1.0, Ridom GmbH, Germany) was used to create neighbor joining tree based on 2513 targets from *E. coli* cgMLST Enterobase with pairwise ignoring

164 missing values. Ridom Seqsphere+ was also used to determine MLST Warwick sequence
165 types.
166 When potentially novel O antigen loci were analyzed, the MiSeq reads were *de novo*
167 assembled with the program A5 (version: 2.1.3) and the contigs were further analysed by
168 means of Geneious. Using all known OAGC as annotation reference (Geneious tool: Annotate
169 from), the partially annotated contigs were extracted and the OAGC region was defined in-
170 between the genes *galF* (UTP-glucose-1-phosphate uridyltransferase) and *hisI* (histidine
171 biosynthesis bifunctional protein), because genes required for the biosynthesis of *E. coli*
172 OAGs are mostly located at this site (24-27). Some ORFs of the new clusters could not be
173 annotated using the known OAGCs as reference. These were then translated into proteins and
174 NCBI pBLAST (standard settings, database: non-redundant protein sequences, algorithm
175 blastp) was performed to search for functional homologues. Homologues of the newly defined
176 OAGC were detected using NCBI nBLAST (standard settings, database: nucleotide
177 collection; optimize for highly similar sequences). The annotated sequences of the new
178 OGACs were uploaded to NCBI (GenBank accession numbers MN172354-MN172357).
179
180

Journal of Clinical Microbiology

Results

STEC isolates analyzed at NRC from 2015 to 2017 and selection of strains for WGS focusing on less frequently found serovars and untypable O antigens.

The NRC receives STEC samples from human disease cases for further subtyping. In 2015 ~ 641, in 2016 ~ 895 and in 2017 ~ 1466 STEC samples (total of 3002) were obtained, which represent about 40-74% of the reported STEC infections per year (2, 9). Of these, about 84% were serotyped by classical microtiter agglutination method, and all were analyzed by PCR for presence of virulence genes, such as *stx1/2* and *eaeA*. 63% of the 2015-2017 STEC isolates belonged to the most frequently identified OAG types including O26, O91, O76, O103, O113, O128, O145, O146, O157. Approximately 15% were Ont or of rough LPS and a further ~20% did not belong to the above mentioned frequently found serovars (Fig. 1A). All isolates harbored *stx*. 41.5% comprised *stx1*, 34.1% *stx2* and 24.4% both *stx1* and *stx2*. 30.5% of the 2015-2017 STEC isolates possessed the *eaeA* gene; specifically 49.0% in combination with *stx1*, 37.4% with *stx2*, or 13.4% with both.

Next, we selected 232 STEC strains from the 2015-2017 isolates by the following criteria: a) less frequently found (less common) OAGs, b) uncommon O:H combinations, and c) Ont/Orough types. For comparison, we also included about 25% strains of more common OAGs. Serovar distribution is shown in Fig. 1 B. Additionally, antibiotic susceptibility testing was performed for epidemiological purpose and analysis revealed that 70% of the selected strains were susceptible and 21% were resistant to more than one of the tested antibiotic (Tab. S1).

WGS-based O antigen determination highly correlates with classical serotyping.

From genome sequence data of the 232 selected strains OAG types were extracted. Here, we mapped the trimmed reads against a set of reference sequences (see Material and Methods

section). By means of classical serotyping, in 67.2% of the strains the OAG was typable and of these 96.8% were confirmed by WGS analysis. Only five strains (3.2%) showed a discordant result.

Specifically, two strains (16-01717 and 16-01865) were classically serotyped as O57, but WGS analysis yielded for the first one OgN1 which was recently assigned as a new OAG (18) and for the second one O2. One strain (16-04148) was determined as O169, but was WGS-typed as O81. The fourth strain (17-05507) was originally serotyped as O109 but WGS analysis revealed O182. The last one (16-04178) defined as O54 was not sufficiently matching with any reference sequence and might belong to a novel OAG cluster (see below).

In 76 (32.8%) of the 232 strains, the OAG was not typable (28.0%) by classical serotyping or was Orough (4.7%). The majority (97.8%) could be classified by WGS analysis and the most frequently found types were: O27 (9 strains), O100 (7 strains), O80 (4 strains), and O153/178 (4 strains) (Tab. S1). About 15% of the non-typable strains belonged to the recently described OgN O antigen clusters (OAGCs) (OgN1, OgN10, OgN12, OgN13, OgN31) (18). Most interestingly, five strains were not matching with known OAG loci and therefore might belong to novel OAGs. Two of the strains harbored a similar OAG locus. In summary, extraction of serotype data correlates very well with classical data and allows classification of so far untypable strains and the identification of novel O genotypes.

WGS-based H antigen determination highly correlates with classical serotyping.

Next, we extracted the H types from WGS data. By classical serotyping in 80% of the strains the HAG was typable and of these 99% were confirmed by WGS analysis. Only two strains showed non-correlating results (16-01506: H19/H14 and 17-05292: H36/H31). By means of classical serotyping, 2.6% HAGs were untypable and 17.9% of the strains were not motile (total of 20.5%). All of these were defined by WGS analysis.

Identification of four novel O antigen gene loci.

As mentioned above, OAGCs of 6 strains were not identified because they did not match to any known OAG loci (Tab. 1). As they might be new OAG loci, *de novo* assembly of the MiSeq reads was performed, resulting contigs were annotated using all known OAGC sequences as reference. (for details see M&M). Some ORFs of the new clusters could not be annotated using the known OAGCs as reference. These were then translated into proteins and pBLAST was performed to search for functional homologues. By this means, it was possible to define new putative O-unit processing genes, specifically *wzx* (gene for O-antigen flippase) and *wzy* (gene for O-antigen polymerase), which are relatively unique for each individual O-type (28). Indeed, four novel OAGCs were defined; two strains carried identical OgN-RKI1 (strains 16-01174 and 16-04846), another two strains carried identical OgN-RKI2 (strains 16-02258 and 17-05936) and the remaining two strains were each assigned to OgN-RKI3 (strain 16-04178) and the other to OgN-RKI4 (strain 17-05676) (Tab. 1). Fig. 2 gives an overview of these new OAGCs.

Comparison of the Wzx and Wzy protein sequences to those from 167 O serogroup strains, 10 OX group reference strains and 15 OgN strains indicated that the sequences of the new OAGCs were unique compared to known OAGCs (29) (Fig. 3).

Phylogenetic analysis revealed that many isolates clustered according to their serotypes.

We further analyzed whether there is a correlation between the O antigen extracted from WGS and the assigned MLST type or/and chromosomal phylogeny determined by means of cgMLST (Fig. 4). For most of the strains showing the same serotype (3-10 strains), MLST type was throughout correlating; for example for O103:H2 (ST17), O128:H2 (ST25), OgN1 (ST26 (Fig. 4). The O157:H7 strains majorly belonged to ST11, however, in two cases MLST type was ST587 and 1804. Fig. 4 shows that these two O157:H7 strains are located in the same phylogenetic branch with the other ST11 strains and therefore all of the strains are

closely related. This was also the case for several other serotypes belonging to different MLST types, like O26:H11 (ST21, ST29), O153/178:H7 (ST278, ST4975), O117:H7 (ST504, ST5292) and O156:H25 (ST300, ST4942) (Fig. 4).

Conversely, strains sharing the same OAG but harboring different HAGs belonged to different MLST types and distinct phylogenetic branches like O18:H2/21 (ST4017, ST40), O36:H19/14 (ST10, ST1176), O55:H7/9/12 (ST335, ST301, ST 101). It is interesting that the branch of strains belonging to MLST type ST10 comprised a large diversity of serotypes, including O89:H9, O113:H4, O82:H4, OgN12:H32, O127:H40, O38:H26 and O36:H19 (Fig.4). Opposing, O8 antigen strains belonged to a variety of different MLST types and occurred at different branches in the phylogenetic tree (ST23, ST88, ST136, ST162, ST767, ST201 and ST4496) and several O8 strains with the same HAG were not even belonging to the same MLST type (Fig. 4). The four novel O genotypes identified here were found at different branches whereby the two strains sharing OgN-RKI2 were closely related, however the two strains sharing OgN-RKI1 were not. To summarize, the MLST type gives an insight into the phylogenetic relationship of a large fraction of STEC strains, however depending on the serotype, it does not completely reflect serotype nor genomic phylogeny.

WGS-based virulence gene determination highly correlates with PCR-based data.

From genome sequences, we further extracted 27 EHEC, EPEC and EAEC virulence gene markers and 6 gene loci (loci for EAEC AAF/I-IV genes; *aat* operon, *ehx* operon). We observed 99-100% concordance with PCR-derived data concerning STEC markers *stx1/2*, *eaeA* and *hlyA* confirming the high suitability of the PCR-based methods. One *stx1* gene was however not confirmed by WGS data. This might be due to the loss of the *stx1* phage in this strain. Further, two *stx2* genes and four *ehxA* genes were found by WGS analysis which were missed by the PCR method. Fig. 5 shows the distribution of selected STEC/EHEC, EPEC and EAEC virulence gene markers detected by WGS analysis. Interestingly, the heat stable

285 enterotoxin 1 (EAST1) gene was present in more than 50% of the STEC strains analyzed
286 here.
287
288 ***Strains causing HUS and those with exceptional virulence gene combinations.***
289 Among the strains analyzed by WGS, 14 were isolated from cases with HUS or fatal cases
290 (Tab. 2). Nine of those belonged to more frequently found STEC OAGs, such as O26:H11
291 (two strains), O103:H2 (two strains), O113:H21, O145:H28 (two strains), O157:H7 and
292 O157:Hnm; all of them present in the HUSEC collection (6). Further five strains belonged to
293 less frequently found STEC OAGs. The serotypes were O55:H7, O80:H2, O174:H21, and
294 O177:H25 (two strains). Except for the O80 and O177 strains all serovars are present in the
295 HUSEC collection (6). Eight of the 14 strains harbored *stx2a*, three strains *stx2c*, one strain
296 *stx2d* and two strains *stx1a*. These latter two strains of serotype O103:H2 did not comprise an
297 additional *stx2* gene (Tab. 2).
298
299 ***Correlations of stx subtypes with O antigens.***
300 We used the WGS data to get an overview about the *stx* gene subtypes coded in our study
301 strains. For 109 *stx1* positive strains the subtype *stx1a* was found in 55.9%, *stx1c* in 42.2%
302 and *stx1d* in 1.8%. The *stx2* gene was detected in 174 strains and subtype distribution was as
303 follows: 32.0% *stx2a*, 29.8% *stx2*b, 13.0% *stx2*c, 4.5% *stx2*d, 12.6% *stx2*e, 4.6% *stx2*f, 2.9%
304 *stx2*g. For example, all O103, O117 and O182 strains carried *stx1*a. *Stx1c* was found in all
305 O38, O43, O78, O112 and O153/178 strains. *Stx2a* was determined for all O26, O145 and
306 OgN31; *stx2*b for all O2, O110 and OgN1; *stx2e* for all O89, O100 and the majority of O8 (9
307 of 12); *stx2*f for all O63 and O132 and *stx2*g for most of O36 (5 of 7) strains. O157:H7
308 comprised only *stx1*a in 2.4%, exclusively *stx2*a in 34.1%, exclusively *stx2*c in 19.5% and the
309 combination of *stx1*a/*2*a or *stx1*a/*2*c in 14.6% and 29.3% of the strains, respectively.
310

311 **Discussion**

312 In this study we preferentially analyzed STEC showing a) less frequently detected STEC
313 OAGs, b) uncommon O:H combinations, and c) Ont/Orough types. As described above
314 usually a 35% fraction of isolates analyzed at NRC belongs to these categories; in this study
315 we doubled this portion to 70% (Fig. 1A). We set out to compare results of classical
316 serotyping and PCR-based detection of main virulence markers with WGS-derived findings
317 and put those into context with STEC belonging to more frequently found OAGs. Validation
318 of WGS-based methods for the here predominately selected strains is especially important
319 since such strains represent a huge variety and so far the vast majority studies is available for
320 the more common STEC types. Uncommon types induce a substantial percentage of severe
321 disease and large outbreaks and therefore deserve special attention (6, 9, 12, 13).

322

323 In the genomic era, OAG serotyping remains an important epidemiological marker of STEC
324 used as a first indication of strain virulence (30). Therefore, it is important to serotype
325 untypable and rough strains, which is now possible by using genome analysis. Our study
326 shows that WGS data can be used to extract STEC serotypes and virulence markers of the
327 selected strains yielding in about 97-99% results concordant to the more classical methods.
328 Importantly, classification of non-typable or rough strains was possible and even allowed
329 identification of four novel OAG genotypes.

330

331 In this study we identified four novel OAG genotypes of six strains which were found located
332 on five distinct phylogenetic branches (Fig. 4, Tab. 1). nBLAST analysis of the novel OAGCs
333 revealed that OgN-RKI1 is abundant in *Shigella boydii* serovar 19 with a nucleotide identity
334 of over 98% and in one published STEC strain with untypable OAG (Tab. S3). This shows
335 that OgN-RKI1 is present in *Shigella* and STEC. The two strains of the study sharing OgN-
336 RKI1 (OgN-RKI1:H49 strain 16-01174 and OgN-RKI1:H20 strain 16-04698) did not show a

close phylogenetic relationship also indicated by their different MLST ST and H types (Fig. 4, Tab. 1). Homologues of OgN-RKI2, OgN-RKI3 and OgN-RKI4 were found only in *E. coli* (Tab. S3). Interestingly, five of the OgN-RKI3 homologues were serotyped as O59, but the O59-OAGC published by Guo at al. 2005 shares only 64% nucleotide identity to the new OgN-RKI3 (24) (Tab. S3). In addition, the *wzx* and *wzy* genes of both OAGCs are different, displaying a nucleotide identity of 72% for *wzx* and 38% for *wzy*. It appeared however that the OgN-RKI3 strain which harbored *stx2a* showed the same MLST ST with the O86:H51 strain 16-05299. One of the OgN-RKI2 homologues was found in *E. coli* strain P7a serotyped as O20 published by DebRoy et al. 2016 (28). However, the O20-OAGC of strain P7a was already described in 2015 by Iguchi and colleagues (14) and the nucleotide identity between the both O20-OAGCs is only 39.5%. The *wzx* and *wzy* genes for O20 used by CGE SerotypeFinder correspond also to the O-20 OAGC of (14) and are distinct from OgN-RKI2 (14). In two of the OgN-RKI2 homologues, the serovar was identified as OXY24 (31) (Tab. S3). One of the OgN-RKI4 homologues was found in an *E. coli* strain with an O2-like OAGC (32). The three O2:H6 strains of MLST ST 141 of this study do not share the same MLST ST and appear at different branches of the phylogenetic tree (Fig. 4). The finding of four novel OAGC in our study corroborates the importance of genome analysis for strain typing. Therefore, description of further OAGC is expected in the future and it is of great interest to harmonize their designation. To evaluate how to handle new serotypes found by WGS studies an international working group exists since 2017, comprising of persons with leading expertise hosted by Penn State University (US, https://sites.psu.edu/ecolishigella/).

The OAG is one of the most variable bacterial cell components. Driven by strong immunogenic selection, the types of sugars, their arrangement within the O unit, and the linkages between O units vary (33, 34). In *E. coli*, the OAG biosynthesis genes are clustered in the chromosome and flanked by the colonic acid gene cluster (*wca* genes) and the histidine

Journal of Clinical Microbiology

363 biosynthesis cluster (*his* genes). The genes for O unit translocation and chain synthesis,
364 specifically *wzx* (encoding O antigen flippase), *wzy* (encoding O antigen polymerase), and
365 *wzm/wzt* (encoding components of the ABC transporter) are highly variable in sequence and
366 therefore especially suitable for serogroup discrimination (14, 35, 36).
367 Our data and those of others highlight that OAGC distribution does not necessarily follow
368 phylogeny as several serogroups are found at distinct branches of the neighbor joining tree
369 (Fig. 4). This supports the notion that OAGC have been spread across *E. coli* by means of
370 horizontal gene transfer and that frequent exchange may occur (14). This suggestion is also
371 illustrated by the completely distinct genomic organization of the four novel O-AGC which
372 we identified in this study. Only the framing of the gene cluster remains identical but other
373 components, such as *wzy* and *wzx* were found at different locations with different neighboring
374 genes (Fig. 2).
375
376 14 strains were associated with severe disease, specifically HUS and/or death and five of
377 those belonged to less frequently found STEC serovars, namely O55:H7, O80:H2, O174:H21,
378 and O177:H25 (two strains). O55:H7 strains are closely related to O157:H7 strains and both
379 belong to MLST ST11 (37, 38). The O55:H7 strain 17-03136 described here also belonging to
380 ST11, is indeed phylogenetically close to O157:H7 strains and harbors *stx2a*/*eaeA* (Fig. 4,
381 Tab. 2). Those strains are considered as emerging pathogens and HUS cases associated with
382 this serovar have been frequently described (6, 9, 39-41). Similar to the *stx2a* and *eaeA*
383 positive O80:H2 strain 16-03025 analyzed in this study, STEC/EHEC strains of this serovar
384 were reported in HUS patients. Due to their multidrug resistance these strains are considered
385 as a new therapeutic challenge (9, 42, 43). The O80:H2 strains analyzed in our study also
386 showed resistance to several antibiotics (Tab. S1). Zhang et al. analyzed phylogeny and
387 phenotypes of clinical and environmental STEC O174, which may harbor distinct *fliC* H
388 types, such as fliCH5, fliCH21, and fliCH46. They found that only serovar O174:H21

associates with HUS; a serovar which we also found in a HUS patient (44). The strain described here was *stx2d* positive and *eaeA* negative. Cundon et al. reported O174 STEC as an emerging pathogen in Argentina. There, they belong to the most prevalent STEC serogroups (45). We described two O177:H25 strains (both *stx2c* / *eaeA* positive) from HUS patients; one of those was classically serotyped as O177:nt and the other as O177:H25. An O177:Hnm and Hnt strain (*stx2*/*eaeA* positive) was previously isolated from a HUS patient (46).

To conclude, our data show that the large variety of STEC strains required to be typed for public health measures can be well managed by means of genome sequence analyses. The novel WGS-based methods moreover enabled further analysis of strains causing severe clinical symptoms and the description of novel STEC O antigen loci highlighting the potential of the method for detailed future investigations of common but also less frequently detected strain types.

Acknowledgements

We thank all lab partners and cooperating institutions of the public health authorities for sending strains. We further acknowledge Ute Siewert, Ute Strutz, Susanne Puchner, Thomas Garn and Karsten Großhennig for excellent technical assistance. We thank Rita Prager for helpful discussions to the project, Jennifer Bender and the MF2 genome sequencing unit of the Robert Koch Institute for the support in Illumina MiSeq sequencing and Sangeeta Banerji for critical reading of the manuscript.

412 **References:**

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

415 2. RKI. 2017. Infektionsepidemiologisches Jahrbuch meldepflichtiger Krankheiten für 2016.
416 Robert Koch-Institut, Berlin.

417 3. ECDC. 2015. Surveillance of seven priority food - and waterborne diseases in the EU/EEA
418 2010-2012. European Centre for Disease Prevention and Contro, Control ECfDPa, Stockholm.

419 4. Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic *Escherichia coli*. Nat Rev Microbiol
420 2:123-140.

421 5. Scheutz F, Teel LD, Beutin L, Pierard D, Buvens G, Karch H, Mellmann A, Caprioli A,
422 Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien
423 AD. 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and
424 standardizing Stx nomenclature. J Clin Microbiol 50:2951-2963.

425 6. Mellmann A, Bielaszewska M, Köck R, Friedrich A, Fruth A, Middendorf B, Harmsen D,
426 Schmidt M, Karch H. 2008. Analysis of collection of hemolytic uremic syndrome-associated
427 enterohemorrhagic *Escherichia coli*. Emerg Infect Dis 14:1287-1290.

428 7. Thomas S, Holland IB, Schmitt L. 2014. The Type 1 secretion pathway - the hemolysin
429 system and beyond. Biochim Biophys Acta 1843:1629-1641.

430 8. Orskov I, Orskov F, Rowe B. 1984. Six new *E. coli* O groups: O165, O166, O167, O168,
431 O169 and O170. Acta Pathol Microbiol Immunol Scand B 92:189-193.

432 9. Fruth A, Prager R, Tietze E, Rabsch W, Flieger A. 2015. Molecular epidemiological view on
433 Shiga toxin-producing *Escherichia coli* causing human disease in Germany: Diversity,
434 prevalence, and outbreaks. Int J Med Microbiol 305:697-704.

435 10. Preußel K, Höhle M, Stark K, Werber D. 2013. Shiga toxin-producing *Escherichia coli* O157
436 is more likely to lead to hospitalization and death than non-O157 serogroups-except O104.
437 PLoS One 8:e78180.

438 11. RKI. 2011. Final presentation and evaluation of epidemiological findings in the EHEC
439 O104:H4 Outbreak Germany 2011. Institut RK, Berlin.

12. Bielaszewska M, Mellmann A, Zhang W, Köck R, Fruth A, Bauwens A, Peters G, Karch H. 2011. Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. Lance Infect Dis 11:671-676.

13. Frank C, Werber D, Cramer JP, Askar M, Faber M, an der Heiden M, Bernard H, Fruth A, Prager R, Spode A, Wadl M, Zoufaly A, Jordan S, Kemper MJ, Follin P, Müller L, King LA, Rosner B, Buchholz U, Stark K, Krause G, Team HI. 2011. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. N Engl J Med 365:1771-1780.

14. Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson NR. 2015. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. DNA Res 22:101-107.

15. Lindsey R, Pouseele H, Chen J, Strockbine N, Carleton H. 2016. Implementation of Whole Genome Sequencing (WGS) for Identification and Characterization of shiga Toxin-Producing *Escherichia coli* (STEC) in the United States. Front Microbiology 23:766.

16. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. 2015. Rapid and Easy In Silico Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol 53:2410-2426.

17. Jenkins C. 2015. Whole-Genome Sequencing Data for Serotyping *Escherichia coli*-It's Time for a Change1. Journal Clinical Microbiology 53:2402-2403.

18. Iguchi A, Iyoda S, Seto K, Nishii H, Ohnishi M, Mekata H, Ogura Y, Hayashi T. 2016. Six Novel O Genotypes from Shiga Toxin-Producing *Escherichia coli*. Front Microbiol 7:765.

19. Chattaway MA, Dallman TJ, Gentle A, Wright MJ, Long SE, Ashton PM, Perry NT, Jenkins C. 2016. Whole Genome Sequencing for Public Health Surveillance of Shiga Toxin-Producing *Escherichia coli* Other than Serogroup O157. Front Microbiol 7:258.

20. Prager R, Strutz U, Fruth A, Tschäpe H. 2003. Subtyping of pathogenic *Escherichia coli* strains using flagellar (H)-antigens: serotyping versus fliC polymorphisms. Int J Med Microbiol 292:477-486.

21. Schmidt H, Russmann H, Karch H. 1993. Virulence determinants in nontoxinogenic *Escherichia coli* O157 strains that cause infantile diarrhea. Infect Immun 61:4894-4898.

22. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

23. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46.

24. Guo H, Kong Q, Cheng J, Wang L, Feng L. 2005. Characterization of the *Escherichia coli* O59 and O155 O-antigen gene clusters: the atypical wzx genes are evolutionary related. FEMS Microbiol Lett 248:153-161.

25. Guo H, Yi W, Shao J, Lu Y, Zhang W, Song J, Wang PG. 2005. Molecular analysis of the O-antigen gene cluster of *Escherichia coli* O86:B7 and characterization of the chain length determinant gene (wzz). Appl Environ Microbiol 71:7995-8001.

26. Hobbs M, Reeves PR. 1994. The JUMPstart sequence: a 39 bp element common to several polysaccharide gene clusters. Mol Microbiol 12:855-856.

27. Wang L, Reeves PR. 1998. Organization of *Escherichia coli* O157 O antigen gene cluster and identification of its specific genes. Infect Immun 66:3545-3451.

28. DebRoy C, Fratamico PM, Yan X, Baranzoni G, Liu Y, Needleman DS, Tebbs R, O'Connell CD, Allred A, Swimley M, Mwangi M, Kapur V, Raygoza Garay JA, Roberts EL, Katani R. 2016. Comparison of O-Antigen Gene Clusters of All O-Serogroups of *Escherichia coli* and Proposal for Adopting a New Nomenclature for O-Typing. PLoS One 11:e0147434.

29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 215:403-410.

30. Valilis E, Ramsey A, Sidiq S, H.L. D. 2018. Non-O157 Shiga toxin-producing *Escherichia coli*-A poorly appreciated enteric pathogen: Systematic review. Int J Infect Dis 76:82-86.

31. Gangiredla J, Mamme MK, Barnaba TJ, Tartera C, Gebru ST, Patel IR, Leonard SR, Kotewicz ML, Lampel KA, Elkins CA, Lacher DW. 2017. Species-wide collection of *Escherichia coli* isolates for examination of genomic diversity. Genome Anounc 5:e01321-1317.

32. Delannoy S, Beutin L, Mariani-Kurkdjian P, Fleiss A, Bonacorsi S, Fach P. 2017. The *Escherichia coli* Serogroup O1 and O2 Lipopolysaccharides Are Encoded by Multiple O-antigen Gene Clusters. Front Cell Infect Microbiol 7:30.

33. Bazaka K, Crawford RJ, Nazarenko EL, Ivanova EP. 2011. Bacterial extracellular polysaccharides. Adv Exp Med Biol 715:213-226.

34. Wang L, Qu W, Reeves PR. 2001. Sequence analysis of four Shigella boydii O-antigen loci: implicaiton for *Escherichia coli* and Shigella relationships. Infection § Immunity 69:6923-6930.

35. Iguchi A, von Mentzer A, Kikuchi T, Thomson NR. 2017. An untypeable enterotoxigenic *Escherichia coli* represents one of the dominant types causing human disease. Microb Genom 3:e000121.

36. Cheng J, Wang Q, Wang W, Wang Y, Wang L, Feng L. 2006. Characterization of *E. coli* O24 and O56 O antigen gene clusters reveals a complex evolutionary history of the O24 gene cluster. Curr Microbiol 53:470-476.

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

38. Whittam TS, Wolfe ML, Wachsmuth IK, Orskov F, Orskov I, Wilson RA. 1993. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. Infect Immun 61:1619-1620.

39. McFarland N, Bundle N, Jenkins C, Godbole G, Mikhail A, Dallmann T, O'Connor C, McCarthy N, O'Connell E, Teacy J, Dabke G, Mapstone J, Landy Y, Moore J, Partridge R, Jorgensen F, Willis C, Mook P, Rawlings C, Acornley R, Featherstone C, Gayle S, Edge J, McNamara E, Hawker J, Balasegaram S. 2017. Resurrent seasonal outbreak of an emerging serotype of Shiga toxin-producing *Escherichia coli* (STEC O55:H7 Stx2a) in the south west of England, July 2014 to September 2015. Euro Surveill 22:30610.

40. Marejková M, Bláhová K, Janda J, Fruth A, Petrás P. 2013. Enterohemorrhagic *Escherichia coli* as causes of hemolytic uremic syndrome in the Czech Republik. PLoS One 8:e73927.

41. Bielaszewska M, Janda J, Blahová K, Feber J, Potzník V, Soucková A. 1996. Verocytotoxin-producing *Escherichia coli* in children with hemolytic uremic syndrome in the Czech Republic. Clin Nephrol 46:42-44.

42. Wijnsma KL, Schijvens AM, Rossen JWA, Kooistra-Smid AMDM, Schreuder MF, van de Kar NCAJ. 2017. Unusual severe case of hemolytic uremic syndrome due to Shiga toxin 2d-producing *E. coli*. Pediatr Nephrol 32:1263-1268.

43. Soysal N, Mariani-Kurkdjian P, Smail Y, Liguori S, Gouali M, Loukiadis E, Fach P, Bruyand M, Blanco J, Bidet P, Bonacorsi S. 2016. Enterohemorrhagic *Escherichia coli* Hybrid Pathotype O80:H2 as a New therapeutic Challenge. Emerg Infect Dis 22:1604-1612.

44. Zhang W, Nadirk J, Kossow A, Bielaszewska M, Leopold SR, Witten A, Fruth A, Karch H, Ammon A, Mellmann A. 2014. Phylogeny and phenotypes of clinical and environmental Shiga toxin-producing *Escherichia coli* O174. Envion Microbiol 16:963-976.

45. Cundon C, Carbonari CC, Zolezzi G, Rivas M, Bentancor A. 2018. Putative virulence factors and clonal relationship of O174 Shiga toxin-producing *Escherichia coli* isolated from human, food and animal sources. Vet Microbiol 215:29-34.

46. Seto K, Taguchi M, Kobayashi K, Kozaki S. 2007. Biochemical and molecular characterization of minor serogroups of Shiga toxin-producing *Escherichia coli* isolated from humans in Osaka prefecture. J Vet Med Sci 69:1215-1222.

Figure Legends:

Fig. 1: Classically determined serotypes of total 2015-2017 NRC STEC strains (A) and of the 232 strains chosen for WGS (B).

Fig. 2: Four novel O antigen gene clusters (OgN-RKI1 to OgN-RKI4) identified in this study.

Fig. 3: Phylogenetic analysis of Wzx and Wzy homologs of the four novel OAGCs OgN-RKI1-4 (red labeled), *E. coli* O serotype strains, OX groups and OgN groups reference strains based on amino acid sequence.

Fig. 4: Chromosomal phylogeny of 232 genome-sequenced STEC strains represented as neighbor joining tree (NJT) and its relation to serogroup and 7 gene MLST type. Ridom SeqSphere+ was used to create the NJT based on 2513 targets from *E. coli* cgMLST Enterobase with pairwise ignoring missing values. Labels are containing the strain number and the MLST-type separated by comma. Different colors are assigned to distinct MLST types. OAGs are depicted in the outer circle. The new OAGs found in this study and are highlighted in yellow with red text. Further new OAGs (OgN) found by WGS are also labeled in red.

Fig. 5: Summary of selected *E. coli* pathovar virulence genes extracted by WGS analysis in the 232 STEC strains analyzed.

564 **Table Legends**:

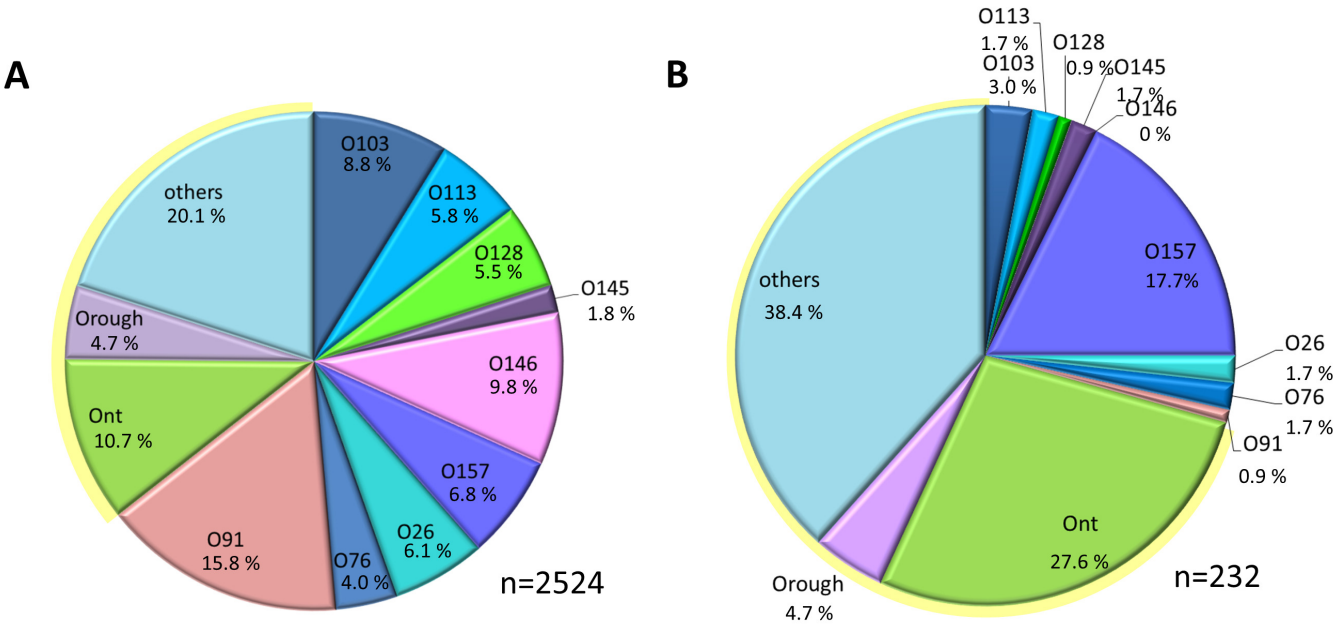
565

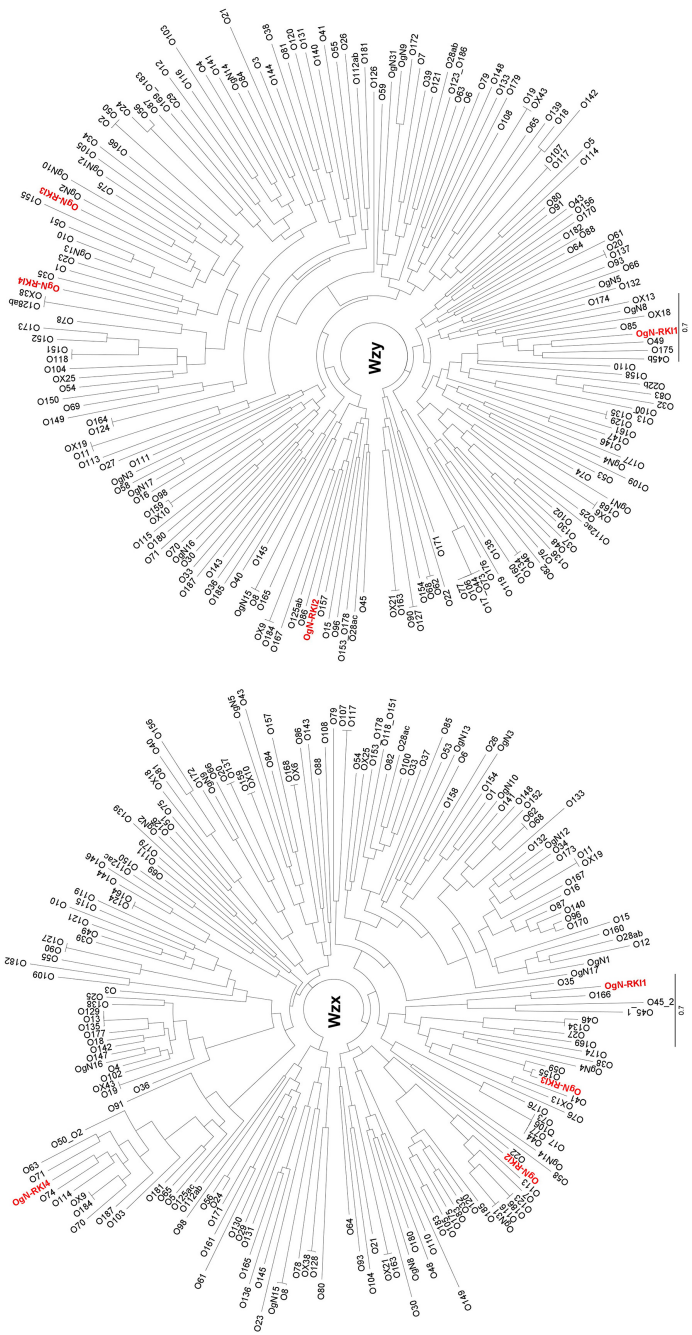
566 Tab. 1: Serotype, MLST ST and virulence gene profile of the six STEC strains with novel
567 OAGCs

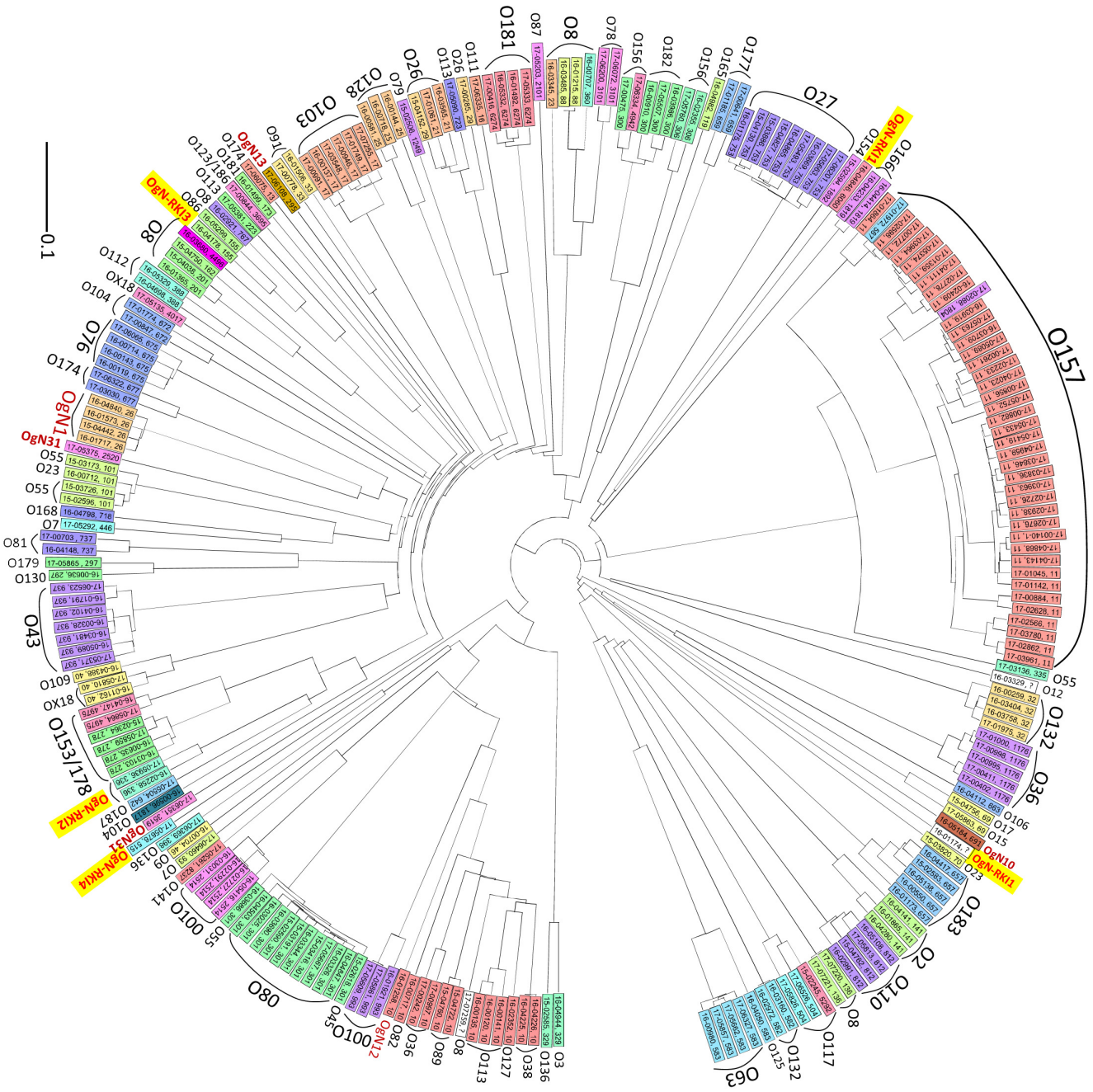
568

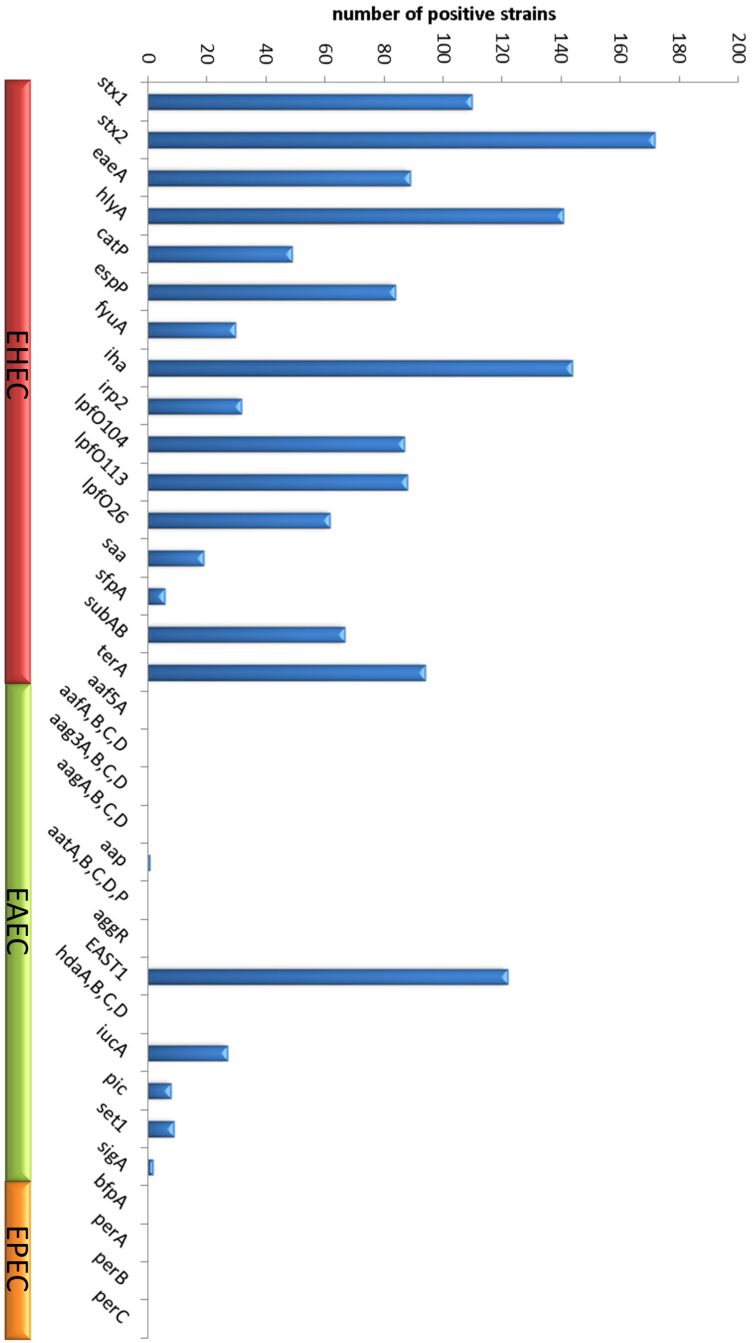
569 Tab. 2: Serotype, MLST ST and virulence gene profile of EHEC strains causing HUS or
570 death.

571









Journal of Clinical Microbiology

RKI No.	O	H	MLST	*stx1*	*stx2*	*eaeA*	*hlyA*	*espP*	*fyuA*	*iha*	*irp2*	*catP*	lpf_{O26}	lpf_{O113}	*subAB*	*terA*	*EAST1*
16-01174	OgN-RKI1	49	9300	-	+/b	-	+	-	-	+	-	-	-	-	+	+	+
16-04846	OgN-RKI1	20	6060	+/c	-	-	-	-	-	-	-	-	-	-	-	-	-
16-02258	OgN-RKI2	16	336	+/c	-	-	-	-	-	-	-	+	+	-	-	-	-
16-04178	OgN-RKI3	21	155	-	+/a	-	-	-	-	-	-	-	-	+	-	-	-
17-05676	OgN-RKI4	29	515	-	+/b	-	-	-	-	+	-	-	-	-	-	-	+
17-05936	OgN-RKI2	16	336	+/c	-	-	-	-	-	-	-	+	-	-	-	-	+

O	H	MLST ST	RKI-No.	clinics	*stx1*	*stx2*	*eaeA*	*hlyA*	*espP*	*fyuA*	*iha*	*irp2*	*catP*	lpf_{O26}	lpf_{O113}	*sfpA*	*subAB*	*terA*	*EAST1*	*iucA*	*pic*	*set1*	*sigA*	*aap*	*aatA*	*aagA, C*
26	11	29	17-00285	HUS	-	+/a	+	+	+	+	+	+	-	+	+	-	-	+	-	-	-	-	-	-	-	-
26	11	21	17-01061	HUS	-	+/a	+	+	+	+	+	+	+	+	+	-	-	+	+	-	-	-	-	-	-	-
55	7	335	17-03136	HUS	-	+/a	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-
80	2	301	16-03025	HUS	-	+a	+	+	+	-	+	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-
103	2	17	17-01749	HUS	+/a	-	+	+	-	-	+	-	-	+	-	-	-	+	-	+	-	-	-	-	-	-
103	2	17	17-03548	HUS	+/a	-	+	+	-	-	+	-	-	+	-	-	-	+	-	+	-	-	-	-	-	-
113	21	223	17-05381	HUS	-	+/a	-	+	+	-	+	-	-	-	+	-	+	-	-	-	-	-	-	-	-	-
145	28	32	16-03404	HUS	-	+/a	+	+	+	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-
157	7	11	17-01864	HUS	-	+/a	+	+	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-
157	7	587	17-01972	HUS	-	+/a	+	+	-	-	-	-	-	-	-	+	-	-	+	-	-	-	-	-	-	-
174	21	677	17-03030	HUS	-	+/d	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-
177	25	659	17-00641	HUS	-	+/c	+	+	+	-	+	-	+	-	+	-	-	+	+	+	-	-	-	-	-	-
177	25	659	17-01185	Death/ HUS	-	+/c	+	+	+	-	+	-	+	-	+	-	-	+	+	+	-	-	-	-	-	-
145	28	32	17-01975	Death	-	+/c	+	+	+	-	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-