Microevolution of epidemiological highly relevant non-O157 enterohemorrhagic *Escherichia coli* (EHEC)

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium

(Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy

of the Freie Universität Berlin

by

M. Sc. Inga Eichhorn

from Rüdersdorf near Berlin

Berlin, 2016

This work was carried out in the period from June 2011 until April 2016 under the supervision of Prof. Dr. Lothar H. Wieler at the Centre for Infection Medicine, Institute of Microbiology and Epizootics, Department of Veterinary Medicine of the Freie Universität Berlin.

1st Reviewer: Prof. Dr. Lothar H. Wieler

2nd Reviewer: Prof. Dr. Heribert Hofer DPhil

Date of defense: 27.09.2016

ABSTRACT

Enterohemorrhagic *Escherichia coli* (EHEC) are the cause of diarrhea, bloody diarrhea, hemorrhagic colitis (HC) and the potentially fatal hemolytic uremic syndrome (HUS) in humans. The key virulence factor of EHEC is the bacteriophage-encoded Shiga-Toxin gene (*stx*). In addition, EHEC can harbor the locus of enterocyte effacement (LEE), which is located on a pathogenicity island, and coffers the ability to cause attaching and effacing (A/E) lesions on epithelial cells as a result of colonization of the intestinal tract. The most common EHEC serotype isolated worldwide is O157:H7. Nevertheless, a variety of non-O157 EHEC have emerged as serious causes of HUS and diarrhea in humans worldwide. The most important non-O157 O serogroups causing one third of the EHEC infections in Germany, are O26, O103, O111 and O145. We hypothesized that these non-O157 differ in their phylogeny and aimed to determine the underlying microevolution leading to the emergence of such non-O157 EHEC.

In this thesis I provide evidence to explain the population structure of the two most-important non-O157 EHEC, O26 and O111 EHEC, and further propose a model of the microevolution of these EHEC, including also the common ancestor of both lineages. The microevolutionary model of O26 and O111 EHEC is based on results of several analyses of the genome of a large collection of non-O157 EHEC strains and the crucial inclusion of atypical enteropathogenic *E. coli* (aEPEC) into this collection. Application of multilocus sequence typing (MLST) on a diverse collection of 250 non-O157 STEC and EHEC revealed a cluster formation of strains of O serogroups O26 and O111 in one single sequence type complex, STC29. The additional presence of aEPEC, that differ from STEC/ EHEC merely in their absence of the *stx*-converting bacteriophage, but share the same O serogroups, among the STEC/ EHEC strains in common sequence types (STs) of STC29, suggests an ongoing microevolutionary scenario, characterized by a bidirectional-conversion, in which the phage-encoded *stx*-gene is transferred between aEPEC and STEC/ EHEC.

To test this hypothesis, we performed whole genome sequencing of 99 selected strains (aEPEC n=20, STEC/ EHEC n=79) designated as STC29 and analyzed the single nucleotide polymorphisms (SNPs) of the maximum common genome (MCG) of those strains in order to gain more detailed information on the underlying population structure and microevolution. The resulting minimum spanning tree (MSTree) of the MCG-based SNP-analysis revealed three distinct clusters. Cluster 1 harbored strains of O serogroup O111 also designated as ST16 with MLST. Interestingly, the distinct Cluster 2 included only O26 aEPEC strains of ST29, while the more heterogeneous Cluster 3 combined STEC/ EHEC as well as aEPEC strain of O serogroup O26 that were only roughly separated into strains of ST29 and ST21.

The analysis of the presences or absence of accessory virulence associated genes (VAGs) confirmed the results of the SNP-analysis, and suggests a parallel evolution of the MCG of those strains and the acquisition of virulence genes. Furthermore, the analysis of insertion sites for mobile genetic elements, with respect to their occupation with phage-related genes and insertion elements, resulted in a similar relation of the analyzed strains. Consequently, the common results of MLST, SNP-analysis of the MCG, the presence of VAGs and occupation of insertion sites led to the development of a microevolutionary model of STEC/ EHEC of O serogroups O26 and O111, which developed as two

distinct lineages from a common aEPEC ancestor of ST29 by lysogenic conversion with *stx*-converting bacteriophages.

Moreover, these analyses and the development of the microevolutionary model were performed on human (n=53) and bovine (n=45) STEC/ EHEC and aEPEC strains, representing the most important hosts of such strains. None of the analyses revealed a separated grouping of strains based on host species. Hence, these strains do not appear to harbor host-specific genomic alterations, neither within the MCG nor in the acquired VAGs, and therefore do not appear to emerge by adaptation to a specific niche. These results strongly support the zoonotic nature of aEPEC and STEC/ EHEC.

In conclusion, STEC/ EHEC of O serogroups O26 and O111 originate from a common ancestor, further aEPEC and STEC/ EHEC of those serogroups share a common phylogeny and are *bona fide* zoonotic agents.

KEYWORDS

Enterohemorrhagic *Escherichia coli* (EHEC), atypical enteropathogenic *Escherichia coli* (aEPEC), microevolutionary model, multilocus sequence typing (MLST), SNP-analysis, lysogenic conversion, *stx*-converting bacteriophage

ZUSAMMENFASSUNG

Enterohämorrhagische *Escherichia coli* (EHEC) verursachen Infektionen des menschlichen Intestinaltraktes und führen zu Durchfall, blutigem Durchfall, hämorrhagischer Kolitis und dem potentiell tödlichen, hämolytisch urämischen Syndrom (HUS). Das Bakteriophagen-kodierte Shiga-Toxin-Gen (*stx*) ist der wichtigste Virulenzfaktor von EHEC. Zusätzlich können EHEC eine Pathogenitätsinsel, den sogenannten ‚locus of enterocyte effacement' (LEE) besitzen, durch den sie bei der Besiedlung des Intestinaltraktes ‚attaching and effacing' (A/E) -Läsionen an Epithelzellen verursachen. Der am weltweit häufigsten isolierte EHEC-Serotyp ist O157:H7, aber eine Vielzahl von non-O157 EHEC wurde weltweit von Patienten mit Durchfallerkrankungen sowie HUS isoliert. Die wichtigsten non-O157 O-Serogruppen, die auch ein Drittel der EHEC-Infektionen in Deutschland verursachen, sind O26, O103, O111 und O145. Wir nehmen an, dass diese non-O157 EHEC unterschiedliche Phylogenien aufweisen und beabsichtigen, das zu Grunde liegende mikroevolutionäre Szenario, dass zur Entstehung dieser Stämme führte, aufzuklären.

In dieser Dissertation konnten wir nicht nur die Populationsstruktur der zwei wichtigsten non-O157 EHEC, nämlich O26 und O111 EHEC, aufklären, sondern auch ein Modell für die Mikroevolution dieser EHEC, einschließlich des gemeinsamen Vorgängers beider Abstammungslinien, vorstellen. Das mikroevolutionäre Modell der O26 und O111 EHEC basiert auf den Ergebnissen einer Vielzahl von Analyseschritten die an den Genomsequenzen einer großen Sammlung von non-O157 EHEC durchgeführt wurden. Hierbei war vor allem die Einbeziehung von atypischen enteropathogenen *Escherichia coli* (aEPEC) in die Analysen von großer Wichtigkeit. Zuerst führten wir eine Multilokus Sequenztypisierung (MLST) von 250 non-O157 EHEC und Shiga-Toxin produzierenden *E. coli* (STEC) durch, und stellten eine Anhäufung von Stämmen der O-Serogruppen O26 und O111 in einem einzigen Sequenztypkomplex fest, STC29. Genau wie STEC/ EHEC-Stämme des STC29, wurden aEPEC-Stämme ebenfalls den einzelnen Sequenztypen von STC29 zugeordnet; diese aEPEC unterscheiden sich von STEC/ EHEC nur durch die Abwesenheit des *stx*-konvertierenden Bakteriophagen und weisen ebenfalls die gleichen O-Serogruppen wie STEC/ EHEC auf. Die Akkumulation dieser Pathotypen in gleichen STen weist auf ein zu Grunde liegendes mikro-evolutionäres Szenario hin, genauer auf bidirektionale Konversion, bei der das Phagen-kodierte *stx*-Gen zwischen aEPEC und STEC/ EHEC übertragen wird.

Um diese Hypothese zu testen, wurden 99 Stämme, des STC29 (aEPEC n=20, STEC/ EHEC n=79), Ganzgenom-sequenziert und anschließend Einzelnukleotid-Polymorphismen (SNPs) des größten gemeinsamen Genanteils aller Stämme (‚maximum common genome', MCG) analysiert um bessere Einblicke in die Populationsstruktur dieser Stämme zu erlangen. In dem resultierenden ‚Minimum Spanning Tree (MSTree)' der MCG-basierten SNP-Analyse zeigten sich drei klar abgetrennte Cluster. Cluster 1 umfasste alle Stämme der O-Serogruppe O111, die durch MLST zusätzlich als ST16 bestimmt wurden. Cluster 2 fasste interessanterweise ausschließlich O26 aEPEC des ST29 zusammen, wohingegen das heterogenere Cluster 3 STEC/ EHEC und aEPEC der O-Serogruppe O26 beinhaltet, die entweder als ST29, oder als ST21 bestimmt wurden.

Die Analyse des Vorhandenseins von Virulenz-assoziierten Genen (VAGen) des akzessorischen Genoms bestätigte die Ergebnisse der SNP-Analyse und deutet darauf hin, dass die Evolution des MCG dieser Stämme und die Aufnahme von VAGen parallel verliefen. Des Weiteren wurden bekannte

Insertionsstellen in diesen Stämmen auf die Integration von mobilen genetischen Elementen, also Phagen oder anderen Insertionselementen, hin untersucht. Die sich ergebende Populationsstruktur war erneut vergleichbar mit den vorangegangenen Ergebnissen.
Die übereinstimmenden Ergebnisse von MLST, SNP-Analyse des MCGs, Vorhandensein von VAGen und Integration von mobilen genetischen Elementen dienten als Grundlage für die Entwicklung eines Mikroevolutions-Modells von STEC/ EHEC der O-Serogruppen O26 und O111. Nach diesem Modell entwickelten sich O26 und O111 STEC/ EHEC in zwei Abstammungslinien von einem gemeinsamen Vorgänger, einem aEPEC des STs 29, vermutlich durch lysogene Konversion mit einem *stx*-konvertierenden Bakteriophagen.
Die Analysen bzw. die Entwicklung des Modells wurden zudem mit humanen (n=53), als auch mit bovinen (n=45) STEC/ EHEC und aEPEC Stämmen durchgeführt und repräsentieren somit die häufigsten Wirte dieser Stämme. Da keine der Analysen eine separate Gruppierung von Stämmen eines Wirtes aufzeigte, weisen die Stämme also keine wirtsspezifische Anpassung des Genoms, weder des MCGs noch bei den akzessorischen VAGen, auf und haben sich folglich nicht durch Anpassung an eine bestimmte Nische entwickelt. Darüber hinaus belegt das Fehlen von wirtsspezifischen Clustern auch den zoonotischen Charakter von aEPEC und STEC/ EHEC der genannten Serotypen.
Abschließend lässt sich sagen, dass STEC/ EHEC der O-Serogruppen O26 und O111 von einem gemeinsamen Vorgänger abstammen, aEPEC und STEC/ EHEC dieser O-Serogruppen eine gemeinsame Phylogenie aufweisen, und *bona fide* Zoonoseerreger sind.

SCHLAGWÖRTER

Enterohämorrhagische *Escherichia coli* (EHEC), atypische enteropathogene *Escherichia coli* (aEPEC), Mikroevolutionsmodell, Multilokus Sequenztypisierung (MLST), SNP-Analyse, lysogene Konversion, *stx*-konvertierende Bakteriophagen

DANKSAGUNG

Ich danke Herrn Prof. Dr. Lothar H. Wieler für die Möglichkeit am Institut für Mikrobiologie und Tierseuchen zu promovieren und für die Überlassung des interessanten Forschungsthemas. Ich danke Ihnen auch, dass Sie mir im Rahmen Ihrer Kooperation mit dem Wellcome Trust Sanger Institut (Hinxton, UK) ermöglicht haben, dort zwei Wochen zu verbringen und neue Techniken und Kollegen kennenzulernen. Vor allem aber danke ich Ihnen für Ihre Ratschläge und Diskussionsbereitschaft während meiner Promotionszeit und die Unterstützung beim Verfassen der Promotionsschrift.

Herrn Prof. Heribert Hofer aus dem Leibniz-Institut für Zoo- und Wildtierforschung (IZW) danke ich für die Bereitschaft das Zweitgutachten meiner Arbeit zu übernehmen.

Torsten Semmler möchte ich für seine bioinformatische Unterstützung, wie bei dem Assembly der Genomdaten, der Erstellung des MCGs und der SNP-Identifikation, danken, sowie für seine Hilfe und Einarbeitung in die verwendeten Programme zur Genomanalyse.

Bianca Kinnemann danke ich für Ihre ausgezeichnete praktische Unterstützung im Labor, vor allem in meiner Anfangszeit am Institut.

Alexander Mellmann aus dem Institut für Hygiene am Universitätsklinikum Münster danke ich für die Bereitstellung von 19 Ganzgenom-Sequenzen von EHEC-Stämmen. Derek Pickard und Gordon Dougan aus dem Wellcome Trust Sanger Institut in Hinxton danke ich für die Ganzgenom-Sequenzierung von *E. coli* Stämmen die in dieser Arbeit analysiert wurden.

Für die Finanzierung meiner Doktorarbeit bedanke ich mich bei dem BMBF-geförderten Verbundprojekt „Food-Borne Zoonotic Infections of Humans“ (FBI-Zoo; Kennzeichen 01KI1012A)). Weiterhin danke ich den folgenden Mitgliedern des Forschungsverbundes für die Isolation und Bereitstellung von *E. coli* Stämmen: Alexander Mellmann und Helge Karch, aus dem Institut für Hygiene am Universitätsklinikum Münster, Jürgen Heesemann aus dem Max von Pettenkofer-Institute an der Ludwig-Maximilians-Universität München und Peter Valentin-Weigand aus dem Institut für Mikrobiologie, Stiftung Tierärztliche Hochschule Hannover. Außerdem danke ich Angelika Fruth aus dem RKI in Wernigerode für die Durchführung der Serotypisierung.

Weiterhin danke ich dem DFG-geförderten Graduiertenkolleg IRGT GRK1673 „Functional Molecular Infection Epidemiology“ für die finanzielle Unterstützung und für die Möglichkeit interessante Weiterbildungskurse zu besuchen.

Allen Kollegen und Kolleginnen am Institut für Mikrobiologie und Tierseuchen möchte ich für das freundschaftliche Arbeitsklima, ihre Hilfsbereitschaft und Diskussionsfreudigkeit bei Problemen danken. Danke, dass ihr meine Promotionszeit so angenehm gestaltet habt.

Von ganzem Herzen danke ich meinen Freunden und meiner Familie für die moralische Unterstützung und die aufbauenden Worte. Claudi, Nobi und Chris danke ich im Besonderen dafür, dass sie immer an mich geglaubt und mich ständig neu motiviert haben.

Meinem Freund Stefan danke ich für sein uneingeschränktes Verständnis, seine Geduld und sein offenes Ohr. Danke, dass du immer für mich da warst und noch bist.

“Flamingoes and mustard both bite.
And the moral of that is—
"Birds of a feather flock together.""

Alice’s Adventures in Wonderland, Lewis Carroll, 1865

TABLE OF CONTENTS

FIGURES

TABLES

ABBREVIATIONS

A/E lesion	‘attaching and effacing’ lesion
aEPEC	atypical enteropathogenic *Escherichia coli*
bp	base pairs
cGMP	Cyclic guanosine monophosphate
DEC	diarrheagenic *Escherichia coli*
DNA	Deoxyribonucleic acid
dsDNA	double-stranded DNA
E. coli	*Escherichia coli*
EAF	EPEC adherence factor
EHEC	enterohemorrhagic *Escherichia coli*
EPEC	enteropathogenic *Escherichia coli*
ETEC	enterotoxigenic *Escherichia coli*
ExPEC	extraintestinal pathogenic *Escherichia coli*
GC-content	guanine-cytosine content
GEI	genomic islands
HC	hemorrhagic colitis
HGT	horizontal gene transfer
HUS	hemolytic uremic syndrome
IE	insertion element
InPEC	intestinal pathogenic *Escherichia coli*
kb	kilo base pairs
kDa	kilo Dalton
LEE	locus of enterocyte effacement
LPS	lipopolysaccharide
MCG	maximum common genome
MLST	multilocus sequence typing
MSTree	minimum spanning tree
NM	non-motile
ORF	open reading frame
PCR	polymerase chain reaction
PAI	pathogenicity island
RNA	Ribonucleic acid
SLV	single locus variant
SNP	single nucleotide polymorphism
ssDNA	single-stranded DNA
ST	sequence type
STC	sequence type complex
STEC	Shiga-Toxin producing *Escherichia coli*
Stx	Shiga-Toxin
tRNA	transfer-RNA
TTSS	type three secretion system
VAG	virulence associated gene
WGS	whole genome sequence

1 INTRODUCTION

1.1 Commensal and pathogenic *Escherichia coli*

In 1885 Theodor Escherich isolated *Escherichia coli* (*E. coli*), which was initially named *Bacterium coli commune,* from the feces of a healthy infant [1]. *E. coli* is a Gram-negative, rod shaped bacteria that is a commensal of the physiological intestinal microbiota of mammals. Besides being present in the gastrointestinal tract of warm-blooded animals, as mammals and birds, *E. coli* has been also isolated from the intestine of other vertebrates like amphibian and fish [2, 3]. The ubiquitous occurrence of *E. coli* marks its ability to adapt to diverse environmental conditions such as water and soil [4-6]. *E. coli* is a well-established laboratory model organism used in genetic engineering, molecular-biology, and the biotechnological production of therapeutic proteins such as insulin [7].

On the other hand pathogenic *E. coli* are the cause of a broad spectrum of human and animal diseases ranging from infections of the gastrointestinal tract, due to *E. coli* broadly classified as intestinal pathogenic *E. coli* (InPEC), to infections of extraintestinal sites caused by *E. coli* designated as extraintestinal pathogenic *E. coli* (ExPEC). The combination of specific virulence features, the phenotype displayed in *in vitro* infection models and the clinical appearance associated with infections are used to discriminate *E. coli* into certain pathotypes [8] (Figure 1).

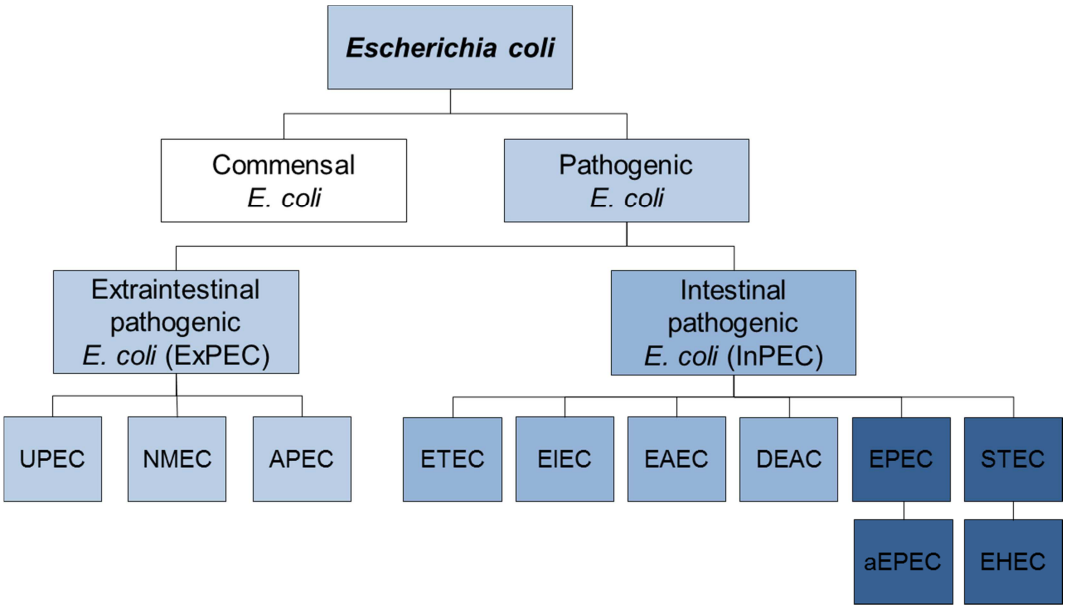

Figure 1 The pathotypes of *E. coli* are subdivided into those which affect extraintestinal sites of the natural host, uropathogenic *E. coli* (UPEC), neonatal meningitis causing *E. coli* (NMEC), and avian pathogenic *E. coli*, and those which affect the intestinal sites of the host, enterotoxigenic (ETEC), enteroinvasive (EIEC), enteroaggregative (EAEC), and diffusely adherent *E. coli* (DAEC), enteropathogenic (EPEC) and atypical enteropathogenic *E. coli* (aEPEC), as well as Shiga-Toxin producing (STEC) and enterhemorrhagic *E. coli* (EHEC).

E. coli is the most common cause of sepsis and bacteremia. The most frequent extraintestinal *E. coli* infection in various hosts is urinary tract infection (UTI) that is caused by uropathogenic *E. coli* (UPEC). ExPEC are also the leading cause of neonatal meningitis (NMEC) and neonatal sepsis. An animal pathotype of ExPEC that causes primarily respiratory tract infection, and associated pericarditis in poultry like chicken and turkey is the avian pathogenic *E. coli* (APEC) [9, 10]. Intestinal pathogenic

E. coli are the major cause of diarrheic infections and are subdivided into six well-characterized pathotypes on the basis of their pathogenic features [11]:

(i) Enterotoxigenic *E. coli* (ETEC) are defined as *E. coli* that express at least one of two plasmid-encoded enterotoxins, either heat-stable enterotoxin (ST) or heat-labile enterotoxin (LT) [11]. Those enterotoxins lead to a deregulation of membrane ion-channels of the epithelial membrane. The ETEC colonization of the small bowel enterocytes is conveyed by proteinaceous fimbrial colonization factors (CFs), so called CFA (colonization factor antigen), CS (coli surface antigen) or PCF (putative colonization factor) which are numbered. CFA/I, CFA/II and CFA/IV are the most common colonization factors found in 75% of human ETEC isolates [8, 15]. With 2.5 million cases and 700,000 deaths in children under 5 years ETEC is the most common bacterial cause of infantile diarrhea, but also the cause of adult diarrhea and traveler's diarrhea in underdeveloped nations [16].
(ii) The biochemical traits of Enteroinvasive *E. coli* (EIEC) are closely related to those of *Shigella* species as well as the dysentery this pathogen causes. EIEC are responsible for shigellosis that leads to ulceration on the epithelium of the colonic mucosa followed by bloody or mucoid diarrhea [12]. This is caused by an invasion process into various cell types e.g. epithelial cells, macrophages and red blood cells. The invasion of the bacteria is mediated through a type three secretion system (TTSS) that is encoded on a ca. 220 kb virulence plasmid, and injects a distinct set of toxins into the host cell. First the epithelial cells are penetrated and the bacteria invade through a rearrangement of the cell cytoskeleton, then the bacteria lyse the endocytic vacuole and access the cytoplasm where they multiply intracellular. They move through the cytoplasm of the infected cells enabled by an actin polymerization at one pole of the bacterial cell and thus are able to extend to adjunct epithelial cells [13]. Due to the similarity of disease outcome caused by EIEC and *Shigella spp*. infections with EIEC are seldom identified as the cause of shigellosis. The disease burden was estimated to 150 million cases and one million deaths per year in the developing world [14].
(iii) Enteroaggregative *E. coli* (EAEC) assemble in aggregative adherence (AA) on cultured HEp-2 cells and appear as characteristic "stagged-brick" pattern in *in vitro* experiments. The AA-phenotype is related to the presence of a 60 MDa plasmid (pAA), which encodes the global virulence regulator *aggR*. AggR regulates the expression of adherence factors, a dispersin protein (aap) that regulates the anti-aggregation phenotype, and a large cluster of genes encoded on the EAEC chromosome [15]. The EAEC pathogenesis is progressed in three major steps and each involves plasmid-encoded traits. First the bacteria adhere to the intestinal mucosa. This initial colonization is facilitated by aggregative adherence fimbriae (AAF). Secondly the mucus production is enhanced, leading to the formation of a mucus-containing biofilm with embedded EAEC on the enterocyte surface. Finally an inflammatory response is induced due to the release of cytotoxins that results in mucosal toxicity and intestinal secretion [16, 17]. EAEC is an emerging cause of diarrhea, and has been implicated in acute and chronic diarrhea among children, adults and HIV infected patients in both industrialized and developing countries. Besides ETEC, EAEC is the second most common cause of traveler's diarrhea. EAEC was isolated with a median of 15% from children living in developing countries and 4% from children living in industrialized countries causing acute diarrheal illness [18].
(iv) Diffusely adherent *E. coli* (DAEC) show a characteristic diffuse adherence (DA) phenotype on cultured HEp-2 and HeLa cells that appears in a scattered pattern over the entire cell surface. DAEC

express various Afa/ Dr adhesins that recognize the brush border-associated decay-accelerating factor (DAF), a molecule that is highly expressed on the apical surface of polarized epithelial cells. The attachment of DAEC to the brush border induces lesions due to cytoskeleton rearrangements which lead to destruction or partially rearrangement of the microvilli and form finger-like projections extending from the surface of infected cells [19]. DAEC cause watery diarrhea that can become persistent in young children in developing countries as well as industrialized countries, but are also associated with recurring UTIs. Older children and adults have been found to carry DAEC intestinally without symptoms [20]. The detection methods for DAEC are still under development therefore no universal method is available for the detection of DAEC in clinical settings [19].

(v) Enteropathogenic *E. coli* (EPEC) induce attaching and effacing (A/E) lesions on epithelial cells. The ability of EPEC to cause A/E lesions is conferred by a 35 kb pathogenicity island (PAI) called the locus of enterocyte effacement (LEE), that encodes for a TTSS, effector proteins, and chaperons. EPEC attach to the intestinal epithelial cells where they efface the microvilli and rearrange the cytoskeleton of the cells due to accumulation of polymerized cytoskeletal proteins. This rearrangement leads to the formation of pedestal like structures directly beneath the adherent bacteria. Typical EPEC harbor a 70-100 kb EAF (EPEC adherence factor) plasmid that encodes for a bundle forming pilus (BFP) used mainly for the adherence to other bacteria. In contrast atypical EPEC (aEPEC) harbor the LEE, but do not possess the EAF plasmid [8, 11]. EPEC are the main cause of persistent diarrhea in children and adults worldwide. An estimate of 79,000 deaths in children under 5 years was calculated for 2011 [21]. Based on results of molecular methods 5% - 10% of the cases of infantile diarrhea in the developing world are caused by EPEC. It was suggested that aEPEC are more prevalent than typical EPEC and are associated with a significantly longer duration of diarrhea [22].

(vi) Shiga-Toxin producing *E. coli* (STEC) characteristically express the cytotoxic Shiga-Toxin (Stx). The two Stx encoding variant genes stx_1 and stx_2 are located on lambdoid bacteriophages. Stx is an AB-toxin: The five identical B subunits of the holotoxin bind to the target cell surface at the glycolipid globotriaosylceramide (Gb3) receptor. The single A subunit is internalized and cleaves ribosomal RNA, leading to the termination of protein biosynthesis [11]. Stx that is produced in the intestine translocates via the bloodstream to the kidney, where it damages renal endothelial cells resulting in renal inflammation, which can lead to the hemolytic uremic syndrome (HUS). HUS is characterized by hemolytic anemia, thrombocytopenia and acute kidney failure, that is potentially fatal [8]. Some STEC induce A/E lesions on epithelial cells as they possess the same LEE PAI described for EPEC. Enterohemorrhagic *E. coli* (EHEC) are a subgroup of STEC and were originally described in association with hemorrhagic colitis (HC) in humans [19]. In 2011 a large food-borne outbreak in Germany was caused by EHEC with nearly 4,000 infected patients of whom 855 developed HUS resulting in 53 deaths [23].

Further details on the pathotypes EPEC and EHEC are discussed in chapter 1.3 and 1.4 as they are the main focus of this thesis.

1.2 Genetic variability of *Escherichia coli* – Horizontal gene transfer

The adaptability and formation of a variety of different *E. coli* pathotypes is conferred by the diversity of their genomes. The chromosome of the laboratory *E. coli* K-12 strain MG1655 was the first completely sequenced *E. coli* genome [24]. Since then a large amount of commensal and pathogenic *E. coli* strains have been sequenced, with an increasing trend due to the declining costs of whole genome sequencing. The genome sequences of *E. coli* strains demonstrate a high structural and genetic diversity, which is also displayed by the wide size-range of sequenced *E. coli* genomes that can be found within the ncbi genome database (http://www.ncbi.nlm.nih.gov/genome) that varies from 4.56 to over 6.0 Mb. The combined, non-redundant set of proteins encoded within the genomes of MG1655 and the initially available genomes of an UPEC and an EHEC strain revealed that only 39.2% (2,996 genes) were common to all three [25]. Furthermore it has been confirmed that only roughly 20% of any given *E. coli* genome will be part of the *E. coli* core, and the remaining 80% are not found in other *E. coli* genomes [26]. In the recent genome research the bacterial genome most often is subdivided into core genome and pan-genome. The core genome sums up only genes that are present in all of the sequenced *E. coli* genomes published up to the present date, and hence are interpreted as genes conserved in all strains of this species. The pan-genome condenses both the core genome, as well as the accessory genome, which is thought to reflect life style adaptation of strains composing a species. This results in a slightly decreasing core genome when more genomes are analyzed, because some previously identified core genes are absent in newly sequenced strains; in contrast, the pan-genome increases as new genes are identified [27]. These facts imply that a critical amount of DNA had been acquired by those *E. coli* strains from other biological sources. These genome sequence-changes influence the pathogenic potential, the metabolic and the phenotypic features of *E. coli* strains and are achieved by either inherent replication process errors trough point mutations, genomic rearrangements, and mobility of insertion sequences or the acquisition of exogenous DNA via horizontal gene transfer (HGT) leading to the transmission of modified genomes from one generation to the next [28]. These genome alterations occur upon selection in response of environmental conditions. The evolution of genomes includes ‘macroevolution’, the development of new species or subspecies over millions of years, and ‘microevolution’, the alteration of genes or traits within short time periods [29]. A bacterial genome comprises a core genome of stable regions and a flexible gene pool composed by homologous recombination, transduction with bacteriophages, acquisition of plasmids and transposons, as well as the accommodation of genomic islands (GEIs), larger genomic regions that include PAIs. Therefore HGT describes the process by which bacteria gain an accessory genome. Bacterial genomes can also be ‘downsized’ due to loss of mobile genetic regions and genomic reduction, as optimization result due to changes of environmental conditions, including the conditions present in the host the bacteria is colonizing or infecting, leading to an adaptation of the bacterial population to these environmental conditions. The main principles of HGT are transformation, transduction and conjugation [30].

Natural transformation is the active uptake of free extracellular DNA that was released from a living or a lysed cell by a bacterial recipient cell, and is then incorporated into the bacterial chromosome. To be competent the recipient cell must be in a physiologically active state and activate a signal cascade of peptides and proteins in order to express the ‘late competence genes’. They encode for proteins

involved in DNA uptake and the following recombination step to integrate the new DNA into the chromosome [31].

Conjugation is the transfer of bacterial DNA through cell-to-cell contact. The donor cell must harbor a conjugative element such as a transposon or plasmid. The donor produces thread-like structures on the cell surface, so-called pili that are processed and excreted by a type four secretion system. These pili are used to contact a recipient cell and to gain close contact between their cell surfaces to form a pore complex that links the two cells. The plasmid DNA is separated into two single-strands of which the non-circular strand is transferred through the pore. When the ssDNA enters the recipient, it is used as a template for lagging strand DNA synthesis resulting in the reconstituted plasmid. The still circular ssDNA in the donor is used for leading strand DNA synthesis to maintain the plasmid [30, 32].

Phage transduction and prophage integration are the major mechanisms of HGT in bacteria and is further exemplary explained for lambdoid bacteriophage. Bacteriophages are viruses that infect bacteria and move DNA from one bacterial cell to another as a vector. In generalized transduction the phage is adsorbed to the bacterial surface and injects its DNA. The phage genome is replicated by a rolling circle mechanism, producing concatemers which are packed into new phage heads. When one phage head is fully packed the DNA is cleaved by phage encoded proteins at specific sequence sections called *pac* site. Similar sites, 'pseudo-*pac*', sites are also present in the *E. coli* genome and can be recognized and processed by the phage. When the phage accidently packs chromosomal host DNA they are called 'transducing particles' and are now able to inject this double-stranded segment of *E. coli* DNA into another host cell. Here it will be either degraded or used as a template in homologous recombination and in conclusion is integrated into the chromosome. During infection 0.3% of the phage particles produced will contain chromosomal host DNA instead of phage DNA [30]. For specialized transduction phage λ (lambda) is the most studied bacteriophage. Lambdoid bacteriophages can either undergo a lysogenic or a lytic life-cycle. In the lysogenic phase the phage genome is integrated into the *E. coli* chromosome via site-specific recombination and its DNA is replicated as part of the host genome. A bacterium that harbors a prophage is called a lysogen. The lytic cycle of a phage can be induced by cellular stress due to antibiotic substances, starvation, or exposure with UV-light. Such influences lead to a DNA-repair mechanism in the bacterial cell. The phage DNA is excised from its bacterial host genome, replicates autonomically, and is packed into new infective phage particles, which are released by lysis of the bacterial host cell [33]. When the phage is incorrectly excised from the chromosome it can abduct genes of the host DNA that flank the insertion site. Prophages might also carry additional genes that are not required for the phages' life-cycle. Those genes are called lysogenic conversion genes, or morons, and are often the virulence associated genes of pathogenic bacteria. Therefore the fitness or phenotype of the lysogen can be altered, leading to the emergence of new pathogens [34].

Recombination is the major DNA repair pathway and also key factor in bacterial evolution which allows the bacteria to increase their virulence features, utilize new energy sources, circumvent the immune response of an invaded host, and acquire antibiotic resistances [35]. The integration of DNA fragments into the bacterial genome requires regions between 50 to 100 bp in length of high similarity to the host chromosome to undergo homologous recombination. The incoming DNA fragments of transduction are linear and double stranded. A trimeric complex RecBCD binds to each double stranded end and

resects the DNA. While the complex translocates along the dsDNA it unwinds the double-strand. The 3' strand is degraded by the nuclease RecB and the 5' strand occasionally cleaved. When the complex recognizes a certain 8 bp short sequence, the strand cleavage preference is changed, leading to the production of a short 3'-5' single-stranded tail that contains RecA every three nucleotides. RecA detects homologous sequences that match the single-strand to pair it with a complementary strand and generate D-loops. The D-loop is extended from the invading 3' end forming a four-way junction (holliday junction). The holliday junction is moved along the duplex and finally cleaved, which leads to the resolvation of the holliday junction resulting in recombination molecules. That can be either crossover or non-crossover products [30].

In site-specific recombination the required homology of two sequences that undergo recombination is relatively small. Site-specific recombination is used for the integration and excision of bacteriophages and PAI. The integrase of phage λ is a recombinase that uses tyrosine as a nucleophile to cleave the bacterial DNA backbone. The tyrosine recombinase recognizes two specific sites in the bacterial chromosome: the phage attachment site *attP* and the bacterial attachment site *attB*. A complex that includes the integrase and the host integration factor is formed the *attP* site before *attB* is bound and the DNA backbone is attacked by the catalytic tyrosine. The covalent binding of each 3' end to an integrase promotor results in a free hydroxyl-group at the 5' end that attacks the DNA-protein bound of the other *att* site, leading to the first strand exchange and formation of a holliday junction. The holliday junction is resolved by a second pair of strand cleavages leading to the second strand exchange and the formation of recombinant products [30, 36]. Site specific recombination generates many different variants and therefore promotes evolution not only by introducing new DNA fragments to the bacterial chromosome, but also by the occasional move or rearrangement of adjunct DNA sequences of the bacterial genome while shifting themselves. They can carry these sequences to another site or cause deletions of neighboring DNA sequences [37].

1.3 Characteristics and prevalence of EHEC

The first two identified outbreaks of EHEC were detected in 1982 in the US states Michigan and Oregon and were associated with watery diarrhea that developed to bloody diarrhea and abdominal cramping. The serotype of the causative EHEC strain was O157:H7, which is now the most common serotype isolated from patients with bloody and non-bloody diarrhea in the US. The outbreak was linked to contaminated undercooked beef patties of a fast-food restaurant chain. In 2014 in the US 4,756 cases of EHEC infections were reported by the Centers for Disease Control and Prevention (CDC) [38] and 1,650 cases were reported by the Robert Koch Institute (RKI) in Germany, of which 29.2% were hospitalized [39]. The infections in Germany included 85 reported cases of HUS. *E. coli* O157:H7 is the leading cause of HUS in the US, Canada and Europe. HUS is defined as a combination of acute renal failure, thrombocytopenia, and microangiopathic hemolytic anemia and is potentially fatal [40]. There is epidemiological distinction between the terms 'EHEC' and 'STEC', as EHEC describes Shiga-Toxin producing strains that are associated with a distinctive clinical syndrome, namely hemorrhagic colitis (HC), mainly caused by *E. coli* with serotype O157:H7. STEC can cause a variety of diarrheal diseases that do not fit this description [41]. "Thus, all EHEC are

STEC, but only some STEC are EHEC, and STEC is a more comprehensive term."[1] While humans develop severe diseases being infected with EHEC, cattle, the main reservoir of STEC, does usually not show any symptoms while being colonized with the bacteria [42]. Yet STEC can be responsible for diarrhea in calves [43]. STEC have been furthermore isolated from the feces of other healthy domestic farm animals, mainly ruminants, such as sheep and goat [44]. Non-ruminants, as pigs, can also be symptomless carriers, but are as well known to develop diarrhea or edema disease [45]. However, edema disease in piglets is caused by certain STEC that produce a particular subtype of Stx2, namely Stx2e. In poultry STEC were only rarely isolated from chicken feces [46], but frequently from turkey [47]. In wildlife animals only a few numbers of isolates from feces were positive for STEC, mainly those from deer (*Odocoileus virginianus*) [48]. Also rabbits (*Oryctolagus cuniculus*) are known to carry and transmit STEC [49]. In general STEC can be isolated from the feces and gut of a variety of animals in the farm environment, like synanthropic rodents and birds, and domestic animals [2, 50]. This large reservoir or infection source respectively of STEC carriers and propagators constitute a potential risk for humans, due to the fact that EHEC are zoonotic agents. EHEC are transmitted from animal feces to the environment, where water and soil can be contaminated [51, 52]. Another important source of infection are inadequately processed foods of animal origin, e.g. undercooked ground beef [53] or unpasteurized dairy products [54]. The transmission to humans takes place via the fecal-oral infection route and therefore causes a wide range of food-borne outbreaks. Beside the O157:H7 serotype, that caused several distinct outbreaks in North America, Southern America, Europe, Africa, China and Japan and sporadic cases worldwide [55-60], more than 150 non-O157 serotypes were also connected with diarrheal diseases in humans. The most frequently isolated of those non-O157 STEC are the O-serogroups O26, O111, O103 and O145, which are equally connected with severe human disease, ranging from watery diarrhea to HC, HUS and death [61, 62]. Nevertheless patients with O157 STEC infections are more likely hospitalized than those with non-O157 infections. Furthermore O157 STEC infections result more often in HUS and are more often related with outbreaks than non-O157 infections, which are more often associated with international travel [63].

Non-O157 and O157 STEC share equal pathogenic factors of which the bacteriophage encoded Shiga-Toxin is the most relevant. A STEC strain may either express Stx1 or Stx2, both of them in combination, or even encode for multiple *stx* genes. Both toxins show a 55–60% amino acid identity [64]. The nucleotide sequence of stx_1 is subtyped into stx_{1a}, stx_{1c}, and stx_{1d}, while stx_2 is subtyped as stx_{2a} to stx_{2g}. Some of those subtypes are related to specific disease [65], e.g. stx_{2e}, as already mentioned, is known to cause edema disease in pigs. Stx1 and Stx2 show the same polypeptide structure of five identical 7.7 kDa B subunits to one 32 kDa A subunit in the mature holotoxin, the same enzymatic activity as N-glycosidases and the subsequent specificity for a certain glycolipid receptor. After binding of the B pentamer to the globotriaosylceramide or Gb3 receptor, it enters the mammalian cell through endocytosis. Here it is transported from the endosome to the Golgi before it enters the endoplasmatic reticulum where the A subunit dissociates. The A subunit is translocated to

[1] cited from Keusch, G.T. and M.B. Skirrow, *Enterobacteria, campylobacter, and miscellaneous food-poisoning bacteria*, in *Oxford Textbook of Medicine*, D.A. Warrell, T.M. Cox, and J.D. Firth, Editors. 2005, Oxford University Press: Oxford. p. 489-503

the cytoplasm and targets the 60s ribosomal subunit. The N-glycosidase activity of the A subunit removes a single adenine from the 28S rRNA and thereby inhibits protein synthesis resulting in cell death [23].

Besides Stx numerous STEC strains also harbor the Locus of Enterocyte Effacement (LEE) PAI that causes attaching and effacing (A/E) lesions on epithelial cells

as a result of colonization of the intestinal tract (Chapter 1.1v). The LEE core region of approximately 34 kb encodes a Type Three Secretion System (TTSS), an outer membrane adhesin called intimin, its translocated intimin receptor Tir, and several proteins for signal transduction. The size of the LEE can vary from 36 to 111 kb [66] and consists of 41 ORFs that are organized in five operons: *LEE1- LEE5*. *LEE1* encodes for the LEE encoded for the regulator (Ler), a transcriptional regulator for the LEE region. *LEE1*, *LEE2*, and *LEE3* harbor *esc* genes, that are also named *sep* genes, and encode for major components of the TTSS. On operon *LEE4* the genes *espA*, *espB*, *espD*, *sepL*, *espF*, and *escF* are located*,* which encode for effector proteins that translocate into the host cell via the TTSS. The *LEE5* operon encodes the genes for intimin expression *eae* and the expression of its translocated intimin receptor *tir* [67]. Besides the effector proteins encoded on the LEE, some effectors are also located outside of the LEE, which are called non-LEE encoded (Nle) proteins, that can often be found on genomic O-islands. Nles are essential for the virulence and colonization of A/E pathogens as well as other processes similar to those of the bacterial host cell. These processes can be anti-apoptotic activities, the interruption of host innate immune responses by preventing NF-κB activation at different levels, and the blockage of cell division. An example is the cycle inhibiting factor Cif that arrests the cell cycle at G1/S and G2/M phases by deamidation of ubiquitin [68]. Another virulence factor of STEC is the plasmid-encoded enterohemolysin gene *ehxA* that is >60% identical to the alpha-hemolysin gene of other *E. coli.* Enterohemolysin is a resiniferatoxin (RTX) cytolysin and lyses erythrocytes *in vivo*, as a source for iron due to the release of heme and hemoglobin [11]. STEC can also express the enteroaggregative heat-stable toxin EAST-1 which is encoded by the *astA* gene. EAST-1 activates the production of cGMP in intestinal epithelial cells, which serves as a mediator of the biological activity of EAST-1. Yet the role of EAST-1 is still not fully revealed, but its hypothesized to play a role in the pathogenesis of infections caused by *E. coli* [69]. Additionally STEC harbor numbers of other virulence factors that are associated with adhesion, like fimbrial proteins e.g. type 1 fimbria and long polar fimbria, and adherence e.g. Iha and OmpA [70].

1.4 Comparison between enterohemorrhagic and enteropathogenic *E. coli*

The diarrheagenic *E. coli* EPEC and STEC share a variety of virulence factors, most importantly the EPEC pathotype defining LEE PAI. Therefore both pathotypes display the characteristic formation of A/E lesions as their mechanism of intestinal colonization. The large number of *nle* effector genes is also a common to both pathotypes. The pathotype EPEC is distinguished into typical and atypical EPEC based on the possession of the EAF plasmid by typical EPEC, and the encoded gene *bfp* facilitating the expression of a bundle forming pilus. The EAF plasmid is not harbored by aEPEC, hence the adherence of the bacteria to host cells is more loose, resulting in a ‘localized adherence-like’ patter. STEC also do not possess the EAF plasmid. Typical and atypical EPEC are two distinct

groups of pathogens, with aEPEC being more closely related to LEE positive STEC, regarding virulence properties, genetic features, serotypes, and reservoirs – with the notable exception of Stx production. While typical EPEC are the leading cause of infantile diarrhea in developing countries, atypical EPEC were mainly associated with infections in the industrialized nations for many years. However recent data suggest that infections with aEPEC exceed those with typical EPEC in developing as well as developed countries. For example a study performed in Iran of EPEC strains isolated from children with diarrhea found 39.3 % *bfp* positive EPEC strains, hence typical EPEC, and 61.7 % lacked the *bfp* gene and were identified as aEPEC strains [71]. Comparable results were found in studies in Brazil and Thailand, and furthermore studies from Australia and Norway isolated aEPEC strainss more commonly from patients with persistent diarrhea than typical EPEC [72].

Typical EPEC have serotypes like O55:H6/NM, O86:H34, O111:H2/NM, O114:H2, O127:H6/H40, O142:H6/H34 [72]. Atypical EPEC serogroups are, besides the most frequent serotype O51, and some lineages which have evolved from typical EPEC strains due to loss of the plasmid encoded virulence genes, are even more diverse than typical EPEC and share the most important O-groups, O157, O26, O103, O111, and O145, with typical EHEC [73]. Atypical EPEC of those serogroups additionally share the same virulence plasmid encoded genes, *tir*-genotypes and the already mentioned LEE and non-LEE encoded genes [74]. While typical EPEC are only isolated from humans, the reservoir of aEPEC includes humans as well as animals. Infections of calves with aEPEC of serotype O26:H11 are well studied [75]. Atypical EPEC of serotype O128:H2 was found in rabbits and dogs, which were also tested positively for aEPEC of serotypes O119:H2 and O111:H25 in other studies [76].

The classification of both pathotypes, aEPEC and STEC, merely depends on the presence or absence of *stx* genes. As the *stx* genes are encoded on transmissible lambdoid phages it is possible that EHEC, while infecting a human host can lose the *stx* gene. This is possible when the lytic phase of the bacteriophage is induced due to mammalian host signals i.e. a hydrogen peroxide increase. The bacterial host cells that harbor the *stx*-converting bacteriophage are vulnerable for lysis through the phage and might discard the phage in to order survive. Such *stx* negative derivatives of original infecting EHEC that lost their *stx*-phage during human infection have been isolated from HUS-patients. The serotypes of that so-called EHEC-LST (EHEC that lost the *stx*-phage) were O157:H7, as the most frequent, followed by O26:H11/NM, O145:H28/NM, and O103:H2/NM [77]. In a follow-up study of HUS patients, the molecular characteristics of *stx* positive and *stx* negative O157:NM and O26:H11/NM strains isolated from stool samples have been found to be closely related [78]. The loss of the *stx*-phage in EHEC O26 was also confirmed with *in vitro* studies [79]. On the other hand is has been demonstrated that O26 aEPEC can be transduced with *stx*-phages of O26 EHEC forming *stx*-producing stable lysogens [79].

The acquisition and loss of *stx*-phages marks aEPEC either as EHEC-progenitors or EHEC-LST. Accordingly EHEC and aEPEC of non-O157 serotypes represent a dynamic system of bidirectional conversion.

1.5 Typing methods

Typing methods are used to discriminate between different bacterial strains of the same species and are important tools in epidemiology for infection control and prevention. Traditional typing methods focusing on the phenotype of bacterial isolates have been used for many years; such typing methods are among others serotyping, phage-typing, biotyping, or antibiogram. Nowadays isolates are typed at a molecular level to analyze the relatedness of bacterial isolates. These molecular methods are either based on nucleotide sequences of certain genes, i.e. alleles of seven housekeeping genes used for Multilocus Sequence Typing (MLST), or analysis of whole genome sequences. The development of benchtop sequencers using next generation sequencing technologies makes whole genome sequencing practical and achievable for research and clinical laboratories [80]. In this thesis the results of the traditional serotyping of *E. coli* strains are compared with the results of molecular typing methods, such as MLST of seven *E. coli* housekeeping genes, and those of a single nucleotide polymorphism analysis based on whole genome sequences. Further the whole genome sequences of the examined *E. coli* strains are analyzed for the presence of virulence associated genes and integration of mobile genetic elements in certain position of their genomes. The order of the following methods described in detail reassembles the order of the presented results (Chapter 3).

1.5.1 Serotyping

E. coli serotypes are a specific combination of O and H-antigens. The antigens of the cell wall (O), and flagellar (H) are used as a classification method. In the past the capsular (K) antigen was additionally typed. The LPS consists of three parts, the lipid A, the core oligosaccharide and the O-poly saccharides (O-antigens). The O-antigen domain presents a notable diversity in structure, because of the variation in combination, position, stereochemistry, and links between the repeating units of sugar residues, along with the absence or presence of acetyl groups. Due to mutation the O-antigen can be lost and the smooth cell membrane becomes “rough” influencing also the saline agglutination as well as phenotype of the colonies and may additionally change the virulence of the bacteria. As not many laboratories are capable to type the K-antigens, serotyping based on O- and H-antigens became the “gold standard” for *E. coli*. Nowadays O-groups O1 to O187 have been defined. The O-serogroups are detected with an agglutination reaction of O-antigen and antisera that is either commercially purchased or prepared in-house [81, 82]. For flagellin, the H-groups H1 to H56 have been described with H-antigen negative (NM) for the non-motile ones. The detection can also be done by serological agglutination reaction with monovalent antisera or a PCR-restriction length polymorphism (RFLP) amplification of the gene *fliC* that encodes for the flagellar [81, 83]. Serotyping was and is still applied to distinguish strains of public health importance or exceptional virulence, before whole genome sequencing became the method of choice to analyze important strains in more depth. Serotyping is a labor intensive method and can be prone to error due to cross-reactivity between the adsorbed O-antigen antisera that is produced in rabbits. Moreover some strains remain not typeable due to auto-agglutination or a rough phenotype [83]. Regarding the non-O157 merely serotyping would not be sufficient, as not all non-O157 have the capacity to cause severe disease in humans, or even a

human infection at all. In conclusion a molecular genetic approach to identify public-high risk non-O157 and to reveal the genetic basis underlying the varied clinical spectrum of disease caused by these strains is needed [84].

1.5.2 Multilocus Sequence Typing (MLST)

MLST is a molecular typing technique originally based on the PCR amplification and sequence analysis of internal fragments of bacterial housekeeping-genes and was first described by Maiden *et al.* (1998) [85]. Since then it has become the “gold standard” for phylogenetic analysis of bacteria. Housekeeping-genes were chosen for this typing method, because they are ubiquitous within the population, and it is assumed that recombination occurs with a much higher frequency than point mutations. Therefore to gain a phylogenetic overview of a population only certain genes are considered rather than looking at the total sequence similarities. The recent protocol for amplification of seven *E. coli* housekeeping-genes and moreover a public available database of sequences from these gene fragments was published by Wirth *et al.* [86]. The housekeeping-genes used are *adk* (adenylate kinase), *fumC* (fumarate hydratase), *gyrB* (DNA gyrase), *icd* (isocitrate/ isopropylmalate dehydrogenase), *mdh* (malate dehydrogenase), *purA* (adenylosuccinate dehydrogenase) and *recA* (ATP/ GTP binding motif). The gained sequence for each fragment is compared to already known sequences in the database and a unique number is given for each allele. When all seven housekeeping-genes are analyzed the strain is characterized by a profile of seven allele numbers. This unique combination of seven allele numbers is designated as a sequence type (ST) with a certain number, i.e. ST1. STs are organized into sequence type complexes (STCs) according the eBurst algorithm [87]: a ST Complex is a set of STs that are all believed to be descended from the founding ST; a STC includes at least three STs and those STs differ from the founding ST in only one of the seven alleles, and are therefore single locus variants (SLVs) of the founding ST. STs that differ in two alleles from the founding ST, but only one from its nearest neighbor ST, and hence are double locus variants of the founding ST, are linked to the SLVs of the founding ST. The cluster of all linked STs is then represented by the STC. STs that do not match these criteria are simply referred to by their ST designation. MLST is used for routine typing as well as evolutionary analysis, as it is sufficient for subgrouping a large population of samples into clusters utilizing only seven loci. Differences and similarities in host specificity and virulence potential in human infections can be revealed by the phylogenetic grouping into discrete STCs.

When compared to serotyping it has been shown that strains of unrelated STs share identical serotypes, which reflects the effect ‘serotype switching’, the HGT of O-antigen encoding genes [88]. Different pathotypes of *E. coli* also cluster together in STs. EHEC and EPEC were found in various distinct STCs, but also within a single ST. EHEC strains were revealed to be nearly as closely related to EPEC strains, as they were to other EHEC. This is especially true for EHEC and EPEC strains of the ‘Diarrheagenic *E. coli* (DEC) Collection’ and for EHEC strains of the ‘HUS-associated EHEC (HUSEC) Collection’ that clustered into equal STs within the STC29 [62, 86].

1.5.3 Single Nucleotide Polymorphism (SNP) analysis

A SNP is a variation in a single nucleotide of the DNA and marks a single base change in a sequence. In this position of the genome one individual or bacterial strain has one nucleotide, e.g. A, and another one has got a different nucleotide, e.g. T. Although in a sequence at each position one of the four bases could be potentially present, SNPs are usually bi-allelic. The reason for this is the low frequency of single nucleotide substitutions at the origin of SNPs and therefore the low probability of two independent base changes in a single position. A positon is considered a SNP if the least frequent allele still occurs with the frequency 1% or more [89, 90]. In protein-coding regions SNPs can be either synonymous, which describes SNPs that have different alleles encoding for the same amino acid, or non-synonymous SNPs, that have alleles encoding for different amino acids [91]. SNPs can be used as markers to calculate the genetic distance and thus access the phylogeny of bacterial strains. Since whole genome sequencing techniques became more popular and benchtop platforms for laboratory use were developed, the costs per bacterial whole genome sequence dropped to an achievable level for research purposes. Hence the SNP-analysis can now be done on whole genome level rather than only on single genes. It already was used to identify nucleotide polymorphisms for high-resolution evolutionary analysis of the STEC O157:H7 genetic subtypes to detect those SNPs associated with either cattle reservoir or increased virulence in human infections [92]. Moreover the derivation of EHEC O157:H7 from EPEC O55:H7 (described in Chapter 1.6) was analyzed in greater depth based on comparison of whole genome sequences. Recombination events, insertions, deletions and mutations (synonymous and non-synonymous SNPs) were identified to create a detailed evolutionary model, of with also the divergence time could be estimated using mutational synonymous SNPs [93].

1.5.4 Presence of Virulence Associated Genes (VAGs) and insertion sites of mobile genetic elements

In contrast to other non-living ecological niches bacteria had to adapt to the mammalian host and its defense mechanisms. These certainly developed alongside with the bacteria, and include among others physical barriers, like mucus-covered epithelia, the production of antimicrobial peptides, iron-secretion mechanisms, and immune responses. As reply bacteria established certain mechanisms and factors to bypass these defenses, the so called virulence factors. Virulence factors enable the bacteria to invade the host and engage specific host cells. The host cells can be altered and/or destroyed, leading to disease. Virulence factors also can evade or neutralize the host defenses. The interaction with the host starts with the localization of an entry site, aiming at a place to multiply in the host body, to finally become persistent in this host or to find ways to reach the next host. Transmission to the next host and multiplication or survival in the environment are crucial for the success of a pathogen [34, 94]. The various genes that encode for virulence factors, virulence associated genes (VAGs), are used to determine the pathotype of *E. coli* (described in Chapter 1.1). Besides the virulence profile of an *E. coli* strain they can also be used to develop evolutionary models of emerging pathogens and describe the phylogeny of a bacterial population by comparing the presence and absence of VAGs of the individual bacteria within a population [95, 96]. Bacteria often acquire VAGs

by HGT of mobile genetic elements. These insert into the bacterial chromosome via recombination that takes place in specific sites of the genome. Characteristically sites for a high recombination rate display a high GC-content and are frequently associated with DNA regions encoding for tRNA [97]. For *E. coli* O157:H7 the *stx*-converting bacteriophage integration sites are fairly conserved and typically limited to four loci (*wrbA*, *yehV*, *argW* and *sbcB*). The *stx*-bacteriophage insertion (SBI) is also used as a typing method in characterization studies investigating *E. coli* O157:H7, but not for non-O157 where the insertion sites of bacteriophages are more varied and less specific [98]. The reason for this is the monophyletic nature of O157:H7 EHEC, while most non-O157 *E. coli* are polyphyletic and only some other O serogroups are represented as a closely related lineage as O157:H7 EHEC (for detailed explanation see chapter 1.6). Nevertheless over 60 insertion sites in the *E. coli* genome are known to be possible regions for recombination and integration of bacteriophages, genomic islands, transposons, and insertion elements [99]. The occupation of those insertion sites by mobile genetic elements could be used in a profile manner to compare relatedness of *E. coli* strains and develop a microevolutionary model how the acquisition of specific mobile genetic elements might lead to the formation of a certain lineage.

1.6 Evolution of O157 EHEC

EHEC of serotype O157:H7 are predominant regarding public health, causing large food-borne outbreaks of intestinal illness. Hence it is the best characterized and studied group of EHEC and evolution and emergence of this pathogen is already described within various publications using different approaches. The evolutionary model that was developed for EHEC O157:H7 is described in detail in the following chapter to give an overview of the bacterial evolution of a certain pathogenic lineage regarding the aim of this thesis to gain insights into the evolution of non-O157 EHEC and the potential development of a comparable model for those pathogens.

Strains of serotype O157:H7 are not closely related to STEC strains of other serotypes, further they are only distinctly or not at all related to strains of O157 in general, that were isolated from animals with enteric disease, as strains of O-serogroup O157 own a broad genomic variability [100, 101]. Hence the O157:H7 serotype is evolutionary a distinct clone. Multilocus enzyme phenotyping revealed that the O157:H7 clone is closely related to a group of O55 EPEC assuming that the O157:H7 clone derived from an O55:H7-like ancestral clone that is already accustomed to trigger diarrheal disease [102]. On basis of multilocus enzyme electrophoresis of 20 gene loci in various O157:H7/NM strains and other O:H-types Feng *et al.* postulated a phylogenetic model about the history of origin and the order of acquisition of virulence genes in the virulent O157:H7 serotype. The model starts with an EPEC-like O55:H7 ancestor that is able to express β-glucuronidase (GUD+) and ferment sorbitol (SOR+), but already has the typical *uidA* -10 mutation (A→T) in the gene for GUD expression, which also is typical for O157:H7/NM strains. This O55:H7 is an aEPEC that harbors the LEE PAI. In the next step the O55:H7 ancestor acquires the stx_2-gene, probably due to transduction by an stx_2-converting bacteriophage. In the next stage the second characteristic mutation of GUD, *uidA* +92 (A→G) occurred, and the somatic antigen changed from O55 to O157, resulting in an O157:H7 stx_2-producing ancestor that possesses a GUD+ and SOR+ phenotype. The exchange of the somatic O-

antigen is assumed to be a result of HGT of an *rfbE*-like region. Next, two lineages developed, of which one lost motility and the other lost the ability to ferment sorbitol (SOR-), which is typical for the clonal lineage of O157:H7. Furthermore the SOR- ancestor acquired a *stx1* gene, presumably by conversion by a bacteriophage. The O157:H7 *stx1*+, *stx2*+, SOR- and GUD+ ancestor then lost the β-glucuronidase acitivity (GUD-) leading to the immediate ancestor of the common, globally spread, O157:H7 clone. Except for the O157:H7 *stx2*+, SOR+ and GUD+ ancestral phenotype, all evolutionary phenotypes have been exemplified with actual strains [95]. Genomic sequences of the representing strains were used to refine the model above. About 70 recombination events have been detected since the separation of the two lineages, which lead to SNPs that affect 5% of the genome. Furthermore, about 200 insertions and deletions were noticed, of which the larger ones mostly involved phage genomes. In the O55:H7 genome nine and in the O157:H7 genome 23 phage-related elements were found with only three common to both genomes. Additionally the divergent time of O157:H7 and O55:H7 was calculated on basis of the number of mutational changes, not including regions that have undergone recombination. As a conclusion, both lineages separated 14,000 to 70,000 years ago [93]. It is easier to implement such a model for a distinct clone like EHEC O157:H7 than for the more distantly linked group of non-O157 EHEC. Nevertheless, if a number of closely related, individual strains that are connected with each other through certain features apart from serotype would be identified, a model of the microevolution of these non-O157 EHEC could be built upon determined results of certain typing methods.

1.7 Aims of the thesis

The emerging pathogen enterohemorrhagic *E. coli* (EHEC) is a cause of bloody diarrhea in humans as well as extraintestinal sequelae, most importantly the hemolytic uremic syndrome (HUS) and thrombotic thrombocytopenic purpura (TTP). EHEC are toxigenic intestinal pathogenic bacteria with a low infection dose that led to several food-borne outbreaks [40]. O157:H7 is the most frequent EHEC serotype causing clinical illness initiated by the consumption of contaminated food products. However also several non-O157 serotypes are important causes of diarrheal disease producing clinical symptoms of similar severity to those of O157:H7. The O-serogroups O26, O103, O111, and O145 are the most frequently isolated, causing one third of the non-O157:H7 infections in humans in Germany [39, 61]. The classification of EHEC strains into separate serotypes implies a phylogenetic diversity and characteristic differences within the group of non-O157. The common characteristic feature of EHEC is the production of the bacteriophage encoded Shiga-Toxin. Additionally, they cause ‘attaching and effacing’ (A/E) lesions on epithelial cells, which are induced by a type three secretion system (TTSS) that is encoded on a pathogenicity island (PAI) called locus of enterocyte effacement (LEE) [66]. The independent acquisition of mobile genetic elements, as bacteriophages and genomic islands, via horizontal gene transfer (HGT) is the primary driving force for a parallel evolution that led to different EHEC phylogenies. In comparison, non-O157 aEPEC are like-wise a common cause of non-bloody and bloody diarrhea, and even can be isolated from patients with HUS [79]. Like EHEC, aEPEC possess the PAI LEE and do not harbor the EAF plasmid, the characteristic feature of typical EPEC [11]. It is assumed that aEPEC were initially EHEC that have lost the *stx*-converting

bacteriophage during human infections, and hence are called EHEC-LST [78]. Yet, aEPEC may be able to convert back to EHEC on acquisition of the *stx* genes. In conclusion, EHEC and aEPEC represent a dynamic system of bidirectional conversion.

The aim of this thesis was to determine the underlying evolutionary background of the most important non-O157 EHEC serotypes in order to develop a microevolutionary model. To achieve this aim the following objectives were conducted:

(I) MLST, as 'gold standard' typing method for *E. coli*, was used to get initial insights into the phylogeny of non-O157 *E. coli* strains.

(II) An additional set of non-O157 aEPEC was also examined with MLST to detect possible common clusters of non-O157 EHEC and aEPEC.

(III) The determined allele based results were used to define a set of closely related *E. coli* strains for further whole genome sequence (WGS) analysis.

(IV) WGS-based SNP-analysis was used to distinguish the phylogenetic relation of these non-O157 EHEC and aEPEC strains to a high level of resolution.

(V) Additionally, the virulence profile of these strains was determined and compared to ideate further insights in the diversity of these non-O157 EHEC and aEPEC strains. Furthermore the integration sites of mobile genetic elements of strains that represent distinct lineages or splitting points within this population were analyzed.

(VI) Finally, based on the summarized results of all analyses conducted with the non-O157 aEPEC and STEC/ EHEC population of STC29, a microevolutionary model of O26 and O111 EHEC was developed.

2 BACTERIAL STRAINS AND METHODS

2.1 Bacterial strains

Different compilations of *E. coli* strains are used in the present thesis. An overview of the strain sets used for different analyzing steps during this study and the respective work flow are schematically displayed in Figure 2.

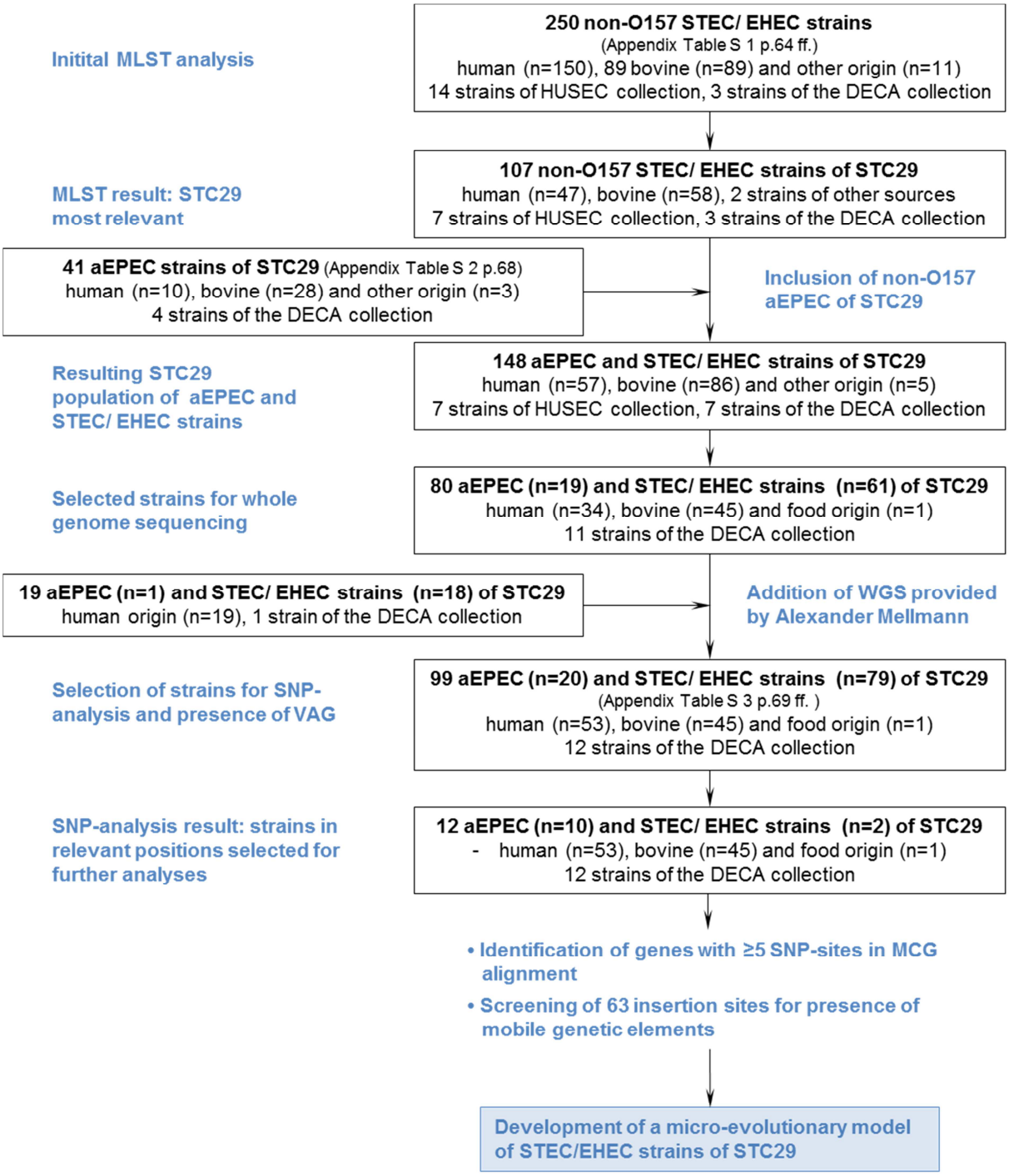

Figure 2 Schematical overview of sets of *E. coli* strains used in the present thesis. Boxes with black font identify the number of *E. coli* strains used in a certain analysis step, indicating also the host species and possible inclusion of strains of published *E. coli* collections (HUSEC collection, DECA collection). The additional text in blue font states the analysis performed with the respective *E. coli* strain set.

2.1.1 non-O157 strains of the MLST Analysis

A total of 250 *Escherichia coli* isolates were investigated (Table S 1 in the Appendix p.64 ff.). Most of the bovine *E. coli* strains of O serogroups O26, O103, O111, and O145 were chosen from strain collections of the Institute of Microbiology and Epizootics (IMT), Freie Universität, Berlin, Berlin, Germany, as well as the *E. coli* MLST database (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) [86]. The human strains of the relevant non-O157 O serogroups were isolated within an exploratory clinical study of the "FBI-Zoo" research consortium (www.fbi-zoo.de). In that study, stool samples of patients with diarrhea were screened for STEC/ EHEC and for *Yersinia*, *Salmonella*, and *Campylobacter spp*. at the University Hospital Münster, the Hannover Medical School, and the Max von Pettenkofer-Institute for Hygiene and Medical Microbiology (Ludwig-Maximilians-Universität München). In addition, 14 strains of the HUS-associated EHEC (HUSEC) collection [62] assigned to the non-O157 O serogroups O26, O103, O111, and O145 were included. Furthermore three strains of the diarrheagenic *E. coli* (DECA) collection (http://www.shigatox.net/stec/cgi-bin/deca) were included. The DECA collection is composed of 78 *E. coli* strains of different serotypes, including O serogroups O26 and O111, isolated from different hosts with diarrheal symptoms, mainly humans, and were sampled worldwide. Overall the greater part of the 250 *E. coli* strains under investigation were sampled in Germany. Strains obtained from other *E. coli* studies, such as studies conducted in Australia, Canada, the United States, and European countries such as Sweden, the United Kingdom, and Belgium, were also included in the study. The strains under investigation had been sampled between the years 1952 and 2009. The strains were isolated mainly from cattle (n=89) and humans (n=150), but strains from camel (n=2), and food source (n=4), as well as single strains isolated from sheep, cat, wild rabbit and water were also included in the analysis. After serotyping, 77 strains were assigned to O serogroup O26 (NM, H11, H19), 96 strains to O103:(NM, H2, H3, H11, H18, H21, H25, H31, H43), 35 strains to O111:(NM, H2, H8, H10), and 42 strains to O145:(NM, H18, H25, H28, H34). Furthermore, 41 aEPEC strains, of human (n=10), bovine (n=28), and ovine (n=2) origin, as well as one strain isolated from a food source (n=1), were included in our analysis. The aEPEC strains belonged to the non-O157 O serogroups and clustered within STs of STC29 (see Table S 2 in Appendix p.68).

2.1.2 Strains of STC29 selected for whole genome sequencing

The MLST analysis of 250 STEC/ EHEC strains resulted in an accumulation of strains in STC29, hence 41 aEPEC of STC29 were additionally included in this analysis. The 148 strains that clustered into STC29 represented the basis for strains further chosen for whole genome sequencing and following analyses; thus 68 *E. coli* strains of the pathotypes aEPEC (n=12) and STEC/ EHEC (n=54) of this set were chosen for whole genome sequencing. In addition 19 whole genome sequences of O26 *E. coli* strains isolated from humans and MLST typed as representatives of STC29, were provided by Alexander Mellmann (Institute for Hygiene, University Hospital Münster, Germany) and included into the SNP-analysis. These genomes included 18 STEC/ EHEC and one aEPEC strain; one EHEC strain was a member of the DECA collection. Furthermore whole genome sequences of eleven strains of the DECA collection, designated as STs of STC29, were included for following analyses. The DECA collection was sequenced by Rasko *et al.* in 2010 as part of the National Institute

of Allergy and Infectious Disease (NIAID) Genome Sequencing Center for Infectious Diseases (GSCID) project (http://gscid.igs.umaryland.edu/wp.php?wp=emerging_diarrheal_pathogens). As a result 99 whole genome sequences of STC29 *E. coli* strains of bovine (n=45) and human origin (n=53), and one of food source were used for SNP-analysis and following analyses (Appendix Table S 3, p.69 ff.).

2.2 Methods

2.2.1 Serotyping

Serotyping was performed at the Robert-Koch Institute in Wernigerode by Angelika Fruth using a microtiter method with antisera against *E. coli* O-antigens 1 to 182 and against H-antigens 1 to 56 according to the method of Prager *et al.* [83].

2.2.2 Pathotyping

Apart from the clinical data, the pathotypes of individual strains were determined on the basis of the presence or absence of the virulence genes stx_1, stx_2, *escV*, and *bfpB*, which were identified by PCR using published primer pairs and multiplex PCR protocols [103]. LEE-positive strains (aEPEC, STEC, EHEC) were identified by the detection of *escV*, a LEE-located translocator gene. In addition, we also screened for the intimin-encoding gene *eae* to confirm the presence of the LEE [104]. aEPEC strains do not harbor the EAF plasmid and therefore are *bfpB* negative. The *stx* genes of strains positive for any stx gene were subtyped according to the work of Scheutz *et al.* (2012) by using the published primers and PCR protocols [65].

2.2.3 Multilocus sequence typing (MLST)

Multilocus sequence typing (MLST) was performed by analyzing internal fragments of seven housekeeping genes (*adk, fumC, gyrB, icd, mdh, purA, recA*) [86]. The alleles and sequence types (STs) were assigned in accordance with the *E. coli* MLST website (http://mlst.warwick.ac.uk/mlst/dbs/Ecoli) using Ridom SeqSphere software (Ridom GmbH, Münster,Germany) in order to generate a database. The genetic relationships between different STs were determined on the basis of the obtained allele profiles using Bionumerics, version 7.5 (Applied Maths, Sint-Martens-Latem, Belgium). Bionumerics was also used to construct minimum spanning trees (MSTrees) using the predefined template ‘MST for categorical data’ as cluster calculation method which is provided in the software and uses a categorical coefficient for the calculation of the similarity matrix, and calculates a standard minimum spanning tree with single and double locus variance priority rule.

2.2.4 Whole genome sequencing

The whole genome sequencing was partly performed at the Wellcome Trust Sanger Institute, Hinxton, UK, using the Illumina TrueSeq library preparation kit (in accordance to the manufacturer's recommendations (Illumina Inc., San Diego, CA, USA)) and 300 bp paired-end sequencing in 96-fold multiplexes was performed on the Illumina HiSeq 2000 (Illumina Inc.) platform. The WGS provided by Alexander Mellmann were sequenced at the Institute for Hygiene, University Hospital Münster, Germany using the Nextera XT chemistry (Illumina Inc.) library preparation for a 100 bp paired-end sequencing run on an Illumina HiScanSQ sequencer in accordance to the manufacturer's recommendations (Illumina Inc.). For whole genome sequencing at the Institute for Microbiology and Epizootics (Veterinary Faculty, Berlin Germany) DNA was isolated from overnight cultures that were incubated 37°C in Luria-Broth (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) using the Epicentre MasterPure™ DNA Purification Kit according the manufacturer's guidelines for cell samples (Epicentre, an Illumina Company, Madison, WI, USA). The libraries were prepared using the Nextera XT library preparation kit (Illumina Inc.) according the manufacturer's recommendations. The 300 bp paired-end sequencing in 30-fold multiplexes was performed on the Illumina MiSeq (Illumina Inc.) platform.
The reads were *de novo* assembled into contigs using CLC Genomics Workbench 7.5 (CLC bio, Denmark).

2.2.5 Determination of the maximum common genome (MCG) and single nucleotide polymorphism (SNP) detection

The genomes of 99 fully sequenced *E. coli* strains were used to determine the maximum common genome (MCG) [105], all genes that were present in each of the 99 genomes of these strains, with at least 70% identity on nucleotide level and a minimum coverage of 90%. This was performed by a hierarchical clustering with usearch [106] and resulted in a set of 3,714 genes as the MCG. After this step the allelic variants of the MCG were extracted from all 99 genomes, aligned individually and concatenated by an in-house BLAST based pipeline. This produced an alignment of 2,649,433 bp for the 99 strains. Within this alignment 11,891 informative sites could be identified and the SNP information of them was used to generate MSTrees as described in Chapter 2.2.3 with Bionumerics, version 7.5. Clusters in the MCG-SNP were determined if the strains that accumulated into one group share more than 90 % of all SNPs. The single clusters should present a clear distinction, respectively more than 1,500 SNPs difference, to the neighboring cluster.

2.2.6 Identification of SNP-locating genes

For strains in important positions of the MSTree, the MCG alignment of two strains was further analyzed in order to identify those genes that differ in more than five SNP-sites between the compared strains and therefore lead to the localization in the respective clusters. The MCG was annotated and prediction of (protein) coding sequences (CDS) was performed by the RAST Server: rapid annotations

using subsystems technology [107]. The annotated alignments of the MCG of selected strains pairs were screened for genes that show more than five SNPs.

Figure 3 shows an example of a gene identified to differ in more than five nucleotide positions in the alignment of the MCG of LH-1 and DEC9D. The SNPs were counted and it was determined whether the SNPs are synonymous or non-synonymous SNPs on basis of the amino acid sequence of the respective MCGs of each strain (Figure 4). Synonymous and non-synonymous SNPs were counted. The respective nucleotide sequence was further used for a blast search against the ncbi nucleotide collection (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using the megablast algorithm to gain highly similar sequences as result in order to confirm the RAST annotation of the respective gene. The megablast default algorithm parameters were used, with match costs of 1, mismatch cost of -2 and linear gap costs. The ncbi database searches were performed between October and December 2015. The resulting gene name and gene product were noted and the gene name further used for a search within the BioCyc Database Collection (http://biocyc.org/) to get information on pathways these genes are involved in, if several genes are involved in the same pathway or if some genes are co-regulated. The whole genome sequenced O26:H11 *E. coli* strain 11368 (NC_013361.1) was used as reference database in the BioCyc Database Collection.

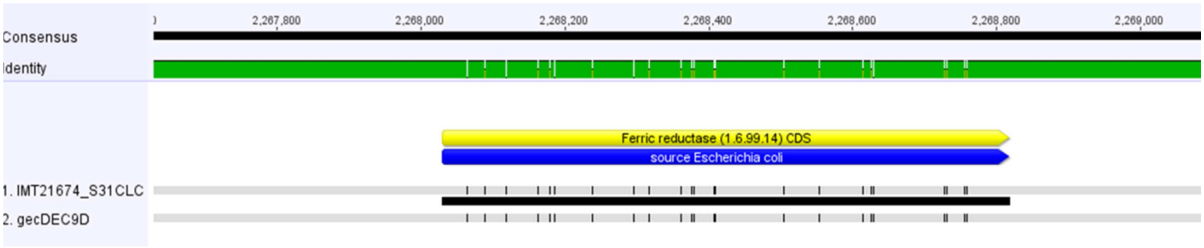

Figure 3 Example for a gene, *fhuF* (Ferric iron reductase), that differs between DEC9D and LH-1 (IMT21674) in more than five SNP-sites (n=15) in the alignment of the MCG of both strains. The black lines indicate the positions of the differences on nucleotide level. Analysis was performed with Geneious 7.0.2.

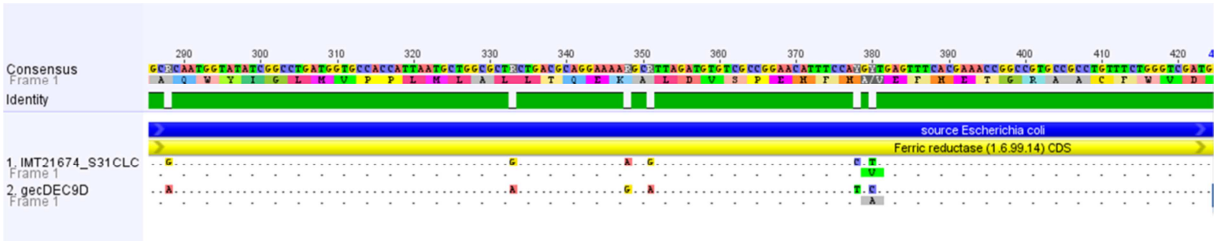

Figure 4 Example for a gene, *fhuF* (Ferric iron reductase), that differs between DEC9D and LH-1 (IMT21674) in more than five SNP-sites (n=15) in the alignment of the MCG of both strains. The gene is the same one as in Figure 3 just in magnified view. Additionally, the amino acid sequence is indicated below the nucleotide sequence, dots indicate agreement between both aligned sequences and a displayed letter (i.e. A or G) is indicating a SNP. The same is true for the amino acid sequence were a dots also indicates agreement between the two sequences and a letter indicates a differing amino acid in both sequences, hence the SNP-including nucleotide-triplet results in a different amino acid and is therefore a non-synonymous SNP.

2.2.7 Screening for virulence associated genes (VAGs)

The 99 whole genome sequences of the *E. coli* strains were screened for 76 VAGs using the online tool 'ViurlenceFinder-1.5' (https://cge.cbs.dtu.dk/services/VirulenceFinder/, [108]). A number of 34 VAGs, that were at least present in one of the analyzed genomes, were taken into account and the presence and absence of each of the gene was noted. Those genes that did not occur in a single *E. coli* genome were not further taken into account. Based on the VAGs a binary virulence-profile was generated and used to construct MSTree in Bionumerics version 7.5 (Applied Maths, Sint-Martens-Latem, Belgium) as already described for MLST (see chapter 2.2.3).

2.2.8 Analysis of insertion sites for mobile genetic elements

A number 63 published insertion sites were chosen to identify insertion of mobile genetic elements in the sequenced strains. Therefore, those genes, namely the insertion sites, were downloaded from the ncbi nucleotide database as genebank files of the *E. coli* O157:H7 str. EDL933 reference strain (NC_002655.2). In Geneious 7.0.2 the insertion sites were assembled against the genome sequence (or respectively the contigs) of each annotated WGS of twelve strains, that were found in prominent positions of the MSTree of the SNP-analysis (see Figure 15, Chapter 3.2.2). The assembly was done using the Geneious Assembler with a high sensitivity method, without any iteration steps. In the next step, the neighboring genes were screened for insertions of mobile genetic elements, which were identified due to an integrase gene next to the insertion site. The integrase and adjunct genes up to the next chromosomally encoded *E. coli* gene were used for a blast search against the ncbi nucleotide collection (http://blast.ncbi.nlm.nih.gov/Blast.cgi) using the megablast algorithm (as described in Chapter 2.2.7). Results of the reference strains O26:H11 *E. coli* strain 11368 (NC_013361.1) and *E. coli* O111:H- strain 11128 (NC_013364.1) were used to give references for each mobile genetic element, respectively bacteriophages or insertion elements. In addition the twelve genome sequences were screened for phage sequences using the phage search tool (PHAST, http://phast.wishartlab.com/). The genomic regions that were identified with PHAST to contain phage sequences were downloaded, and given the name of the phage reference sequence provided by the output file. These phage region sequences again were assembled against the genome sequences of the twelve *E. coli* strains using the Geneious Assembler as described. Then the gene next to the phage region was identified and the respective insertion site noted to be occupied with a phage most related to the respective phage reference provided by PHAST. Hence, the resulting table states the gene name of the occupied insertion site with either the phage reference or the reference of the insertion element (IE) of the inserted genes.

The occupation of each of the 63 insertion sites was used to construct a dendrogram in Bionumerics, version 7.5 (Applied Maths, Sint-Martens-Latem, Belgium). Here the occupation of an insertion site with a bacteriophage was indicated by number two, occupation with any other insertion element by number one, and unoccupied insertion by number zero. The comparison settings for the calculation of the dendrogram were a categorical, value based similarity coefficient and unweighted pair group method with arithmetic mean (UPGMA) as cluster analysis method. In addition, the occupation of insertion sites was used to generate MSTree in Bionumerics as described in chapter 2.2.3.

3 RESULTS

3.1 MLST of non-O157 EHEC strains

A total of two hundred and fifty *E. coli* strains that were typed as STEC and EHEC strains, thus positive for at least one *stx* gene, and characterized as one of the most common non-O157 serotypes O26, O103, O111 and O145 were further analyzed with MLST. The 250 strains (Table S 1, p.64) originated from different hosts including the most prominent ones human (n=150) and cattle (n=89), but also single strains from other sources such as food, water and non-bovine animals (sheep, cat, camel, wild rabbit). Most strains were isolated in Germany (n=207) and other European countries (n=30, including n=24 from the UK), but also from Africa and North America (n=6). For O-serogroup O103 a total of 96 strains were further analyzed, including the serotypes O103:(H11; H18; H2; H21; H25; H3; H31; H34; NM) with O103:H2 (n=80) representing the most frequent occurring serotype. O-serogroup O26 was represented by n=76 strains with serotypes O26:(H11; H19; NM) and O26:H11 (n=44) as the most frequent one. For O145 a number of 42 strains was considered for MLST including the serotypes O145:(H18; H25; H28; H34; NM) with the non-motile serotype O145:NM as the most numerous one. O-serogroup O111 (n=36) included the following serotypes O111:(H10; H2; H8; NM) with O111:NM (n=28) as most frequent serotype (Table 1).

	H-antigen	H10	H11	H18	H19	H2	H21	H25	H28	H3	H31	H34	H8	NM	not assigned	total
O-antigen	O103	0	1	1	0	80	2	1	0	1	1	1	0	2	6	96
	O26	0	44	0	1	0	0	0	0	0	0	0	0	24	7	76
	O145	0	0	1	0	0	0	1	9	0	0	1	0	30	0	42
	O111	1	0	0	0	3	0	0	0	0	0	0	3	28	1	36

Table 1 Overview the number of the O- and H-antigen combinations of 250 further analyzed EHEC and STEC strains. The most frequent serotypes are highlighted in grey.

The chosen set of strains contained also fourteen strains of the HUSEC Collection that belong to the non-O157 serotypes under investigation [62].

The phylogenetic relationship of the 250 STEC and EHEC strains that results from MLST analysis is illustrated in a minimum spanning tree (MSTree) (Figure 5). The majority of the strains (n=234, 93.6 %) were assigned to four ST Complexes: STC10, STC20, STC29 and STC32. Of those, STC29 incorporated most of the strains (n=107, 42.8 %) within 15 STs. The major STs of STC29 are the founding ST29, including six strains, and the adjunct ST16 (n=28) and ST21 (n=58), which are single locus variants (SLVs) of ST29. The other 13 STs are mainly represented by a single strain each. ST16 and its SLVs ST294, ST1107 and ST1792 form a cluster of strains with serotypes of O111:(H2; H8; NM) and represent 91.43 % of all O111 strains. Furthermore all 27 O111:NM strains analyzed within this study can be found within these three STs. Only three O111 strains did not cluster within STC29. Those are located within the unrelated ST9 (O111:H2, isolated from cattle), ST43 (STC10, O111:H10, a human isolate of the HUSEC collection (HUSEC001)) and ST165 (O111:H not assigned, isolated from a wild rabbit with enteritis). Strains of serotypes O26:H11/NM clustered mainly into ST21 and ST29 and their SLVs, representing the vast majority of all investigated O26 strains (n=74, 96.1 % of all O26 strains). Three O26 strains, that were isolated from cattle, were unrelated to STC29 and

belonged to ST129 (O26:H not assigned), which differs in three alleles from ST29, ST1125 (O26:H19), which is different in four alleles and ST986 (STC10, O26:H11). Only one strain of O-serogroup O103 can be found within STC29 and represents ST723 (O103:H11, isolated from human). The majority of the O103 strains (n=86, 89.58 % of all O103 strains) clustered within STC20. STC20 combines eight STs with ST17 (n=73) and ST386 (n=8) being the largest ones, while the other STs are mainly represented by single strains. Besides the formation of STC20 distinct strains of O-serogroup O103 are spread all over the MSTree. These O103 strains all harbor individual serotypes due to various H-antigens (H18, H2, H21, H25, H31, H34, NM) (Figure 6).

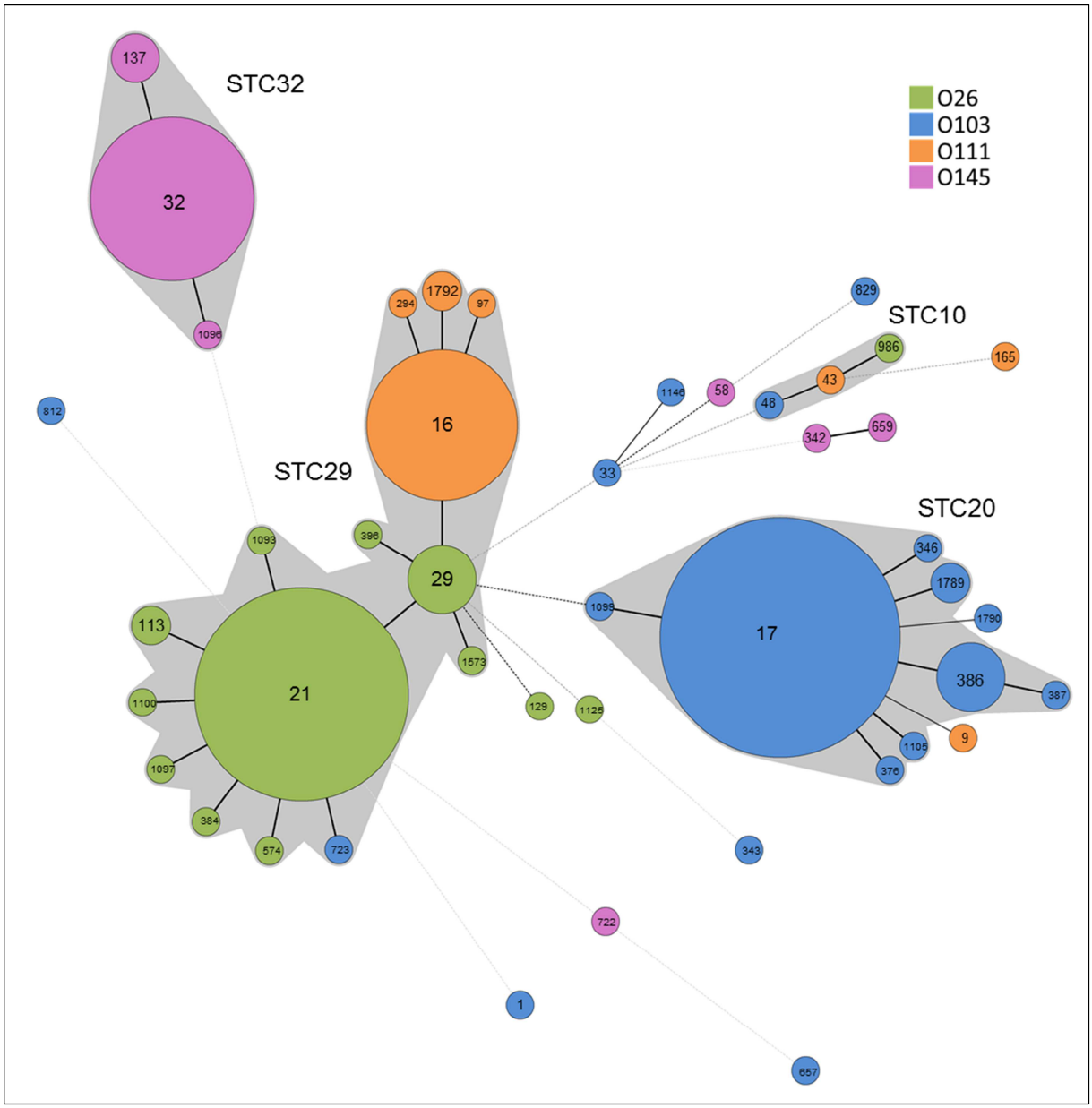

Figure 5 Minimum spanning tree (MSTree) displaying the population structure of 250 Shiga-Toxin harboring *Escherichia coli* strains (STEC and EHEC) based on allele sequence combinations of genes *adk, fumC, gyrB, icd, mdh, purA* and *recA*. The O-serogroups of the strains are indicated by different colors. The size of the circles represents the number of strains within each ST. Bold lines between the STs represent a distinction in one allele, more than one allele difference is indicated with dotted lines. Three main Sequence Type Complexes (STCs) can be recognized, namely STC29 (107 strains), STC20 (86 strains) STC32 (38 strains) and STC10 (three strains), leaving just 16 strains (6.4%) unassigned to these STCs.

Most strains with O-serogroup O145 clustered into the three STs, ST32, ST137, and ST1096, that form STC32 (n=38, 90.48% of all O145 strains). This leaves four O145 strains not assigned to any STC, but allocated over the MSTree. Again they have distinct flagellar (H) antigens, with an O145:H18 strain (ST58, isolated from cattle), and three human strains of serotypes O145:H34 (ST722), O145:H25 (ST342) and a non-motile strain O145:NM (ST659).

The MLST result of non-O157 STEC and EHEC strain revealed a cluster formation of O26 and O111 strains within STC29, indicating a phylogenetic relationship between strains of different O-serogroups. In conclusion this complex is the focus of further investigations to get insights into the phylogeny of these related non-O157 strains.

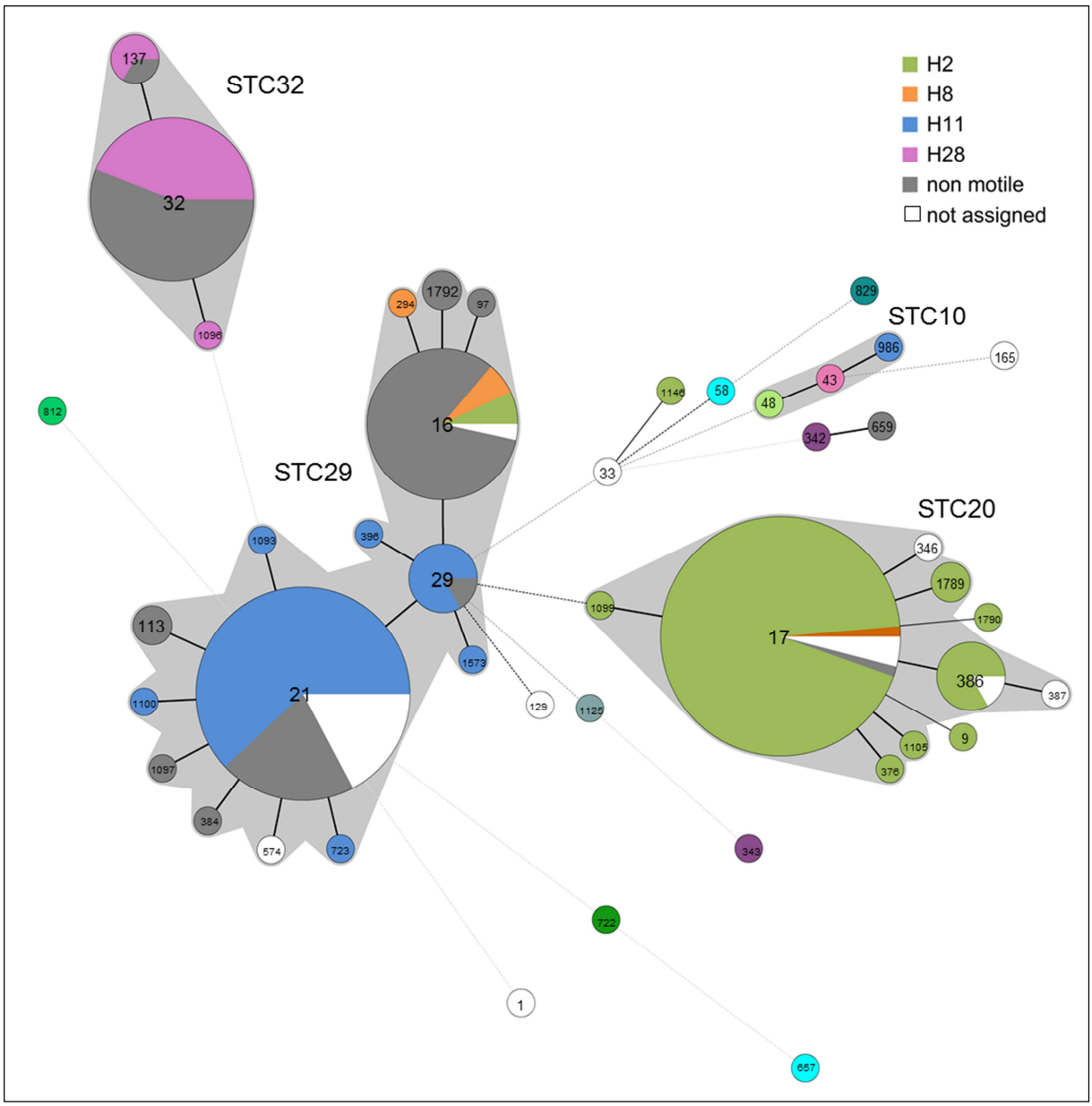

Figure 6 MSTree displaying the population structure of STEC and EHEC strains based on *E. coli* MLST results. The H-types of the strains are indicated by different colors. The size of the circles represents the number of strains within each ST. Bold lines between the STs represent a distinction in one allele, more than one allele difference is indicated with dotted lines. ST Complexes are highlighted in with grey background.

3.1.1 Inclusion of non-O157 aEPEC strains

STC29 comprises STEC and EHEC strains of the important non-O157 O-serogroups O26 and O111. It has been shown that aEPEC strains often display these O-serogroups. Thus STC29 constitutes an appropriate setting to analyze the relation between STEC/ EHEC and aEPEC strains. For this reason we included aEPEC strains of the analyzed non-O157 serotype into the MLST analysis.

The pathotypes within STC29 were determined using PCR amplification of the characteristic genes *stx*$_1$, *stx*$_2$, *escV* and *bfpB*. The gene *escV* is used as a marker for the presence of intimin that is located on the LEE PAI. The *bfpB* gene is a marker for the EAF plasmid that is harbored by typical EPEC. Hence, if a strain is *bfpB* negative and no additional *stx* gene was identified, the pathotype aEPEC was assigned.

A set of 41 aEPEC strains (Table S 2, p.68) of bovine (n=28), human (n=10), and ovine origin (n=2), as well as one strain of unknown origin, were typed as STs clustering within STC29. These aEPEC strains were isolated in Germany (n=12), the United Kingdom (n=17) and other European countries (n=5), but aEPEC strains from Australia (n=3) and USA (n=3) were included as well. In addition to three EHEC strains of the DECA Collection already included into the analysis (DEC8A, DEC8B, DEC10E), four DECA aEPEC strains were included (DEC9A, DEC9B, DEC9C, DEC9E). All aEPEC strains of STC29 had an O26 O-serogroup with the resulting serotypes O26:H11 (n=20) and the non-motile O26:NM (n=17), with one exception of a bovine O111:NM strain (ST16). For three O26 strains the H antigen had not been typed.

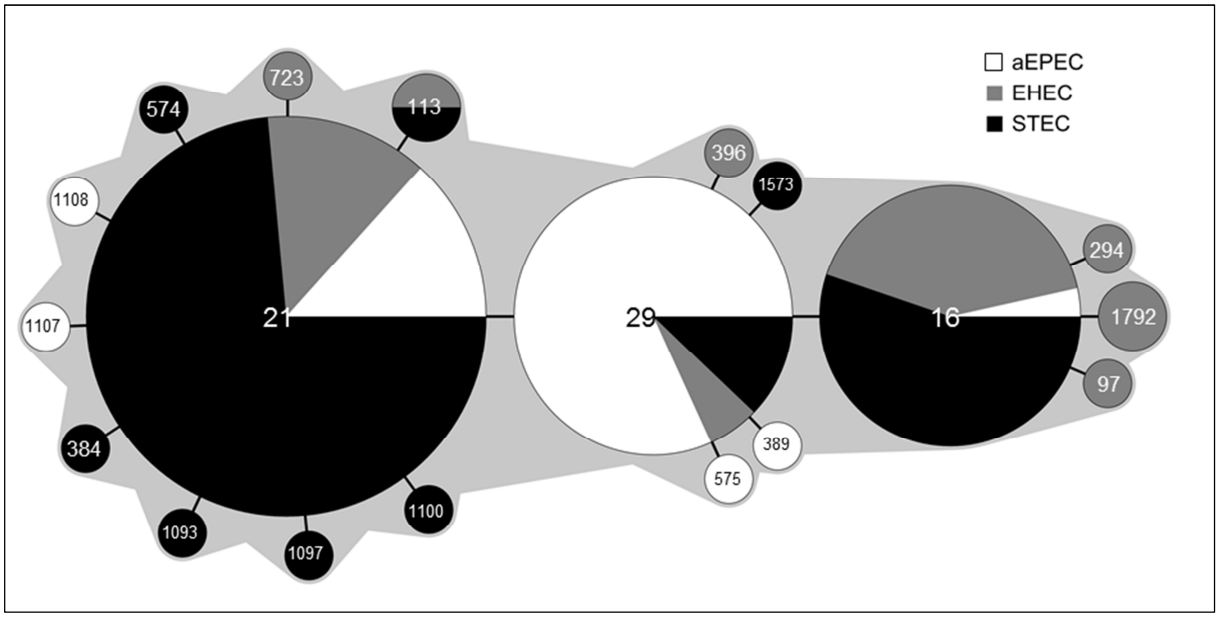

Figure 7 MSTree showing the population structure of 148 *E. coli* strains designated to Sequence Type Complex STC29. This graph illustrates the pathotype of each strain indicated by different colors. STC29 is constituted of 41 aEPEC and 107 STEC/ EHEC strain.

The greater number of the aEPEC strains clustered within STs formed by the STEC/ EHEC strains (n=37, 90.24%). Yet, also two additional STs occurred as SLVs of ST21, namely ST1107 and ST1108, and two STs as SLVs of ST29, respectively ST389 and ST575. Each of the four additional STs is represented by a single strain (Figure 7). The founding sequence type, ST29 and its SLVs, included

most of the aEPEC strains (n=28, 68.29%). Hence, within our collection of strains of ST29 and its SLVs the number aEPEC outweighs the number of just eight STEC/ EHEC strains (22.2% STEC/ EHEC, 77.8% aEPEC). ST21 and its SLVs subjoined twelve aEPEC to the STEC/ EHEC strains and ST16 included a single aEPEC strain. In conclusion, a total of 148 strains are the basis of the further comparative analysis to unravel the microevolution of aEPEC and STEC/ EHEC in this thesis.

3.1.2 Strain features of Sequence Type Complex STC29

Since the phylogenetic relationship of aEPEC and STEC/ EHEC strains of the O-serogroups O26 and O111 is properly reflected within STC29, this STC can be used to analyze the relation of other properties of the included strains. To visualize a possible host specificity of strains within STC29, the host origin of each strain was color labeled in the respective MSTree (Figure 8).

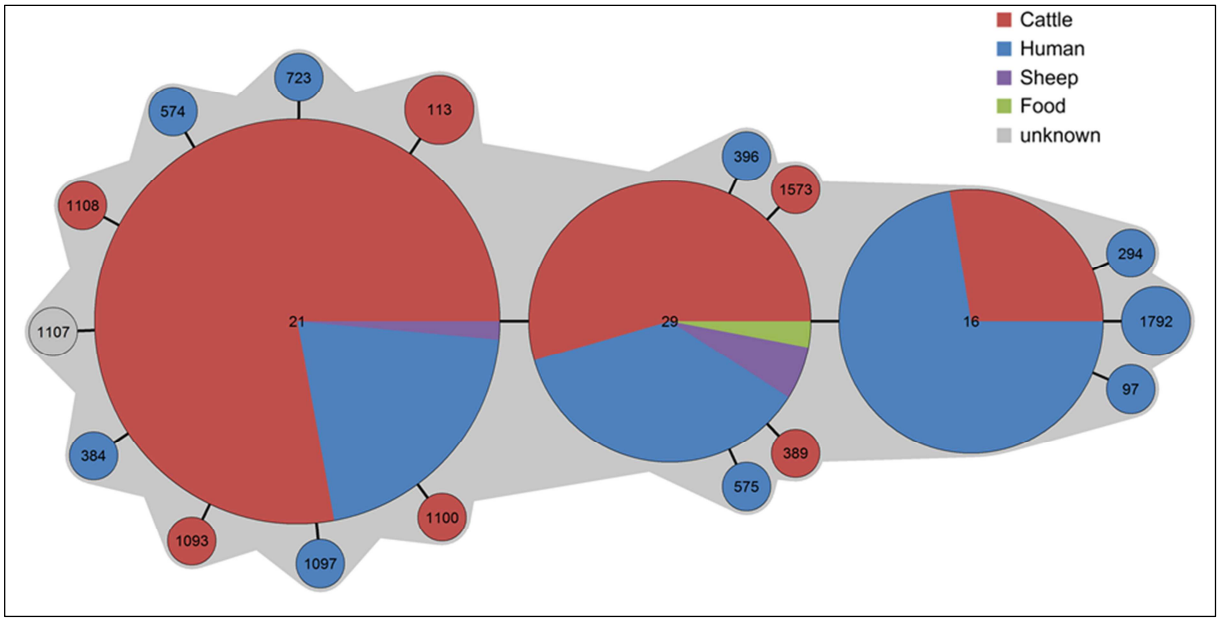

Figure 8 MSTree showing the population structure of 148 *E. coli* isolates of pathotypes aEPEC and STEC/ EHEC designated to STC29. This graph illustrates the host origin of each strain indicated by different colors. The strains mainly originate from human (n=57) and bovine (n=86) hosts and do not form any clusters.

No cluster formation could be detected for the host origin of the strains within STC29. In total STC29 comprised 57 strains isolated from humans, 86 strains isolated from cattle, three strains isolated from sheep, and a single strain of food origin, as well as one isolate of which the source is not known. Human and bovine strains are jointly located in all of the three main STs: ST16, ST21 and ST29. ST21 and its SLVs include nineteen strains isolated from humans, and 58 strains isolated from cattle. Additionally, one strain isolated from a sheep and one of unknown source can be found within ST21. ST29 and its SLVs comprise thirteen human strains, twenty bovine strains, as well as two strains isolated from sheep and one isolated from a food source. Within ST16, and its three SLVs, 25 human and eight cattle strains are located. Hence, the strains do not show a host specific cluster formation in the allele based MSTree.

Strains positive for any *stx* gene were further subtyped with PCR amplification following the protocol published by Scheutz *et al.* [65]. The resulting *stx* subtypes were highlighted in the MSTree of STC29 to identify possible cluster formation of strains that show an identical set of these characteristic virulence associated genes (Figure 9).

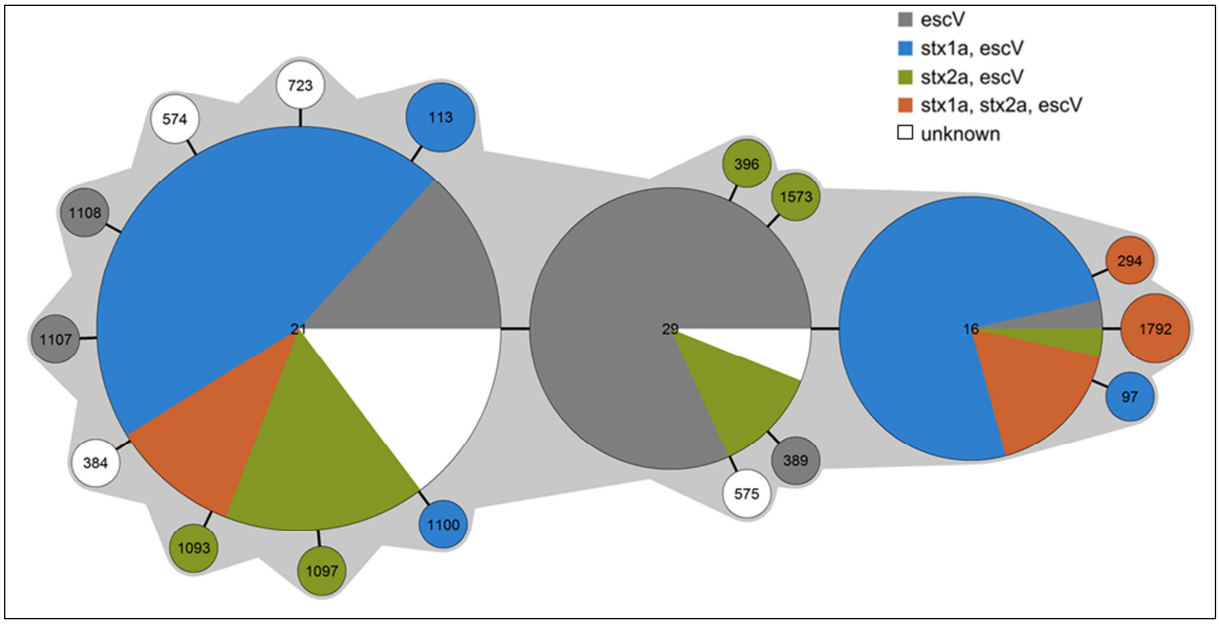

Figure 9 MSTree showing the population structure of 148 *E. coli* isolates of pathotypes aEPEC and STEC/ EHEC designated to STC29. This graph illustrates the presence of characteristic genes for the pathotypes aEPEC (*escV* as representative gene for the locus of enterocyte effacement LEE) and STEC, harboring at least one *stx* gene. The gene combinations are highlighted with different colors.

All strains of STC29 that could be analyzed for the presence of escV were detected as *escV* positive ones, implying that all strains harbored the intimin gene as a representative for the LEE PAI. Further the aEPEC strains of STC29 are visible in the MSTree due to the presence of this gene and the absence of any *stx* gene. The Shiga-Toxin subtypes stx_{1a} and stx_{2a} were identified for the STEC/ EHEC strains of STC29. The *stx* genes are either present as a single stx_{1a}, or a single stx_{2a} in one strain, but also as combination of both *stx* genes harbored by one strain. Within ST21 and its SLVs all combinations of the tested virulence associated genes were present among the strains. A number of 34 strains that harbored stx_{1a} in addition to *escV* can be found within these STs, further thirteen strains positive for stx_{2a} and *escV*, and seven strains that harbored both stx_{1a} and stx_{2a}. The twelve aEPEC of these STs are represented by the solitary presence of *escV.* Within ST29, and its adjunct STs, the majority of aEPEC strains of the examined population cluster together, which is represented by a number of 27 strains that are *escV* positive, but *stx* negative. Not a single strain with an stx_{1a} gene was found among the strains of ST29 and its SLVs, but six strains were identified as positive for stx_{2a}. Within ST16 and its SLVs the majority of strains harbored either a single stx_{1a} (n=23) or both stx_{1a} and stx_{2a} (n=8). Two strains of these STs were positive for a single stx_{2a} gene and one strain did not harbor any *stx* gene, thus represents the only aEPEC strain of ST16. One strain of ST16 was, in addition to stx_{1a} and $stx_{2a,}$ positively typed for the stx_2 subtype stx_{2d} (not shown in Figure 9). Some of the STC29 strains (n=16) could not be further subtyped, as the strains were not in possession of our Institute nor was information regarding the *stx* subtypes of this strains published.

3.2 SNP-Analysis of strains of non-O157 strains of STC29

MLST analysis of the most important non-O157 O-serogroups O26, O103, O111 and O145 revealed cluster formation of the majority of O26 and O111 STEC/ EHEC strains, namely STC29. When the strains of STC29 were further examined, their diverse population structure became evident. Within the STs of STC29, strains of various host origins clustered together. The determination of characteristic virulence associated genes did not reveal any cluster formation within STC29, instead strains of the pathotypes aEPEC and STEC/ EHEC were found together within single STs. Therefore STC29 represents the phylogenetic background of an apparently diverse group of strains and constitutes a set of strains suitable to analyze the microevolution of this *E. coli* phylogroup in greater detail.
Hence, a group of 68 of the analyzed strains was chosen for whole genome sequencing, including twelve aEPEC and 56 STEC/ EHEC strains of human (n=23) and bovine (n=44) origin, and one strain isolated from food source. Additionally, 19 whole genome sequences of STC29 of human O26 strains (18 EHEC strains, one aEPEC strain), provided by Alexander Mellmann (Institute for Hygiene, University of Muenster), were included. Further, the publicly available genome sequences of twelve DECA strains [109], MLST typed as STs of STC29, were included into the further analysis (Table 2), resulting in a total of 99 whole genome sequences (WGSs) (Table S 3, p.69 ff.).

Strain designation	STC	ST	Serotype	Pathotype	Host species	Year	City/Region	Country	Continent
DEC8A	STC29	16	O111a:NM	EHEC	Human	1977	unknown	USA	North America
DEC8B	STC29	294	O111:H8	EHEC	Human	1986	Idaho	USA	North America
DEC8C	STC29	21	O111:NM	EHEC	Cattle	1986	South Dakota	USA	North America
DEC8D	STC29	21	O111:H11	aEPEC	Human	1953	unknown	Cuba	Central America
DEC9A	STC29	29	O26:H11	aEPEC	Human	1961	Wisconsin	USA	North America
DEC9B	STC29	29	O26:NM	aEPEC	Human	1979	New Hampshire	USA	North America
DEC9C	STC29	29	O26:NM	aEPEC	Human	1952	unknown	Switzerland	Europe
DEC9D	STC29	29	O26:H11	aEPEC	Human	1967	unknown	Denmark	Europe
DEC9E	STC29	29	O26:H11	aEPEC	Human	unknown	Mexiko	USA	North America
DEC10A	STC29	21	O26:H11	EHEC	Human	unknown	unknown	UK	Europe
DEC10B	STC29	21	O26:H11	EHEC	Human	1986	Brisbane	Australia	Australia
DEC10D	STC29	21	O26:H11	aEPEC	Human	1952	unknown	France	Europe

Table 2 Information on the twelve strains of the DECA Collection included into the analysis [109].

In order to perform a SNP-analysis on the 99 WGSs of STC29 strains, the maximum common genome (MCG) of these strains was identified. The MCG of the aEPEC (n=20) and STEC/ EHEC strains (n=79) of STC29 consisted of 2,649,433 bp, encoding 3,714 genes. Those diverse genes included, *inter alia*, 218 proteins of the inner or outer membrane, membrane bound proteins or proteins associated with the biogenesis of the membrane. 105 genes were associated with bacterial surface structures, such as flagellar, fimbriae, curli and pili, including adhesins and regulators of those surface structures, as well as proteins involved in their biosynthesis. 22 genes were associated with bacteriophages, either they encode for proteins involved in the assembly of the phage particle, capsid and tail, or have regulator functions like the antiterminator Q and enzymes, such as integrase and endolysin. Since all of the STC29 strains involved in this analysis harbor the LEE PAI, the MCG includes also nineteen genes of this TTSS, with outer and inner membrane protein channels, and secreted proteins, i.e. *espA*, as well as intimin. For 298 genes the function is not further known, and

such genes are only annotated as encoding for hypothetical proteins. The others genes of the MCG encode mainly for proteins involved in metabolic cycles or biogenesis.
Within these 3,714 genes of the MCG a number of 11,891 SNP-sites were identified, enabling in depth analysis of the strains phylogeny. The relationship of the strains is hereafter displayed in MSTrees.

3.2.1 Distribution of STs within the SNP-analyzed population of EHEC and aEPEC of STC29

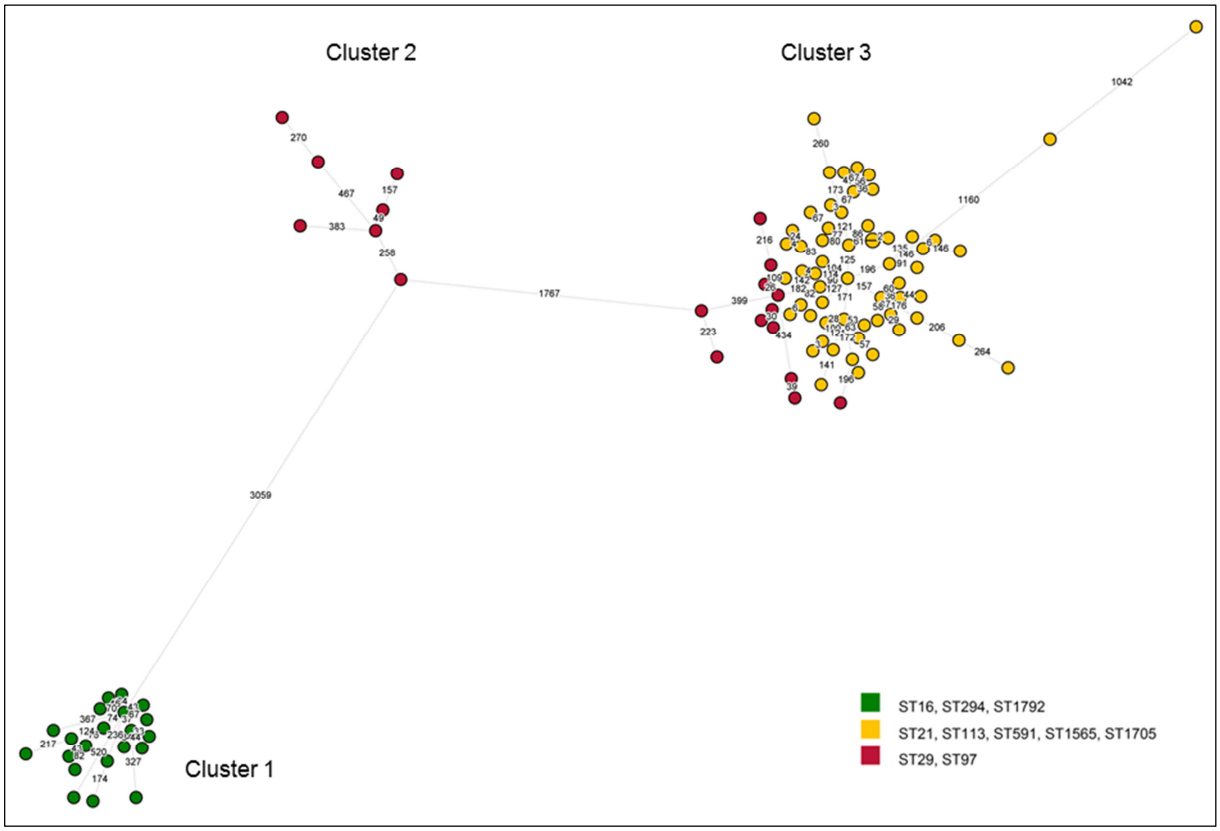

Figure 10 MSTree based on 11,891 SNPs that were detected within the 3,741 genes of the MCG of 99 STEC/ EHEC and aEPEC strains of STC29. The STs are highlighted for the groups of the three major STs and their adjunct SLVs. One circle represents a single strain. The dashed lines between the circles mark the number of SNPs that separate one strain from the next closest relative; the number of SNPs is also displayed by the number on each line. Three distinct groups of strains were identified and designated as Cluster 1, Cluster 2 and Cluster 3.

The SNP-analysis of the MCG of 99 STEC/ EHEC and aEPEC strains of STC29 revealed the formation of three clusters when displayed as an MSTree (Figure 10). Cluster 1 condensed all strains of ST16 and its SLVs (n=21) with distances of 33 to 520 SNPs difference between the single strains. Cluster 2 is composed of seven strains of ST29 and located in the center of the MSTree. The strains of Cluster 2 are 49 to 467 SNPs distinct from another. Cluster 1 and Cluster 2 differ in 3,059 SNPs. Cluster 3 consists partly strains of ST29 and a single ST97 stain (total n=12), while the majority of Cluster 3 in this SNP-analysis is represented by the total of all strains of ST21 and its SLVs (n=59). The ST29 strains of Cluster 3 differ in 18 to 434 SNPs. A single strain of ST97 clusters within the strains of ST21 and its SLVs and has 291 SNPs difference to the nearest strain of ST29. The strains of ST21 and its SLVs differ in three to 260 SNPs from one another. Two strains of ST21 do not show

any difference in their SNP-profile at all: P3146/08-2 and P3146/08-3, the nearest neighbor of those two strains is P3146/08-1. All three strains were isolated from healthy cattle in Giessen, in 2008 and are *stx*$_{1a}$ positive strains of serotype O26:H11. Two strains of Cluster 3 are more distantly located from the rest of the major group of strains and differ in 1160 and 1861 SNPs from the nearest ST21 strain: DEC8D and DEC8C (human, ST21, EHEC, O111:H11/NM).

3.2.2 Distribution of O-serogroups within the SNP-analyzed population of EHEC and aEPEC of STC29

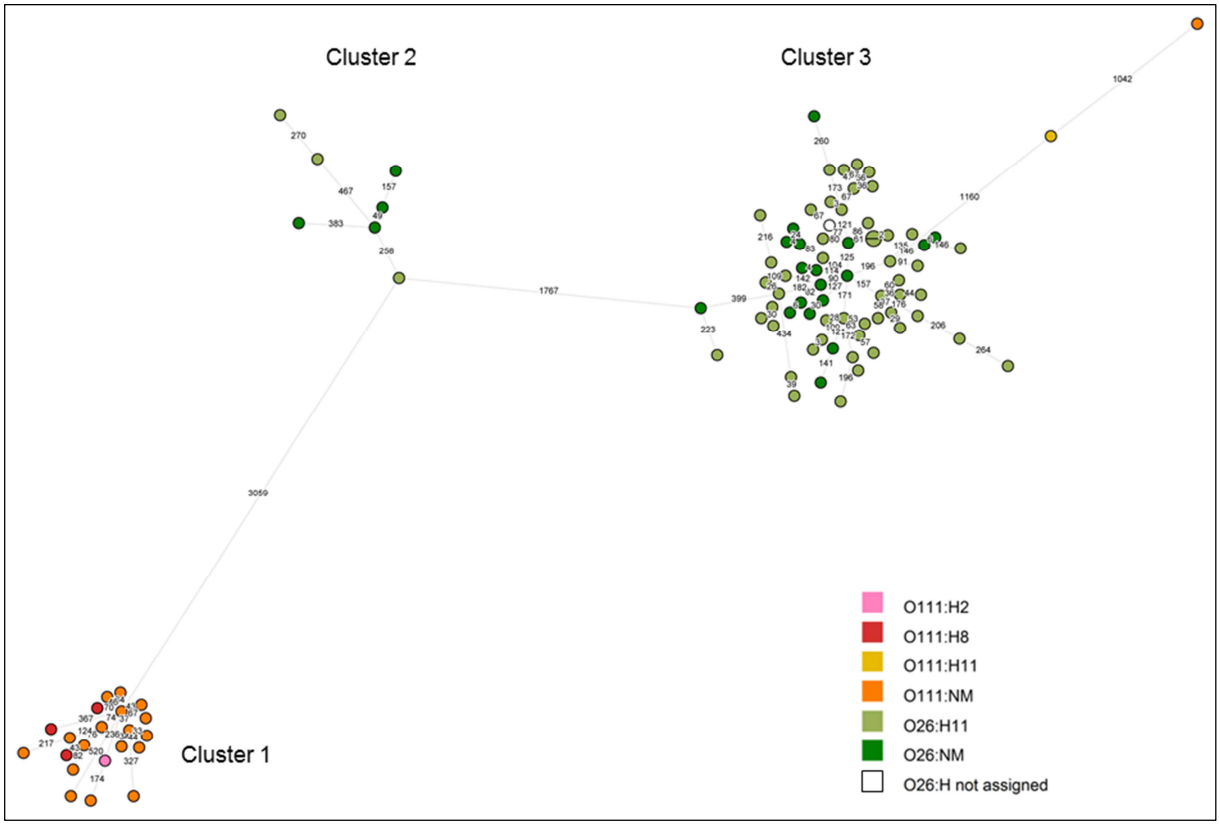

Figure 11 MSTree based on 11,891 SNPs that were detected within the 3,741 genes of the MCG of 99 STEC/ EHEC and aEPEC strains of STC29. The O-serogroup of each strain is highlighted for O26:(H11, NM) and O111:(H2, H8, H11, NM). Cluster 1 represents all O111 strains, while O26 strains are distributed into Cluster 2 and Cluster 3.

When the serotype of each strain is highlighted in the MSTree of the SNP-analyzed strains, an O111 cluster, including O111:H2 (n=1), O111:H8 (n=8), and O111:NM (n=16) strains, is resembled by Cluster 1, which is formed by the strains assigned to ST16 and its SLVs (Figure 11). These results match the finding of the MLST analysis, that strains of O-serogroup O111 are mainly represented by ST16 and its SLVs. Two strains with serotypes O111:H11 and O111:NM are not located in Cluster 1, but are distantly adjunct to Cluster 3 and were identified as DEC8D and DEC8C (see Chapter 3.2.1), DECA strains that originally had been identified by multilocus enzyme electrophoresis. Cluster 2 and Cluster 3 are composed of strains of O-serogroup O26. Cluster 2 comprises three strains of serotype O26:H11 and four strains of serotype O26:NM. In Cluster 3 39 strains of serotype O26:H11, 17 non-motile strains of O26:NM, and one strain with an unassigned H-type group together. These results are consistent with those of the MLST analysis of aEPEC and STEC/ EHEC strains of STC29.

3.2.3 Distribution of host origin and pathotypes within the SNP-analyzed population of EHEC and aEPEC of STC29

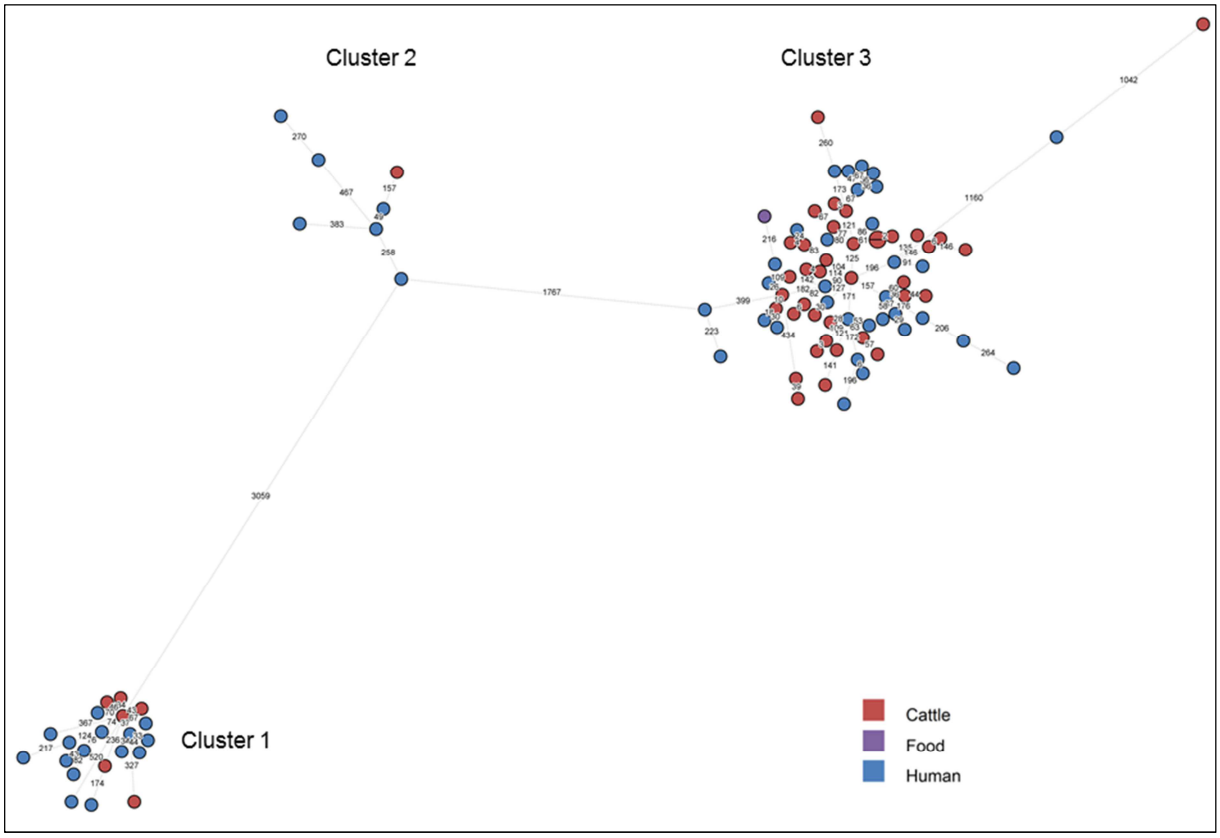

Figure 12 MSTree based on 11,891 SNPs that were detected within the 3,741 genes of the MCG of 99 STEC/ EHEC and aEPEC strains of STC29. The host origin, cattle (n=45), human (n=53) and food (n=1), of each strain is highlighted. Human and cattle strains are spread over the whole MSTree without detectable cluster formation.

The host origin of the strains, when highlighted in the MSTree of the SNP-analyzed population, did not reveal any cluster formation. Cluster 1 consists of six strains isolated from cattle and 15 strains of human origin. Cluster 2 includes six human strains and one strain isolated from cattle. In Cluster 3 human (n=32), and bovine (n=38) strains clustered together with a single strain isolated from food source. Thus, this analysis gives strong evidence of the zoonotic nature of the analyzed strains.

When the pathotype of each strain within the SNP-analyzed population of aEPEC and STEC/ EHEC strains of STC29 is highlighted (Figure 13), an accumulation of seven aEPEC in Cluster 2 is detectable. This is consistent with the finding of the MLST analysis that the majority of aEPEC strains can be found in ST29. A single aEPEC strain is located at the center of Cluster 3 and surrounded by O111 STEC/ EHEC strains. Another aEPEC strain, with a second adjunct one that differs in 223 SNPs is located between the Cluster 2 and the majority of strains of Cluster 3. The remaining ten aEPEC strains are distributed among the STEC/ EHEC strains in Cluster 3.

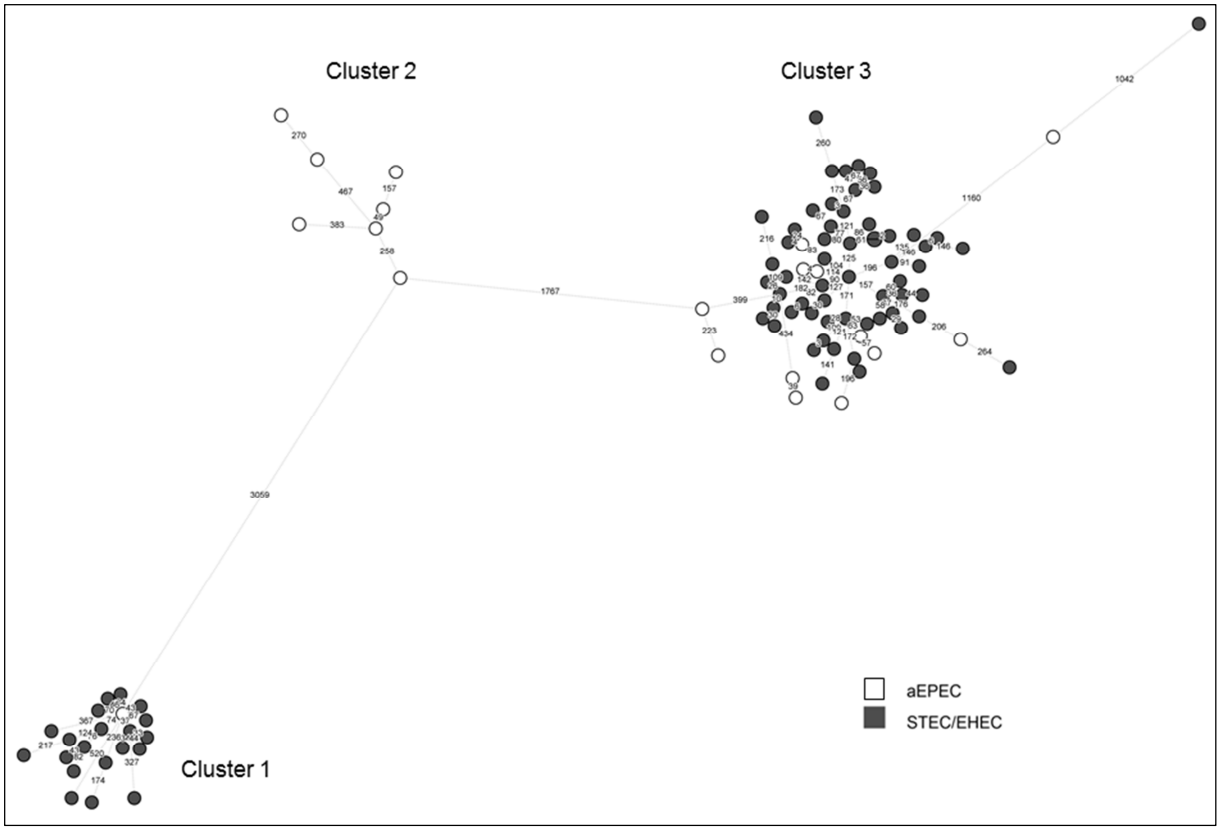

Figure 13 MSTree based on 11,891 SNPs that were detected within the 3,741 genes of the MCG of 99 STEC/ EHEC and aEPEC strains of STC29. The pathotype, aEPEC (n=20) and STEC/ EHEC (n=79), of each strain is highlighted.

3.3 VAGs of strains of non-O157 strains of STC29

A number of 34 virulence associated genes (VAGs) were further analyzed for their presence and absence in the WGS of 99 aEPEC and STEC/ EHEC strains of STC29 (Table S 4, p.72 ff.). The applied online-database “VirulenceFinder-1.5” screens for the presence of 76 VAGs [108]. The 34 VAGs, which were selected for further analysis, were present in at least one of the analyzed genomes. Those VAGs encoded, among others, for the subtypes of Shiga-Toxin, TTSS proteins, like intimin, *nle* effectors, enterohemolysin, and other toxin genes, such as toxin B and EAST-1 (Table 3).

VAG	**Encoded Protein**	No. positive aEPEC	No. positive STEC/EHEC
stx1a	Shiga toxin 1, variant a	0	61
stx2a	Shiga toxin 2, variant a	0	30
eae	Intimin	20	79
tir	Translocated intimin receptor protein	20	79
nleA	Non-LEE encoded effector A	19	79
nleB	Non-LEE encoded effector B	20	76
nleC	Non-LEE encoded effector C	17	69
iha	Adherence protein	10	42
espA	Type III secretion system	20	79
espB	Secreted protein B	19	59
espF	Type III secretion system	20	67
espI	Serine protease autotransporters of Enterobacteriaceae	3	1
espJ	Prophage encoded type III secretion system effector	20	78
espP	Putative exoprotein precursor	10	38
efa1	EHEC factor of adherence	15	64
ehxA	Enterohaemolysin	9	58
astA	EAST-1 heat-stable toxin	18	72
cif	Type III secreted effector	20	79
lpfA	Long polar fimbriae	20	79
epeA	Serine protease autotransporters of Enterobacteriaceae	0	1
toxB	Toxin B	4	29
iroN	Enterobactin siderophore receptor protein	0	1
etpD	Type II secretion protein	0	6
iss	Increased serum survival	19	61
cba	Colicin B	0	11
cma	Colicin M	0	16
celb	Endonuclease colicin E2	1	41
tccp	Tir cytoskeleton coupling protein	12	15
mchF	ABC transporter protein MchF	1	7
mchC	MchC protein	1	2
mchB	Microcin H47 part of colicin H	0	2
pfrB	P-related fimbriae regulatory gene	20	79
gad	Glutamate decarboxylase	18	63
katP	Plasmid-encoded catalase peroxidase	4	35

Table 3 Virulence associated genes (VAGs) analyzed for their presence and absence in the whole genome sequences of 99 aEPEC and STEC/ EHEC of STC29. The number of positively tested aEPEC (total n=20) and STEC/ EHEC (total n=79) are listed for each of the 34 VAGs.

All of the strains harbored the LEE PAI, therefore all 20 aEPEC and 79 STEC/ EHEC were positively tested for the genes *eae*, encoding for intimin, *tir*, the translocated intimin recptor protein, *cif*, encoding for a secreted effector, and *espA,* a serine protease. Further, all strains possessed the *lpfA*, which encodes for the long polar fimbriae and *prfB*, a P-related fimbriae regulatory gene.

The relationship of the strains can be displayed in MSTrees based on the binary VAG-profile of each strain (Figure 14). Cluster 1 of the SNP-analyzed population structure combined all strains of ST16, and roughly represented all analyzed strains of O-serogroup O111. In the VAGs-profile based MSTree the same O111 strains formed a cluster once again (Figure 14A Cluster 1). They mainly harbored stx_{1a} (n=13), or both stx_{1a} and stx_{2a} in combination, just one ST16 strain possessed only stx_{2a} and another strain was *stx* negative. The strains of ST16 and its SLVs did not have the genes *iss*, for increased serum survival, and *espB*, a secreted protein of the TTSS, that were harbored by all other strains of STC29. All of the 21 strains of ST16, and adjunct STs, possessed the *iha* gene that encodes for an adherence protein; this gene was further harbored by only 31 other strains of the analyzed STC29 population. Except of two ST16 strains, that were isolated from different hosts and had a different H-type, all other strains of Cluster 1 harbored *celb*, for Colicin E2; just 23 other strains of the analysis harbored this gene. As already detected with SNP-analysis, the two ST21 EHEC strains of O-serogroup O111, DEC8C and DEC8D, did not cluster together with the other O111 strains in Cluster 1, but rather could be found among the O26 strains of ST21 and its SLVs (Figure 14B). The central Cluster 2, formed by seven aEPEC strains of ST29 in the SNP-analysis, is circled in Figure 14 since it is not displayed as a separated group, as in the SNP-based MSTree (Figure 10). Nevertheless, a central accumulation of aEPEC strains was detectable in the MSTree of the VAG-profile. The strains that formed Cluster 2 in the SNP-based MSTree, are surrounded by aEPEC strains of ST21 in MSTree of the VAG-profiles, resulting in a larger aEPEC cluster in the center of this MSTree (Figure 14C). The seven aEPEC strains of Cluster 2 all possessed the *tccp* gene that encodes for a Tir cytoskeletal coupling protein. Five neighboring aEPEC strains of ST21 and ST29 also harbored *tccp*. Seven other strains of ST29 cluster together more distinctly in the upper part of the MSTree, separately from the other strains of the ST21. All of those strains were stx_{2a} positive. Additionally, these strains harbored the genes *cba* and *cma*, encoding for Colicin B and Colicin M, which were found in just four other strains, in case of *cba*, and nine other strains for *cma*, of the analyzed genome sequences. Furthermore, six of these strains possessed the gene *etpD* that is involved in a type two secretion system and not harbored by any other strains of the analyzed STC29 strains. When the host origin of each strain is highlighted in the MSTree based on the VAG-profile (Figure 14D) again no cluster formation for any host origin was detectable, and therefore no host specific VAG-profile was identified.

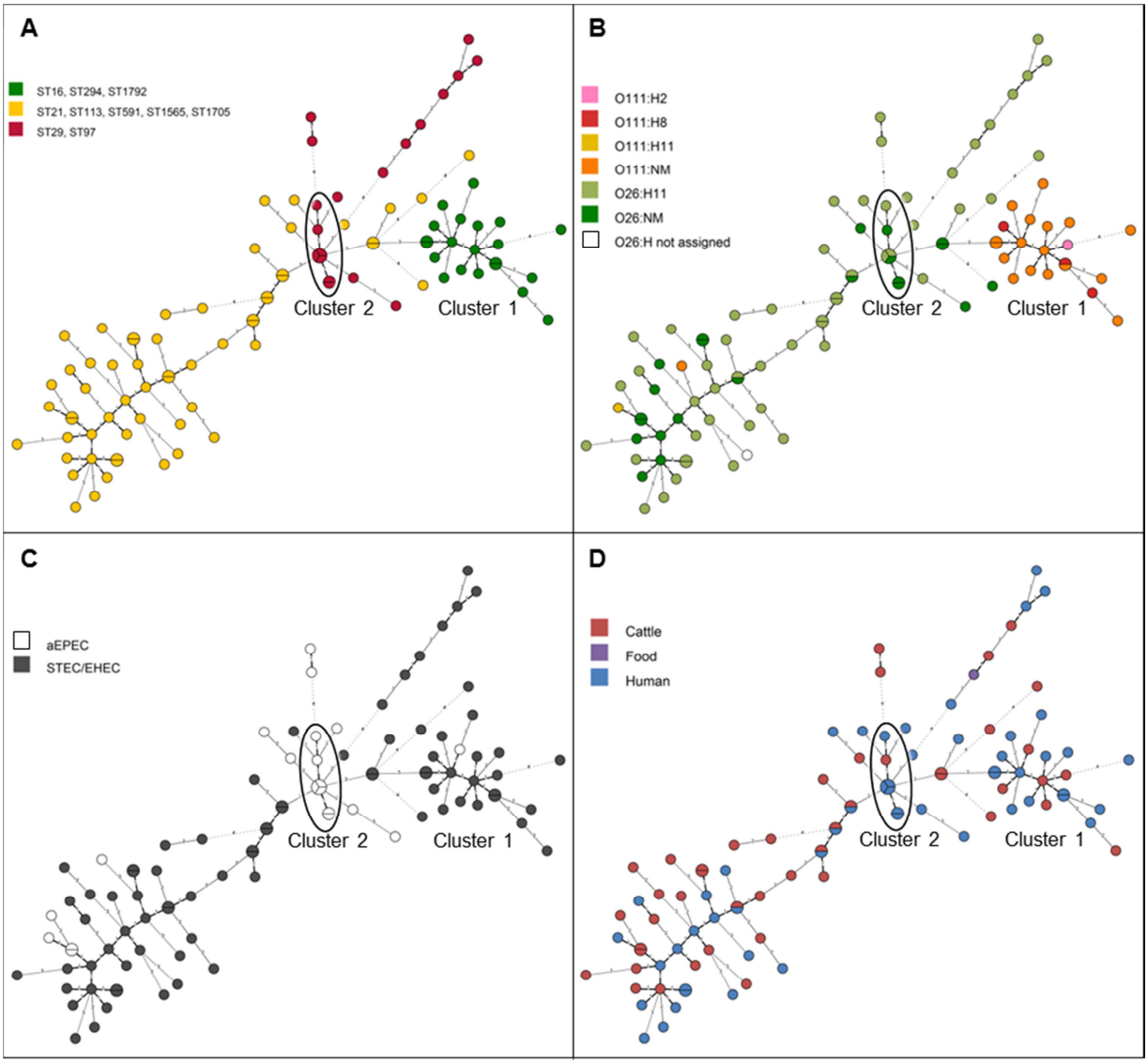

Figure 14 MSTrees based on the presence and absence of 34 VAGs within the genome sequences of 99 aEPEC and STEC/ EHEC strains of STC29. Bold lines represent a difference in one VAG, thin lines represent differences in two or three VAGs, dashed lines mark differences in four VAGs between the neighboring strains. The size of each circle indicates the number of strains that share an identical VAG-profile, the number of strains in each circle is further displayed by the node disk. (A) The three major STs, ST16, ST21 and ST29, and their SLVs, designated for each strain are highlighted in different colors. Cluster 1 is composed of ST16 strains, and equals Cluster 1 of the SNP-analysis. The circled seven strains equal Cluster 2 of the SNP-analysis. (B) The serotype of each strain is highlighted with different colors, including 23 strains with O-serogroup O111 and 76 strains with O-serogroup O26. (C) The pathotype of each strain is highlighted as either aEPEC (total n=20) and STEC/ EHEC (total n=79). (D) The host origin of each strain is highlighted either as human (total n=53), cattle (total n=45) or food source (total n=1).

3.4 Further analysis of selected strains of the analyzed STC29 population

The population of the analyzed aEPEC and STEC/ EHEC strains of STC29 results in comparable phylogenetic structures when displayed as MSTrees, regardless whether built on allele based MLST, SNP-analysis of the MCG or based on the binary VAG-profile. Therefore some of the strains can be found in relevant positions in the center of the whole MSTree or in branching points. Those strains might stand exemplarily for an evolutionary step of the development of non-O157 *stx*-producing *E. coli* of O-serogroups O26 and O111 (Figure 15). A set of seven aEPEC of ST29 in the center of the

MSTree (DEC9A to DEC9E, LH-8 and IMT19623) is therefore chosen for further analysis, as they might represent the founding and initial strains of the STC29, and further might function as pre-EHEC strains. The aEPEC strain RW2070 of ST16 was chosen as a representative of an intermediate stage for the development of STEC/ EHEC strains of Cluster 1. Next, the aEPEC strain LH-1, which is located close to the cluster of ST21 strains (previously identified as Cluster 2), is chosen for further analysis as it might function as an intermediate stage. The STEC strains 0739/03, which is a *stx*$_{2a}$ positive strain of ST29, as well as the *stx*$_{2a}$ positive strain IMT20337 of ST21 are also included for further analysis. IMT19981 was included as a representative for an aEPEC strain of ST21.

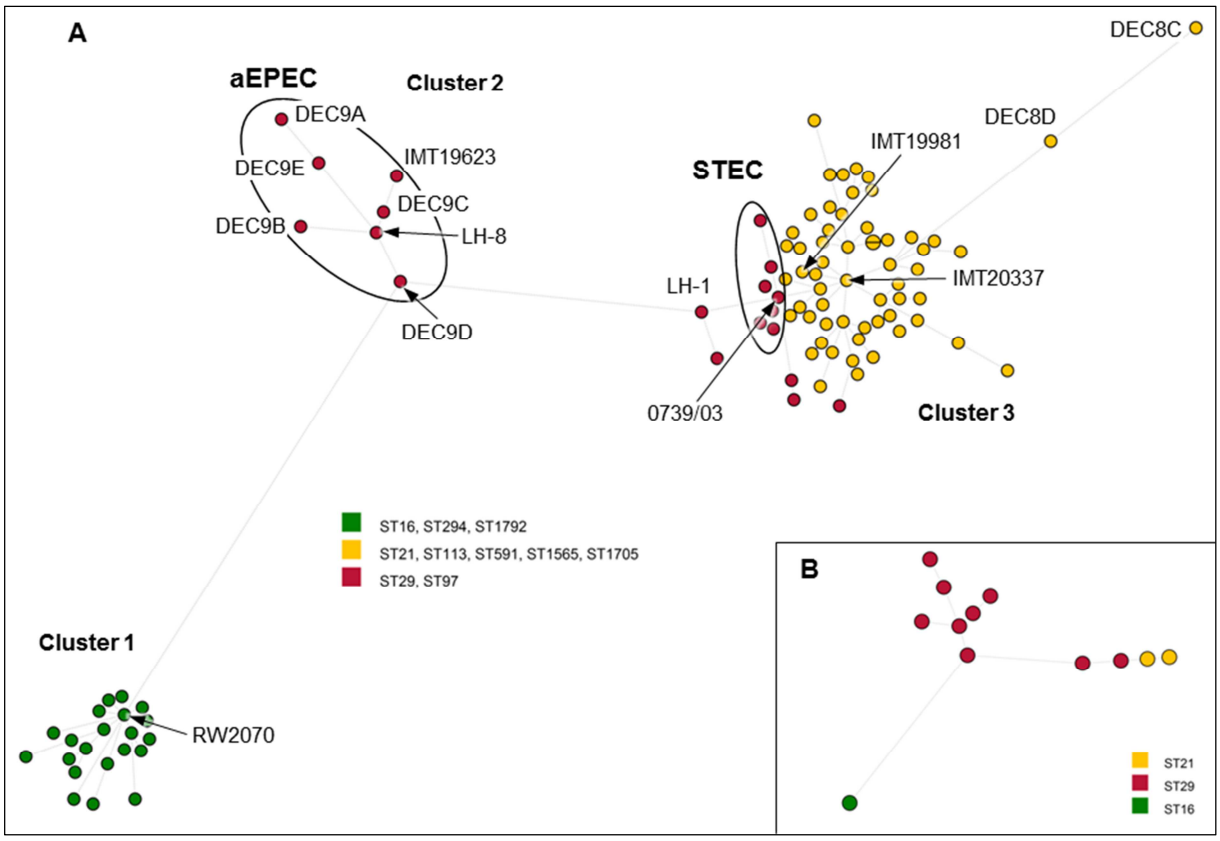

Figure 15 (A) MSTree based on 11,891 SNPs that were detected within the 3,741 genes of the MCG of 99 STEC/ EHEC and aEPEC strains of STC29. The STs are highlighted for the groups of the three major STs and their adjunct SLVs. The locations of important strains are either indicated with arrows or directly designated with their strain number. (B) The resulting MSTree of twelve strains that were chosen for further analysis. DEC8D and DEC8C are not included in this extracted MSTree.

3.4.1 SNP-sites of selected strains of the analyzed STC29 population

For certain strains the localizations of SNP-sites are of interest, respectively such genes of the MCG that distinguish strains from one another due the presence of SNPs. Therefore the alignment of the MCG of two selected strains is searched for genes that present more than five SNP-sites compared to one another. This indicates that the respective gene is different in the two analyzed strains, at least on nucleotide level. The SNPs were counted, and differentiated depending on whether they are synonymous or non-synonymous SNPs according to the amino-acid sequence of the gene. The identity of the nucleotide sequence of each gene, compared between both strains, was noted. The

nucleotide sequence of the gene was used for a blastn search in the ncbi nucleotide database (performed between October and December 2015), and the results were used to gather further information on the molecular function and biological process the gene-encoded protein is involved in. The strain pairs that were chosen for comparison of the MCG genes are: DEC9D and RW2070, to get further insights in the genetic differences between Cluster 1 and Cluster 2; DEC9D and LH-1, to detect the genes differentiating Cluster 2 and Cluster 3; DEC8C and RW2070, to figure out why DEC8C (and DEC8D) does not cluster within Cluster 1, as the rest of the O111 strains.

The comparison of the MCG of DEC9D (ST29, O26:H11, aEPEC, human) and RW2070 (ST16, O111:NM, aEPEC, cattle) revealed a total of 3,059 SNPs. Within the alignment of the MCG of both strains 1,933 of these SNPs, that are located in 53 genes, were further analyzed. These further analyzed genes include 1,174 synonymous SNPs and 759 non-synonymous SNPs (Table S 5, p.75). Thirteen genes of the MCG of DEC9D and RW2070 that were found to differ in more than five nucleotides are located on the LEE PAI (Table 4). The gene *espA* has two synonymous and twelve non-synonymous SNP-sites. The EspA protein forms filamentous surface appendages (EspA filaments) that, in cooperation with the TTSS, form a molecular syringe between the bacteria and the host cell [110]. The secreted protein EscF, a protein required for the EspA filament assembly, differs only in six nucleotides that do not influence the amino-acid sequence between DEC9D and RW2070. In contrast within the nucleotide sequences of the secreted protein EspD, which plays a major role in the pore complex formation in the host cell membrane at the tip of the of the molecular syringe, 77 synonymous and 86 non-synonymous SNP-sites are located, leading to an identity of 90.5 % of the genes in the compared strains. EscD is an inner membrane component and necessary for the secretion of translocators and effectors. Within the compared strains *escD* differs in 37 synonymous and 16 non-synonymous SNP-sites, therefore both strains share an identity of 95.5 % of the nucleotide sequence of this gene. The *sepZ* gene includes even more non-synonymous SNPs (n=71) and five synonymous SNPs, leading to an identity of the genes in DEC9D and RW2070 of just 81.7 %. Further, *ler*, the LEE regulator, which activates gene expression of the LEE [111], is found to include 17 synonymous and three non-synonymous SNP-sites. Besides the described secreted proteins and further regulators of the LEE, also chaperone proteins such as CesT and CesD, are found be to locations of several SNP-sites. Besides the genes located on the LEE PAI, also metabolic genes include SNP-sites. These are, among others, *hisC*, *hisF* and *hisG* that are involved in the biosynthesis of the essential amino acid histidine. The gene *hisC* includes 139 synonymous and 58 non-synonymous SNP-sites, resulting in an identity of 87.5 % of the gene in both strains, and 79 synonymous and 11 non-synonymous SNP-sites with a resulting identity of 88,2 % for *hisF*. Furthermore genes that encode for the assembly of structural parts of the bacterial cell, i.e. flagellin, and the capsular polysaccharide are also affected by SNP-sites. The genes *fliD,* that is involved in the flagellin assembly, and *fliZ*, that is an expected activator for class 2 type flagellar operons, have an identity of 98.2 % and 98.4 % in both analyzed strains. The genes *wcaD* and *wcaF* that are involved in the production of extrapolysaccharide colanic acid, which confers protection against desiccation, heat, acid, osmotic and oxidative stress to cell and plays a role in biofilm formation, are locations of several SNP-sites [112]. Thus, these genes show an identity of 87.7 % and 81.6 % in the compared MCGs of both strains. This is also true for the genes *wza* and *wzb*, which are located on the same operon as

wcaD and are also involved in the colanic acid biosynthesis for the capsular formation. These genes include less SNPs and thus have a gene identity of 97.2 % and 97.6 % in the compared MCG of the strains DEC9D and RW2070.

Gene	Product	length [bp]	Identity [%]	synSNPs	nonsynSNPs
BX75_08170	Type III secretion protein	321	95	13	4
cesD	CesD protein	457	93,9	20	6
cesD2	T3SS chaperone CesD2	408	97,8	7	2
cesT	CesT protein	465	95,4	1	20
ECO9942_15685	T3SS component	429	94,1	28	10
escD	Type III secretory protein EscD	1.183	95,5	37	16
escF	T3SS structure protein EscF	222	97,3	6	0
espA	Secretion protein EspA	578	90,5	20	35
espD	Secretion protein EspD	1.143	90,5	77	86
grlR	Negative regulator GrlR	351	97,3	2	12
ler	LEE encoded regulator	373	96,4	17	3
mpc	T3SS regulator Mpc	357	94,4	25	6
sepZ	SepZ protein	302	81,7	5	71

Table 4 Number of synonymous SNP-sites (synSNPs) and non-synonymous SNP-sites (nonsynSNPs) in 13 genes of the LEE PAI that are part of the MCG and that differ between DEC9A and RW2070 in ≥5 SNPs. The gene name, protein product, length of the gene and the identity of the nucleotide sequence of the genes are shown.

The alignment of the MCG of DEC9D (ST29, O26:H11, aEPEC, human) and LH-1 (ST29, O26:NM, aEPEC, human) had a total of 1,767 SNPs of which 979 SNPs, located in 62 genes, were further analyzed (Table S 6, p.76). In contrast to the previous comparison between DEC9D and RW2070 no gene located on the LEE PAI harbored more than five SNP-sites. Yet, six genes, involved in the assembly of fimbrial structures, were identified as locations of serval SNP-site. Thus *fimA*, the major subunit of *E. coli* type 1 frimbriae, had 13 synonymous and ten non-synonymous SNP-sites. The gene *fimC* that is involved in the biogenesis of the type 1 pilus fiber as it binds to and forms stable complexes with the other subunits, includes seven synonymous SNP-sites. The gene *fimE* encoding for the fimbrial regulator that regulates *fimA*, has six synonymous SNPs, and the gene *fimG*, an adaptor component of the short tip fibrillum, includes four synonymous and two non-synonymous SNP-sites. The gene *fimH* encoding for the adhesin that is attached at the distal end of the pilus, includes 13 synonymous and three non-synonymous SNP-sites. Furthermore the Quorum sensing *E. coli* regulator D, *qseD*, that downregulates the flagellum regulon and decreases motility has eight synonymous and one non-synonymous SNP-site(s). The other genes of the MCG alignment of DEC9D and LH-1 that are affected by SNP-sites are mainly metabolic genes, such as *deoA*, *deoB* and *deoC*, which are necessary for the growth of *E. coli* on deoxythymidine and deoxyuridine as sole carbon source, and include almost only synonymous SNP-sites. Additionally, the gene *yjjG*, which encodeds for a dUMP phosphatase generates deoxyuridine, which in turn can be used by DeoA to liberate the deoxyribose-1-phosphate required for salvage of thymine, was found to include four synonymous SNP-sites. Furthermore seven genes involved in the 4-hydroxyphenylacetate degradation are found to locate several SNP-sites. 4-hydroxyphenylacetate is a common product of aromatic acid fermentation, and also produced in the animal intestine during the degradation of plant materials. The identified genes are *hpaB* and *hpaC* and are located on one operon, and *hpaD*, *hpaE*,

and *hapH* that are located on a second operon. Furthermore *hpaX*, that transports 4-hydroxyphenylacetate, and *hpcE*, also involved in the 4-hydroxyphenylacetate degradation cycle are found to include ≥5 SNP-sites. Two genes involved in the degradation of D-fructuronate, namely *uxuA* and *uxuB*, and *uxuR*, a transcriptional factor that represses the transcription of the cluster of operons including *uxuA* and *uxuB*, shows several SNP-sites. The metabolic genes *creB*, a transcriptional regulator, that induces *creD*, which is responsible for the tolerance to Colicin E2, differ both, due to SNP-sites, in the analyzed MCGs of DEC9D and LH-1.

The third alignment of the MCG, those of the strains RW2070 (ST16, O111:NM, aEPEC, cattle) and DEC8C (ST21, O111:NM, EHEC, cattle), was analyzed in greater detail The MCGs of both strains differed in 3,916 nucleotide positions of which 2,271 SNP-sites located in 115 genes were further analyzed (Table S 7, p.77). In the MCG alignment of both strains twelve genes of the LEE PAI were found to include more than five SNP-sites (Table 5). The twelve identified LEE genes are the same that had already been identified in the MCG alignment of DEC9D and RW2070, hence the function and relevance of these genes is already described in the previous section of this chapter. The gene encoding for the negative regulator GrlR was not identified in the alignment of RW2070 and DEC8C to include more than five SNPs.

Genename	Product	length [bp]	Identity [%]	synSNPs	nonsynSNPs
BX75_08170	Type III secretion protein	321	96.9	11	4
cesD	CesD protein	457	93.9	20	8
cesD2	T3SS chaperone CesD2	408	98.4	7	3
cesT	CesT protein	467	95.3	19	3
ECO9942_15685	T3SS component	429	94.1	28	10
escD	Type III secretory protein EscD	1.182	97	33	12
escF	T3SS structure protein EscF	222	97.3	6	0
espA	Secretion protein EspA	579	90.5	20	35
espD	Secretion protein EspD	1.143	90.4	74	86
ler	LEE encoded regulator	373	96.4	17	3
mpc	T3SS regulator Mpc	355	94.1	24	6
sepZ	SepZ protein	303	82	16	66

Table 5 Number of synonymous SNP-sites (synSNPs) and non-synonymous SNP-sites (nonsynSNPs) in twelve genes of the LEE PAI that are part of the MCG and differ between RW2070 and DEC8C. The gene name, protein product, length of the gene and the identity of the nucleotide sequence of the genes are shown.

Comparable to the MCG alignment of DEC9D and RW2070 genes involved in the flagellar biosynthesis were identified to include several SNP-sites in the MCG alignment of RW2070 and DEC8C, namely *fliD*, including three synonymous and eight non-synonymous SNP-sites (98.2 % identity) and *fliZ,* including five synonymous and four non-synonymous SNP-sites (98.4 % identity). Additionally, the genes for the biogenesis of fimbriae, described in the MCG comparison of DEC9D and LH-1, were also identified in the MCG alignment of RW2070 and DEC8C as locations of numerous SNP-sites. These genes are *fimA*, including 13 synonymous and ten non-synonymous SNP-sites (97.1 % identity), *fimC*, with five synonymous and two non-synonymous SNP-sites (99.4 % identity), *fimE*, including six synonymous SNP-sites (98.9 % identity), *fimG*, with four synonymous and two non-synonymous SNP-sites (98.8 % identity), and *fimH*, with 13 synonymous SNP-sites (98.2 %

identity). The Quorum sensing *E. coli* regulator D, *qseD*, was also found to include six synonymous and nine non-synonymous SNP-sites (98.9 % identity). Additionally, the genes *yehA* and *yehC*, that are located on one operon, and encode for type 1 fimbrial proteins which promote biofilm formation, are identified with several SNP-sites, respectively *yehA* includes 14 synonymous and nine non-synonymous SNP-sites (98.5 % identity) and *yehC* includes ten synonymous SNP-sites (98.5 % identity). Metabolic genes as for the degradation of 4-hydroxyphenylacetate, which were already described for the MCG alignment of DEC9D and LH-1, were also found to include several SNP-sites in the alignment for RW2070 and DEC8C. Those genes are *hpaB*, *hpaC*, *hpaD*, *hpaE*, *hpaH*, *hapX* and *hpcE*. The same is true for the previously described genes of the MCG comparison of DEC9D and LH-1, involved in the degradation of D-fructuronate (*uxuA*, *uxuB*, and *uxuR)* and for the growth of *E. coli* on deoxythymidine and deoxyuridine as sole carbon source *(deoA*, *deoB* and *deoC*) which were again found to include several nucleotide differences in the alignment of RW2070 and DEC8C. Furthermore, two genes encoding for enzymes involved in the adenosyl-cobalamin salvage from cobinamide, namely *cobT* and *cobU*, were found to include SNP-sites: *cob*T had 14 synonymous and six non-synonymous SNP-sites (98.1 % identity), while *cobU* included seven synonymous and three non-synonymous SNP-sites (98.8 % identity). The two regulators of the *gal* regulon expression GalS and GalR, that are responsible for the transport and catabolism of D-galactose and are encoded by *galS* and *galR*, were affected by SNP-sites in the MCG alignment of the analyzed strains. The gene *galS* included seven synonymous and two non-synonymous SNP-sites (99.2 % identity) and *galR* had six synonymous and seven non-synonymous SNP-sites (97.4 % identity). Besides these metabolic genes, genes encoding for multidrug efflux pumps were found to include SNP-sites. Those genes are *mdtC*, encoding for a subunit of the MdtABC Resistance Nodulation and cell Division (RND)-type drug exporter, which includes twelve synonymous SNP-sites (99.1 % identity), *mdtM*, encoding a multidrug efflux protein that belongs to the major facilitator superfamily (MFS) of transporters, which has 17 synonymous and two non-synonymous SNP-sites (99.0 % identity) as well as *mdtP*, enconding for a predicted outer membrane factor family component of a certain multidrug efflux pump involved in sulfur drug resistance, that includes five non-synonymous SNP-sites (99.8 % identity).

In summary, most SNP-sites in the MCG alignment of the strains DEC9D of Cluster 2 and RW2070 of Cluster 1 are located in genes that are involved in pathogenicity, in fact those of the LEE PAI, genes that encode for proteins of the flagellar assembly and colanic acid biosynthesis for the capsular formation, as well as genes involved in histidine biosynthesis. For the comparison of the MCG DEC9D of Cluster 2 and LH-1 of Cluster 3 most SNP-sites were located in genes involved in the biogenesis of type 1 fimbriae and metabolic genes encoding for enzymes for 4-hydroxyphenylacetate and D-fructuronate degradation, and genes which are necessary for growth on deoxythymidine or deoxyuridine and regulators for the tolerance to Colicin E2. The comparison of two O111 strains, RW2070 of Cluster 1 and DEC8C of Cluster 3, revealed most SNP-sites in the MCG alignment in genes encoding for secreted proteins and regulators of the LEE PAI, flagellar assembly, metabolic genes encoding for enzymes for 4-hydroxyphenylacetate and D-fructuronate degradation, adenosyl-cobalamin salvage from cobinamide and enzymes necessary for growth on deoxythymidine or deoxyuridine, regulators of the transport and catabolism of D-galactose, as well as multidrug efflux pump proteins. The genes of all three strains MCG comparisons were explained in further detail, as

they were found to be located on one PAI or operon or involved in the same metabolic pathway. Nevertheless, various other genes were found to include more than five SNP-sites, but were not connected to other SNP-site locating genes, and thus were not described in this chapter although they might as well justify the location of the strains in a certain cluster.

3.4.2 Insertion sites of mobile genetic elements of selected strains of the analyzed STC29 population

Twelve aEPEC and STEC/ EHEC strains of STC29 located in prominent position in the MSTree of the SNP-analysis (Figure 15) were chosen for deeper analysis of the integration sites of bacteriophages, integrative elements (IE) and PAI. The selected strains are marked in the MSTree of the SNP-analysis in Figure 15. A number of 63 genes that are known to be sites of recombination in *E. coli* were mapped to the WGSs of the selected strains. The surrounding genetic region in the annotated nucleotide sequence was screened for mobile genetic elements. If such a mobile genetic element was identified, the respective sequence was used for a blastn search in the nucleotide database of ncbi to identify the mobile genetic element according to a reference sequence (performed between June and October 2015). Furthermore, the phage search tool 'PHAST' was used to screen the WGSs for phage presence. The resulting phage sequences, that were already identified with the best matching reference sequence given by the online tool, were mapped against the genome sequence of the respective strains and used to identify the insertion site of the phage (Table S 9 to Table S 20, p.79-84). In the next step, the collected data on the occupation of each of the 63 insertion site for each of the analyzed genomes was then used to generate MSTrees. The mobile genetic elements were grouped into phages and all other genetic elements. On basis of the results a dendrogram of the population was built (Figure 16). In this dendrogram the insertion of phage genes in the respective integration site are marked in black and insertion of PAI and IE are marked in grey; white fields indicate that no insertion was found in this genetic position.

The resulting dendrogram shows that the aEPEC strains of Cluster 2, previously detected within the MSTree of the SNP-analysis and designated as ST29 (marked in red), namely DEC9B, IMT19623, LH-8, DEC9E, DEC9D, DEC9A, and DEC9C form a distinct branch. Within the second larger branch, RW2070, the only included strain of the ST16 Cluster 1, is located, but yet distinct from the rest of the strains within this branch. Those strains belonged to Cluster 3 based on the SNP-analysis. These strains are LH-1, an ST29 aEPEC strain located in Cluster 3 that is closest related to 0739/03, an *stx2a* positive strain of ST29. 0739/03 shows the closest connection to the branch formed by the two strains of ST21: IMT20337, an stx_{2a} positive strain, and IMT19981, an aEPEC strain. The dendrogram reflects the results of the SNP-analysis and the VAGs analysis. Therefore these strains can be used to build a model of the evolutionary scenario of STC29 strains.

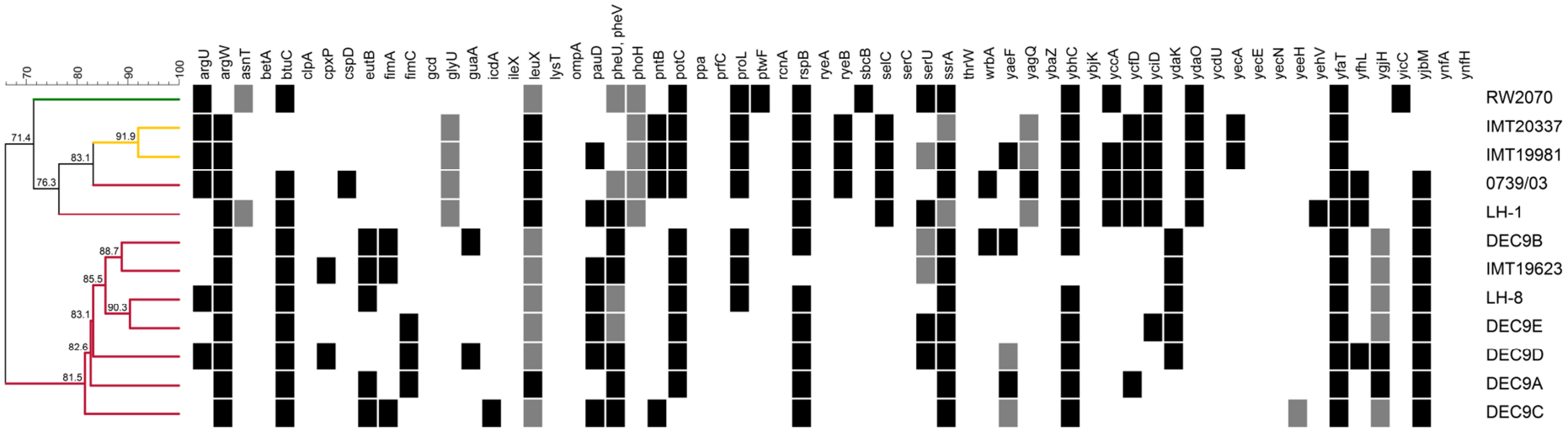

Figure 16 Dendrogram of the twelve selected strains of the analyzed STC29 population of aEPEC and STEC/ EHEC genome sequences generated on basis of the occupation of 63 insertion sites with Bionumerics (v7.5, Cluster analysis with a categorical (values) similarity coefficient, using UPGMA as calculation method). The 63 insertion sites are given on top of each column. The grey fields mark an occupation of the insertion site with a PAI or IE and the black fields mark insertion of phage genes. White fields indicate that no insertion element was detected in this position. The colors of the branches indicate the ST the strain belongs to: red marks strains of ST29, yellow strains of ST21 and green strains of ST16.

3.5 Evolutionary model of the analyzed STC29 population

The strains of STC29 that have already been analyzed in depth in the previous chapters were now further utilized to develop a model of the microevolution of these strains. The occupation of insertion sites of the twelve selected strains can be additionally visualized as an MSTree to nicely display the relation of these strains (Figure 17).

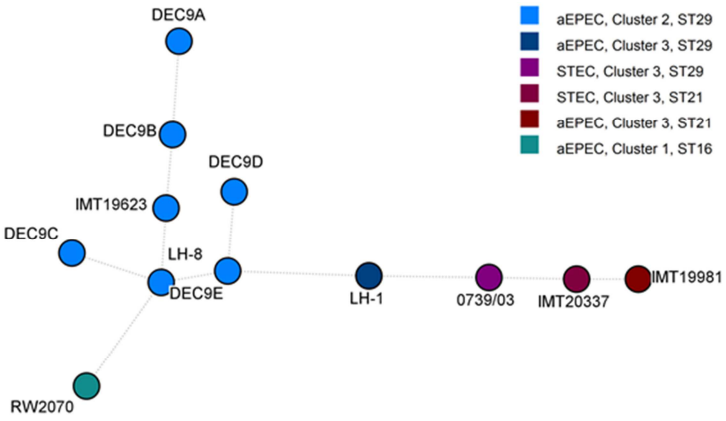

Figure 17 MSTree of twelve selected strains of the analyzed STC29 population of aEPEC and STEC/ EHEC genome sequences generated on basis of the occupation of 63 insertion sites. The single strains are labeled and the color indicates their pathotypes and cluster location in the SNP-based MSTree.

The MSTree (Figure 17) indicated again that the aEPEC strains of Cluster 2, which are once more located between Cluster 1 and Cluster 3 when analyzed for the occupation of their insertion sites, play a major role in the evolution of the other strains and might therefore represent the founding strain of STC29 strains. DEC9D will be chosen a representative for these aEPEC strains of Cluster 2 as it is in a central position in the SNP-analysis based MSTree (Figure 15). The strains further included into the model represent one point in the microevolutionary scenario in the development of strains of STC29. The LEE PAI of these strains is always found to be integrated into the tRNA encoding gene *pheV*, except for IMT20337 and IMT19981. The insertion site of the *stx*-converting bacteriophage in the *stx* positive strains was also determined. The stx_{2a}-phage in 0739/03 (IMT15915) is integrated into *wrbA*, while in IMT20337 it is integrated into *yciD,* and *wrbA* does not have any inserted mobile genetic elements in this strain. IMT19981 is enclosed into this model as a strain that might have lost the *stx*-phage, but is could become a lysogen again when transduced with an *stx*-phage. As all other strains in Cluster 1 (O111, ST16), except of one strain that harbors an stx_{2a}-phage, harbor an stx_{1a}-convertering phage, one stx_{1a} positive strain is included into the final model as a representative for these strains. This strain, 06-03988 (O111:NM, human), has the stx_{1a}-phage integrated into the gene *ssrA*. The final schematic model (Figure 18) gives an overview of the microevolutionary scenario of STC29 strains based on the analyses described above. Two separated lineages developed from a common ancestor, an aEPEC strain of ST29 (A1). One lineage had a change of the O-antigen cluster from O26 to O111, as well as change of the ST due to an allelic change in the *recA* gene from ST29 to ST16 (C1). This lineage developed further by integration of stx_{1a}-convertering phage into its genome (C2). Other strains might develop from the initial aEPEC strain of ST16, due to integration of additional stx_{2a}-phages. The second lineage that developed from the initial ST29 aEPEC strain due genomic

changes, i.e. integration and loss of mobile genetic elements, the acquisition of other VAGs, like *efa1* and *iha* that are involved for adherence, and changes in metabolic genes, led to 1,767 SNPs difference between the MCG of A1 and the next evolutionary step A2. In the next step an stx_{2a}-convertering phage is integrated into the genome (A3), and then the ST is changed from ST29 to ST21 due an allelic change in the gene *adk* (B1). The step B2 is added to symbolize the possible loss of the *stx*-phage that leads to a pathotypes change to aEPEC and illustrates the bidirectional conversion of these strains.

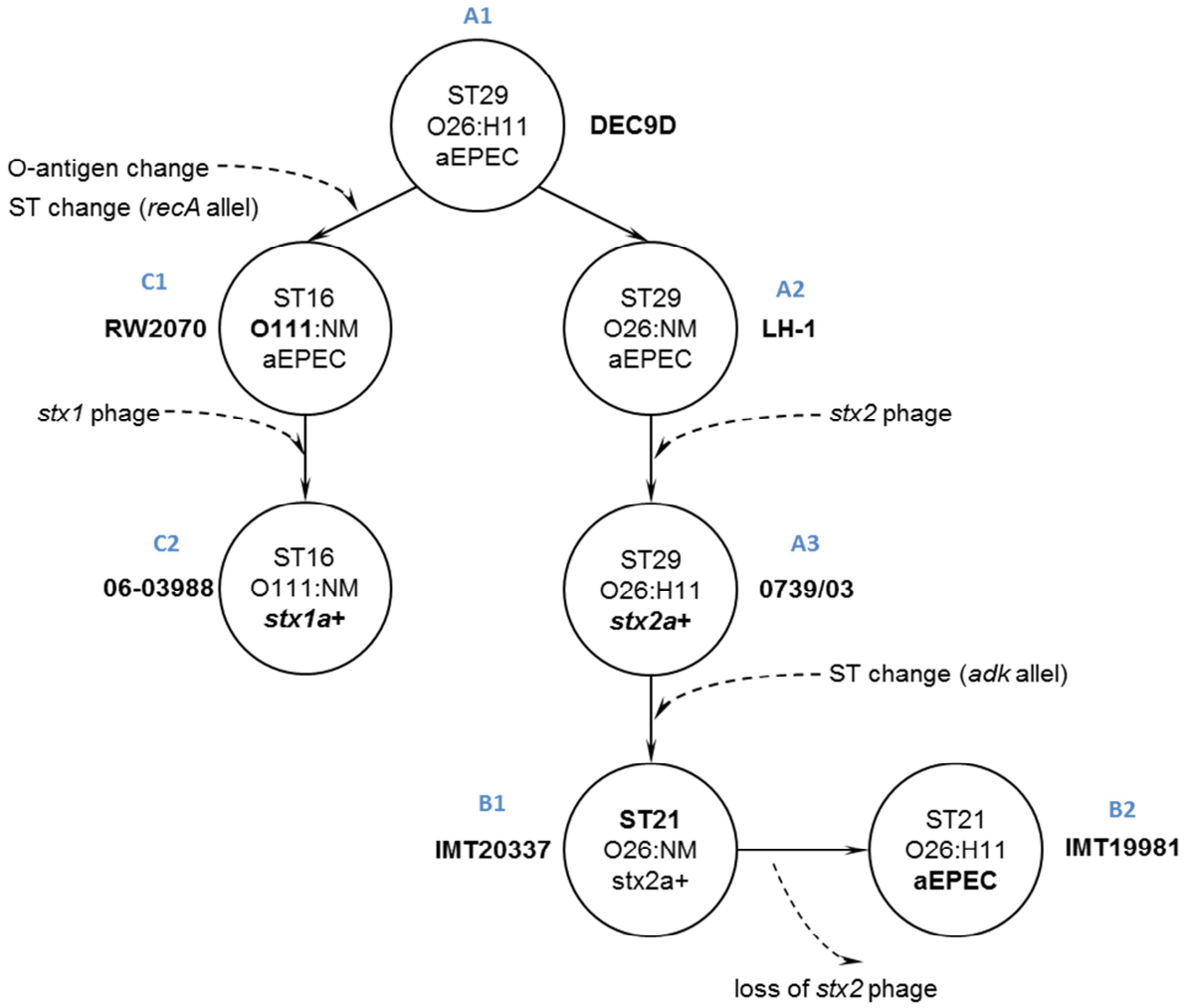

Figure 18 Proposed microevolutionary model for the emergence of stx_{1a}-positive strains of O-serogroup O111 and stx_{2a}-positive strains of O-serogroup O26 from a common aEPEC ancestor based on MLST, SNP-analysis, absence and presence of VAGs and occupation of insertion sites. The characteristics of each evolutionary step are shown in the circles. The associated reference strain for each evolutionary step is given in bold letters next to each step.

4 DISCUSSION

4.1 Development of a microevolutionary model of the most crucial non-O157 *E. coli*

Shiga-Toxin producing *E. coli* of non-O157 serotypes are causing diarrheal diseases and clinical symptoms that are comparable to those of O157:H7. The O-serogroups that occur most frequently are O26, O111, O103 and O145, and are also responsible for one-third of the non-O157:H7 infections in humans in Germany and outbreaks worldwide [28, 39, 113-115]. Once it had been recognized that non-O157 serotypes of *E. coli* were also involved in human disease, the detection rate of such strains increased, due to either a higher awareness of the respective laboratories, or the development of better detection methods, but possibly also due to a higher incidence of human and animal infections with these pathogens [56, 63, 116].

In this thesis the underlying population structure of STEC/ EHEC strains of the most important non-O157 serotypes was analyzed for a large *E. coli* collection of diverse strains using innovative typing methods based on whole genome sequences which enabled us to establish a microevolutionary model of these highly relevant non-O157 STEC/ EHEC. The importance and improvement of this model compared to previously published studies, based on comparatively narrow approaches, are the following:

First, we used a large set of diverse *E. coli* strains, including *E. coli* strains of more than one serotype, whereas other studies focused largely on one specific serotype, such as O26:H11/NM [117, 118], O111 [119], O103:H2 [28, 120] or O145:H28 [115, 121]. In addition, in the strains examined in this thesis were isolated from humans as well as cattle, the most common host of STEC/ EHEC, rather than separately analyzing strains isolated from one single host only [75, 113, 122, 123].

Second, we have included aEPEC, a closely related pathotype of STEC/ EHEC, which enabled us to analyze the relatedness of both pathotypes, while other research groups have mainly focused on one of these pathotypes only.

Third, we performed multiple analyses on these strains including MLST, SNPs in the MCG of those strains, their VAG profiles and the integration of mobile genetic elements to describe the population structure in greater detail.

Due to this multi-analyses approach we were able to develop a microevolutionary model of the most important non-O157 *E. coli*. In the evolutionary-model each proposed evolutionary step is represented by at least one biological strain and thus gives sound evidence for the existence of each of the respective steps.

In the proposed model an O26 aEPEC strain of ST29 was determined as the common ancestor of STEC/ EHEC strains of O-serogroups O26 and O111, considering the aEPEC strains of ST29 that constituted the central cluster, Cluster 2, in the MSTree resulting from the SNP-analysis. These data are corroborated by other studies proposing that O26:H11 and O111:H11 *E. coli* evolved from a common ancestral strain by an antigenic shift from O111 to O26 [124-126]. However, this is the first study that identifies a valid common ancestor of both lineages. Selecting an aEPEC strain as the ancestral progenitor of O26 and O111 STEC/ EHEC is further reasonable as they can be converted via lysogenic conversion with *stx*-converting bacteriophages that integrate into the genomes of aEPEC

and thus gain the ability to produce Shiga-Toxin. The possibility of *stx*-phage to lysogenize *E. coli* strains *in vivo* and *in vitro* has been shown by several research groups and is further discussed in Chapter 4.2 [127-130].

The development of two lineages from a common ancestor of O serogroup O26 appears to have occurred due to recombination events in the O-antigen-coding regions and explains the differences in the somatic antigen of both lineages. Such a lateral gene transfer of the *rfb*-like region is known to have caused the evolution of O157:H7 EHEC strains from a common O55:H7-like ancestor [95]. The O-antigen is an integral part of the lipopolysaccharide (LPS) in the outer membrane of gram-negative bacteria. Although the LPS is a known endotoxin and therefore virulence factor of *E. coli*, it does not seem to have much influence on the pathogenicity of the aEPEC and STEC/ EHEC of O-serogroups O26 and O111. Both O-antigens shown only a 45 % identity on nucleotide level and are not mentioned as somehow similar in the *Escherichia coli* O-antigen Database (EcoDAB, http://nevyn.organ.su.se/ECODAB/). Nevertheless the strains with such different O-antigens share a phylogenetic background that seems to have developed without much influence of either O-antigen as a pathogenicity factor.

The next step in the proposed model in the evolution of O26 STEC/ EHEC is suggested by aEPEC of ST29 that cluster together with STEC/ EHEC strains of ST29 and ST21 in Cluster 3. This step is suggested by the observation that they differ in a greater number of SNP sites (n=1,767) compared to the nearest ST29 aEPEC neighbor of Cluster 1, and therefore likely represent the next step in the evolutionary scenario, prior to the integration of an stx_{2a}-phage into their genome. Lysogenic conversion with an *stx*-phage leads to the development of an intermediate evolutionary step, represented by ST29 STEC/ EHEC of Cluster 3. Cluster 3 appears to be more heterogeneous than the other two clusters in that it is formed by strains designated as ST29, or its SLVs, and ST21, or its SLVs, respectively. In conclusion, these strains of ST29 are different to those of Cluster 2 and are more related to those of ST21. Nevertheless, they are separate and more related to one another than to the other strains of ST21 and its SLVs. Therefore we propose them representing an intermediate stage in the evolution of O26 STEC/ EHEC of ST21.

These stx_{2a}-positive strains of ST29 were identified as newly emerging human pathogens in Germany [117, 131]. Bielaszewska *et al.* (2013) suggested that the VAG-profile of four genes and the MLST results indicate two lineages of O26 EHEC: first, strains of ST29 that are stx_{2a} positive and contain *hylA* and *etpD*, but not *espP* and *katP* and belong to the new emerging clone in Germany, and second, the strains of ST21, that differ in their VAG-profile and phylogeny from the ST29 strains [117]. These findings match our VAG analysis and MLST results and are in accordance with the SNP-analysis. They support the hypothesis that ST29 strains of Cluster 3 represent an important intermediate state in the evolution of STEC/ EHEC strains of O-serogroup O26. It should be noted that Bielaszewska *et al.* did not include aEPEC in their study, but in this thesis the indicated virulence profile was also identified in ST29 aEPEC, and verifies the close relationship of aEPEC and STEC/ EHEC.

The concordance of the resulting population structure represented by SNP- and VAG-analysis indicates a parallel evolution of the MCG and the VAGs acquired by these bacteria. The model of parallel evolution of virulence has been suggested by Reid *et al.* (2000) in an analysis of

house-keeping genes and VAGs of pathogenic *E. coli* [132]. The authors found that older *E. coli* lineages had acquired the same virulence factors in parallel, including the LEE PAI, the plasmid-encoded hemolysin, and phage-encoded Shiga-Toxins. They concluded that such parallel evolution indicates a selective advantage that has favored an ordered acquisition of VAGs and results in a progressive build-up of molecular mechanisms that increase virulence leading to the emergence of new virulent clones. Furthermore the authors hypothesized that molecular interactions between mobile genetic elements, which also may encode VAGs, might lead to an ordered progression of change [132]. In our analyses the population structure that results from changes in the MCG, as seen in the presence of SNPs, is mirrored by the presence and absence of certain VAGs. In conclusion, the MCG of these *E. coli* strains developed in parallel to the acquisition of VAGs.

In order to account for the possible loss of the *stx*-phage by STEC/ EHEC strains of O-serogroup O26, as proposed by Karch *et al.*, who designated these aEPEC strains as EHEC-LST [133] (further discussed in Chapter 4.2), such an evolutionary step also has been integrated into the proposed model.

In the second lineage developing from the common ST29 aEPEC ancestor, the next evolutionary step is represented by an aEPEC of O-serogroup O111, which is characterized as ST16. The association of O111:H8 and O111:NM *E. coli* strains with ST16 was already detected during the compilation of the HUSEC collection and the analysis of non-O157 EHEC in the Czech Republic [62, 134]. Remarkably only one O111 aEPEC strain was identified among the examined population, probably due to the presence of a cryptic prophage in the genome of most O111 STEC/ EHEC strains. This prophage is deficient in excision from the bacterial genome and hence not capable of lysis of the host bacteria and the production of new phage particles [135]. Accordingly, the next step in the evolutionary scenario of the development of O111 STEC/ EHEC is the integration of an stx_{1a}-converting bacteriophage into the genome of the O111:NM aEPEC.

The last line of evidence that the proposed model is correct and accurately mirrors the evolutionary scenario of O26 and O111 STEC/ EHEC strains of STC29 and most importantly the determined common ancestor of both lineages, O26 aEPEC of ST29, would be a time-scale analysis to investigate the temporal course of the separation into the two lineages. One indication that the aEPEC strains of ST29 actually represent the common ancestor of both lineages is their date of isolation. The DEC9 strains of the aEPEC cluster are the ‘oldest’ strains in the analyzed population having been isolated between the years 1952 and 1979 (for comparison: most strains of the analyzed population were isolated in the 2000s (n=56)).

An evolutionary model of O26 EHEC created by Bletz *et al.* (2013) based on 48 selected SNPs described the development of four SNP clonal complexes (SNP-CC) of O26 EHEC. These authors proposed an initial development of an O111 EHEC (the whole genome sequenced reference strain O111:NM str. 11128, NC_013364.1) and a first evolutionary step of the O26 EHEC from an unknown common ancestor. However, again these authors also did not include aEPEC in their analyses. The analyses performed in this thesis provide evidence that the identified central cluster of aEPEC of ST29 represents such a common O26/O111 ancestor. Furthermore SNP-CC1 and SNP-CC2, as described by Bletz *et al.*, are composed of EHEC of ST29 or ST396, a SLV of ST29, and only in SNP-CC3 a change of the ST occurs from ST29 to ST21. The last evolutionary step, SNP-CC4, is composed of

EHEC of ST21 harboring either *stx*$_{1,}$ or both *stx*$_1$ and *stx*$_2$ in combination. The study of Bletz *et al.* [118] and the work in this thesis correspond nicely. However, the inclusion of aEPEC in this thesis provides a much better insight into the possible underlying evolution of STEC/ EHEC strains of O-serogroup O26 as both pathotypes are highly related. In addition, analyzing all occurring SNPs of the MCG rather than a few selected ones results in a much higher resolution of the population structure. Another advantage of our analyses incorporating an aEPEC and STEC/ EHEC population is the inclusion of bovine as well as human strains. In the resulting MSTree, where the host origin of each strain is highlighted, no cluster formation was observed. Therefore, regardless of the host origin, the strains share highly related SNP-profiles. The lack of host-specific SNPs provides strong evidence of the strains being *bona fide* zoonotic agents. This finding also indicates the zoonotic risk of aEPEC and STEC/ EHEC of O-serogroups O26 and O111, which can be transmitted from cattle to human or *vice versa*. In regard to the evolutionary scenario this also means, that the development and emergence of certain O26 strains is not occurring due to adaptation within a stable niche over a long time period.

4.2 Importance of aEPEC within the population analysis of STC29 *E. coli*

The classification of aEPEC and STEC into different pathotypes relies merely on the presence or absence of an *stx*-converting bacteriophage in the genome of STEC strains that enables these strains to produce the cytotoxic as well as apoptotic Shiga-Toxin. Apart from this *stx*-phage which is not harbored by aEPEC strains, their virulence characteristics, such as the PAI LEE and the resulting ability for formation of A/E lesions on epithelial cells, as well as serotypes are common to both STEC and aEPEC strains. In contrast to typical EPEC, STEC and aEPEC do not possess the EAF plasmid harboring the *bfp* gene, and thus do not produce the bundle forming pilus that mediates a localized adherence on cultured epithelial cells. Based on these common features, in this thesis, aEPEC of serotypes O26:H11 and O26:NM were included in the MLST analysis. *E. coli* strains belonging to this serotype are EHEC and aEPEC. In addition, O26:H11/NM EHEC are the most common non-O157 EHEC that are associated with HUS in humans. O26 aEPEC are known to cause mainly diarrhea, but have also been isolated from HUS patients [78]. In contrast, aEPEC of serotype O26:H11 have also been isolated from stool samples of healthy infants in Germany, although with a low frequency.

Mellmann *et al.* (2005) identified O26:H11/NM aEPEC in stool samples of HUS patients and concluded that these strains originate from EHEC strains of the same serotypes that have lost the *stx*-converting bacteriophage. They hypothesized that loss of the *stx*-phage may be a strategy for better survival in the host [78]. The *stx* gene is encoded on a lambdoid bacteriophage which is integrated into the bacterial chromosome. The temperate phage can switch into the lytic life cycle when its bacterial host undergoes stress, e.g. exposure to UV-light, sublethal doses of mitomycin C or antimicrobial substances among others. In response to such stress conditions, the bacterial host activates an SOS response which also results in the activation of the phage encoded anti-terminator Q. The anti-terminator Q is required for expression of genes involved in lysis of the bacterial cell and the excision of the bacteriophage from the bacterial genome and results in the production of new infective phage particles and Shiga-Toxin.

Hence, phage induction results in a number of events affecting both the bacteria and host: First, antibiotic treatment of patients infected with EHEC can result in an even higher Shiga-Toxin production and thus can lead to a more severe disease outcome [136, 137]. Second, a bacterial cell harboring a lambdoid bacteriophage is prone to lysis, and therefore the loss of the *stx*-converting bacteriophage might lead to a higher chance of survival and offer a selective advantage for the bacteria. Third, the lytic life cycle of the bacteriophage and the lysis of bacterial host results in the release of infective phage particles into the surrounding milieu which are able to infect other bacteria and even lysogenize them by integrating into the bacterial genome. Both, the loss of the *stx*-converting bacteriophage and the lysogenic conversion due to phage integration are mechanisms of bacterial evolution, and have been described as bidirectional conversion in various publications. Based on these observations, the O26 aEPEC isolated from patients with HUS were designated by Mellmann *et al.* (2005) as EHEC-LST, EHEC that have lost the *stx*-phage during infection [77].

An additional observation was that O26 EHEC-LST are isolated more frequently than those of O-serogroup O157, indicating that the loss of the *stx*-phage is associated with the serogroup rather than the time interval between initial and follow-up stool sampling. In addition, the authors identified EHEC-LST as the only pathogens in some patients with HUS and concluded that those strains might be more persistent than the *stx* positive strains [78]. In addition to the idea that *stx* negative, *eae* positive strains are EHEC-LST, they might also be progenitors of typical EHEC, as they can also be lysogenized by an *stx*-converting bacteriophage. The fact that transduction with *stx*-converting bacteriophage can occur *in viv*o was shown by Acheson *et al.* (1998). These authors infected mice with a derivate of an stx_1-converting phage that had a deletion of the *stx* gene which had been replaced with an antibiotic-resistance gene in order to identify the resulting antibiotic-resistant lysogens. Transductants that were recovered from the stool samples of these mice provided evidence for intraintestinal phage production by lysogens [138]. Comparable studies have shown the lysogenic conversion of commensal *E. coli* in the gastrointestinal tract of sheep [129] and of laboratory *E. coli*, as well as porcine O45 EPEC *in viv*o in piglets [139]. However, in the latter study, they were not able to show the lysogenic conversion *in vitro*, and therefore hypothesized that the *in vivo* conditions might favor transduction of *stx*-converting phages and lysogenic conversion of wild-type *E. coli*. The production of lysogens *in vivo* is dependent on the presence of phage susceptible bacteria in the intestine [128]. Furthermore, it was found that 10 % of humans intestinal *E. coli* are susceptible to *stx*-converting bacteriophages [130].

The lysogenic conversion of *E. coli* strains *in vitro* was shown with stx_2-converting bacteriophage, and to a lesser extent, with stx_1-converting bacteriophage from EHEC O26 that were able to transduce aEPEC of O-serogroup O26 [79]. Schmidt *et al.* (1999) used a derivative of an stx_2-converting bacteriophage from EHEC O157:H7 to transduce enteric *E. coli* wild-type strains of various serotypes, including O26:NM, and detected lysogenization with the bacteriophage in one-third of the examined wild-type strains. Additionally, they were able to transduce STEC strains that already harbored stx_1 and stx_2 genes, alone or in combination [127]. A single bacterial host cell can harbor multiple *stx*-converting bacteriophages, hence can be transduced with *stx*-phages multiple times, resulting in the presence of various *stx* genes [140].

In conclusion, previous studies of EHEC-LST, as well as *in vivo* and *in vitro* lysogenic conversion with *stx*-converting bacteriophages corroborate that our collection of aEPEC and STEC/ EHEC strains of STC29 represent a population likely prone to bidirectional conversion. In particular, the central Cluster 2, which is formed by aEPEC of ST29, is of interest for bidirectional conversion as it might represent either progenitors or EHC-LST of O111 STEC/ EHEC strains of Cluster 1 and O26 STEC/ EHEC strains of Cluster 3.

It is apparent that the aEPEC strains of Cluster 2 occur to have more different SNP-profiles than the strains of the other clusters, indicating a more distinct relation to one another. With regard to the total amount of SNPs found in the analyzed population, the aEPEC strains of Cluster 2 share more than 90 % of the SNPs. In fact between 92.6 % and 99.6 % of all SNPs are shared by these aEPEC strains. In comparison, Cluster 1 and 2 share only 74.3 % of all SNPs and Cluster 2 and 3 share just 85.5 % of the SNPs. In the already mentioned study by Bletz *et al.* (2013), the phylogeny of O26 EHEC strains was analyzed on the basis of 48 selected SNPs. SNP-clonal complexes were assigned to strains that shared ≥90 % of those 48 SNPs [118]. Hence, the designation of the aEPEC strains of ST29 as a distinct cluster is valid. Another noticeable fact of this Cluster 2 is the accumulation of five DECA strains that group together with just two other strains of dissimilar origin. In this case, a co-isolation of clonal strains can be excluded as those strains originate from different geographical origins and therefore are actually distinct strains. In contrast, the two strains of Cluster 3, sharing an identical SNP-profile, are presumably just two isolates of the same strain, thus probably belonging to the same identical clone.

The most important VAG for initial distinctive analysis of this population is the presence of the *stx* gene. The most frequently detected *stx* gene was stx_{1a}, alone or in combination with a stx_{2a} gene. Furthermore, with the exception of two strains, all strains of O-serogroup O111 that are designated as ST16 and form Cluster 1 possess stx_{1a}. Only one aEPEC was detected in this cluster. In general the stx_{1a} gene is the most prevalent in O111 EHEC [119] as it is encoded on a cryptic, immobilized prophage in the genome of O111 EHEC strains. This phage is not able to excise itself from the *E. coli* genome [135]. In contrast, the strains of O-serogroup O26 either do not possess any *stx* genes, as in the case of the aEPEC of Cluster 2, or harbor a stx_{2a} gene, as the ST29 strains in Cluster 3. The latter presumably gained their *stx* through lysogenic conversion. The strains of ST21 belonging to Cluster 3 either do not harbor an *stx* gene, or they harbor either *stx1* or *stx2* genes, or a combination of both. These findings correspond with those of the MLST analysis when the presence of *stx* genes was highlighted in the respective MSTree of STC29. These observations again strengthen the hypothesis that aEPEC of ST29 are likely to be the progenitor of other STEC/ EHEC strains of STC29, or EHEC-LST.

4.3 MLST analysis is more descriptive than serotyping

The MLST results of 250 non-O157 performed in this thesis show a clear phylogenetic relation of O-serogroups O26 and O111 as both cluster together in STC29. Within the 42 *E. coli* strains of the HUSEC-Collection established by Mellmann *et al.* (2008) most of the included strains (n=10, 23.8%) belong to STC29, including ST16, ST21, ST27, ST29 and ST396, and comprise EHEC strains of the

following serotypes: O111:H8/NM, O26:H11/NM, Orough:H11 [62]. The integration of STC29 EHEC into a reference collection underlines the importance of these non-O157 strains.

ST21 and ST29 have already been recognized by diverse research groups to represent STEC and EHEC of O-serogroup O26 [117, 131]. For STEC and EHEC strains of serotypes O111:H8/NM several studies identified ST16 as the most prominent ST [62, 134, 141], a finding that can be confirmed with our data showing the majority of all O111 *E. coli* strains clustering into ST16 and its SLVs.

As in the MLST analysis, the MSTree of the SNP-analysis revealed that the majority of O111 *E. coli* strains formed a distinct cluster, here designated as Cluster 1, and was further equivalent to ST16. The association of O111:H8 and O111:NM *E. coli* strains with ST16 was detected earlier during the compilation of the HUSEC collection and the analysis of non-O157 EHEC in the Czech Republic [62, 134]. In contrast, one strain of serotype O111:H11 and an adjunct strain of serotype O111:NM were located along Cluster 3 that was formed by strains designated as ST21 or its SLVs. This could be either due to the fact that H-antigens give more homogeneous representation of the relationship of those strains, as discussed later in this chapter. Hence, these strains are more related to other strains of this H-type than to others of different H-stypes in Cluster 1. On the other hand, the second strain of serotype O111:NM adjunct to the O111:H11 strain, might indicate that two distinct lineages of *E. coli* strains of O-serogroup O111 exist, and that the second O111 cluster was not detected as strains of serotype O111:H11 are underrepresented in the present analysis.

In addition to the cluster formation of O26 and O111 *E. coli* strains within STC29, only one strain of O-serogroup O103 was located with STC29, while the majority of these strains could be found within STC20, and most frequently were typed as ST17. Iguchi *et al.* (2012) detected a close relation of O103:H11 strains with STEC O26:H11 and suggested that those lineages evolved from a common ancestor with one or more exchanges in the O-antigen biosynthesis cluster. The O103:H11 STEC strains belonged to ST21 and its SLV ST723 [142]. In this thesis only one O103:H11 strain, which also clustered into STC29, namely ST723, was included. We suggest that if more strains of serotype O103:H11 would have been included into the MLST analysis additional strains would cluster into STC29 and would be allocated with strains of O-serogroup O26 in ST21. Iguchi *et al.* (2012) further found STEC of serotype O103:H25 assigned to ST343. Here, only one STEC strain of this serotype was included and accordingly typed as ST343. Strains of serotype O103:H2 were identified as ST17 (STC20), which matches the findings in the present thesis that the majority of the O103 STEC strains were of serotype O103:H2. In addition to O103:H2, strains of serotypes O103:H8 and O103:NM also clustered into ST17. O145:NM STEC strains have been shown to belong to ST32 or its SLV ST137. A HUSEC of serotype O145:H28 was also typed as ST32 [62, 134, 141], which matches with the MLST analysis performed in thesis which indicated cluster formation of O145:H28 strains in STC32, including ST31, ST137 and ST1096.

Taken together, the resulting MLST of 250 of the epidemiologically most important non-O157 O-serogroups O26, O103, O111 and O145 condenses the findings of several research groups about the regular STs in a context that has not been realized so far. The common classification of *E. coli* strains into O-serogroups is thought to be an indicator of distinct relationships of these strains; therefore, the O-serogroup would imply a detached phylogenetic lineage with differences in the biological characteristics of the respective strains. Hence, the usage of O-serogroups as descriptive

characteristics for the pathogenicity of STEC and EHEC strains may result in the hypothesis that they may represent distinct phylogenetic lineages, although in fact the analyzed strains might be polyphyletic.

In contrast to the polyphyletic O-serogroups, the flagellar (H) antigens have been described as monophyletic and thus thought to give a more robust representation of the phylogenetic relationship among the strains [125]. For the strains included in the present study, some H-types display a more distinct relationship than others. Strains of H-type H28 cluster together only with non-motile (NM) strains in STC32, whereas H-type H2 is found within different STs, namely ST1140, ST1099, ST9 and ST1790, with the majority of the latter strains with the flagellar H2 antigen forming STC20. Moreover, most STC, like STC20, include more than one H-type and therefore indicate a phylogenetic relationship that is not adequately resolved by analyzing the sole flagellar antigen. Those strains analyzed here that are distributed outside the major STCs were found to harbor various H-types (H10, H19, H11, H18, H21, H25, H31, H34, and H43) and were designated as distinct STs.

Accordingly, these results are consistent with the idea that H-types are more descriptive than O-serogroups and lead to the conclusion that the most proper phylogenetic background would be reflected by the O:H serotype. While this has been hypothesized for decades, the results presented here verify that the same O:H serotypes can also be found in distinct lineages of *E. coli*, even of strains of the same pathotypes, indicating a convergent evolution that results in the same serotypes. The same serotypes can be found in strains with highly diverse chromosomal composition and are likely to include different gene content and pathogenic potential. This is of major importance, as risk assessment of STEC strains based on the conventional O:H serotype alone is prone to misinterpretation and therefore should be improved by additional determination of the MLST sequence type.

4.4 Comparison of the performed population structure analysis with a phylogenetic analysis of ETEC strains

A recent publication on the phylogeny and evolution of enterotoxigenic *E. coli* (ETEC) by Mentzer *et al.* (2014) supports the relevance of the microevolutionary model developed in this thesis and the innovative nature of the applied, genome-based methods [105]. These authors examined a broad collection (n=362) of ETEC strains isolated worldwide between the years 1980-2011. The strains were selected on the basis of their colonization factors and toxin profiles, including strains lacking such factors. Hence, the ETEC collection in that study is at least partly comparable to our strain collection of STEC/ EHEC that were isolated worldwide, and which included aEPEC, strains which lack the Shiga-Toxin production. In the study of Mentzer *et al.* the ETEC collection was serotyped and whole genome sequenced for phylogenetic analyses. The authors performed SNP-analysis based on the previously determined MCG of the ETEC strains in the same manner we performed the SNP-analysis of the MCG of STEC/ EHEC and aEPEC strains. The authors further found that the SNP-based phylogenetic tree agreed with previous MLST studies of those ETEC strains, just as we observed in this thesis. In addition to SNP-analysis, they analyzed the colonization-factor profile and toxin-profile of the ETEC strains and compared the combined results as a specific virulence-profile with the SNP-based

phylogenetic tree. By this approach they were able to identify clusters of specific lineages in accordance with the SNP-based phylogenetic tree. Some lineages comprised ETEC strains with a mix of colonization factors and toxin genes, suggesting that in these lineages gene exchange might be common.

In the presented thesis, we analyzed the VAG-profile of STEC/ EHEC and aEPEC strains, which included both toxin genes and colonization-factors, and observed a comparable resulting population structure based on the VAG-profile as with the SNP-analysis. Furthermore, the strains of Cluster 3 showed quite diverse VAG-profiles compared to the other identified clusters where the VAG-profiles were more coherent. As Mentzer *et al.* concluded, and as we did show here, once the virulence factors were acquired, the clades (lineages or clusters) subsequently spread, implying a parallel evolution of the MCG and the VAGs. Furthermore, Mentzer *et al.* detected lineages that included only colonization-factor negative strains, but also colonization-factor negative strains that were spread across the phylogenetic tree. This finding is comparable to the aEPEC Cluster 2 in the SNP-based MSTree in this thesis and the aEPEC that were located among the strains of other clusters.

Overall, the data of Mentzer *et al.* clearly demonstrated that ETEC harbor identifiable lineages that are globally spread, but have been stable over substantial periods of time in endemic areas, which is contrary to the hypothesis that ETEC represent any *E. coli* lineage that could acquire, express and retain plasmid-harboring colonization factors and/or toxins. In the presented thesis, we found strong evidence that STEC/ EHEC strains of serogroup O26 and O111 developed from a common aEPEC ancestor and formed two distinct lineages that are specific in their MLST sequence type, SNP-profile, VAG-profile and integration of mobile genetic elements.

5 CONCLUSION

In this thesis a microevolutionary scenario of the most-important non-O157 STEC/ EHEC serotypes was developed using next-generation sequencing and innovative analyzing methods. Notably, we were able to identify the common ancestor of the closely related STEC/ EHEC lineages of O serogroups O26 and O111. This was especially achieved by the inclusion of aEPEC into the performed analyses steps. Additionally, the analyses underlined the close relation of the pathotypes STEC/ EHEC and aEPEC, as they share not only serotypes, but also MLST sequence types, SNPs of the MCG and VAG-profiles. Our model further is valid not only for human STEC/ EHEC strains, but also for bovine strains. All strains were analyzed with the same methods, revealing that the strains do not harbor any host-specific SNP- or VAG-profiles and therefore are *bona fide* zoonotic agents that constitute a risk of transmission between both hosts. With regard to the underlying evolutionary scenario this also means, that the development and emergence of certain STEC/ EHEC strains of the examined serotypes are not a result of adjustment of the MCG or acquisition of VAGs due to adaptation to stable niche over a long time period. The SNP-based MSTree and the MSTree built on basis of the absence and presences of VAGs interestingly resulted in comparable cluster formation which let us conclude that the evolution of MCG and acquisition of VAGs of STEC/ EHEC and aEPEC strains was a parallel development.

Applying an initial MLST analysis on a large collection of non-O157 STEC/ EHEC strains enabled us to recognize, first, that these strains have a heterogeneous relation and second, that strains of O-serogroup O26 and O111 clustered together in one distinct ST complex, STC29. Accordingly STC29 includes the most important STEC/ EHEC, but also aEPEC strains, and hence constituted a valuable capacity, which was suitable for further population analyses in order to develop an evolutionary model. Moreover, the MLST analysis already gave a solid representation of the population structure of the examined strains and, as MLST is a straight-forward method, allows the identification of emerging STEC/ EHEC strains in addition to *stx*-subtyping and serotyping in diagnostic approaches. Most importantly, the role of aEPEC in the emergence of O26 and O111 STEC/ EHEC, as elucidated in this thesis, should be considered for infection control measures to prevent possible lysogenic conversion with *stx*-converting bacteriophages as major vehicle driving the emergence of STEC/ EHEC lineages.

Overall this thesis uncovers crucial microevolutionary steps of the two most-important STEC/ EHEC lineages with direct Public Health consequences.

6 REFERENCES

1. Escherich, T., *The intestinal bacteria of the neonate and breast-fed infant. 1885.* Rev Infect Dis, 1989. 11(2): p. 352-6.
2. Ferens, W.A. and C.J. Hovde, *Escherichia coli O157:H7: animal reservoir and sources of human infection.* Foodborne Pathog Dis, 2011. 8(4): p. 465-87.
3. Hansen, D.L., et al., *Sources and Sinks of Escherichia coli in Benthic and Pelagic Fish.* Journal of Great Lakes Research, 2008. 34(2): p. 228-234.
4. Ishii, S., et al., *Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds.* Appl Environ Microbiol, 2006. 72(1): p. 612-21.
5. McLellan, S.L., *Genetic diversity of Escherichia coli isolated from urban rivers and beach water.* Appl Environ Microbiol, 2004. 70(8): p. 4658-65.
6. Holden, N.J., et al., *Prevalence and diversity of Escherichia coli isolated from a barley trial supplemented with bulky organic soil amendments: green compost and bovine slurry.* Lett Appl Microbiol, 2014. 58(3): p. 205-12.
7. Kamionka, M., *Engineering of therapeutic proteins production in Escherichia coli.* Curr Pharm Biotechnol, 2011. 12(2): p. 268-74.
8. Kaper, J.B., J.P. Nataro, and H.L. Mobley, *Pathogenic Escherichia coli.* Nat Rev Microbiol, 2004. 2(2): p. 123-40.
9. Dho-Moulin, M. and J.M. Fairbrother, *Avian pathogenic Escherichia coli (APEC).* Vet Res, 1999. 30(2-3): p. 299-316.
10. Johnson, J.R. and T.A. Russo, *Extraintestinal pathogenic Escherichia coli: "the other bad E coli".* J Lab Clin Med, 2002. 139(3): p. 155-62.
11. Nataro, J.P. and J.B. Kaper, *Diarrheagenic Escherichia coli.* Clin Microbiol Rev, 1998. 11(1): p. 142-201.
12. Ud-Din, A. and S. Wahid, *Relationship among Shigella spp. and enteroinvasive Escherichia coli (EIEC) and their differentiation.* Braz J Microbiol, 2014. 45(4): p. 1131-8.
13. Parsot, C., *Shigella spp. and enteroinvasive Escherichia coli pathogenicity factors.* FEMS Microbiol Lett, 2005. 252(1): p. 11-8.
14. Kotloff, K.L., et al., *Global burden of Shigella infections: implications for vaccine development and implementation of control strategies.* Bull World Health Organ, 1999. 77(8): p. 651-66.
15. Kaur, P., A. Chakraborti, and A. Asea, *Enteroaggregative Escherichia coli: An Emerging Enteric Food Borne Pathogen.* Interdiscip Perspect Infect Dis, 2010. 2010: p. 254159.
16. Morin, N., et al., *Characterization of the AggR regulon in enteroaggregative Escherichia coli.* Infect Immun, 2013. 81(1): p. 122-32.
17. Johnson, T.J. and L.K. Nolan, *Pathogenomics of the virulence plasmids of Escherichia coli.* Microbiol Mol Biol Rev, 2009. 73(4): p. 750-74.
18. Huang, D.B., et al., *A review of an emerging enteric pathogen: enteroaggregative Escherichia coli.* J Med Microbiol, 2006. 55(Pt 10): p. 1303-11.
19. Croxen, M.A. and B.B. Finlay, *Molecular mechanisms of Escherichia coli pathogenicity.* Nat Rev Microbiol, 2010. 8(1): p. 26-38.
20. Jafari, A., M.M. Aslani, and S. Bouzari, *Escherichia coli: a brief review of diarrheagenic pathotypes and their role in diarrheal diseases in Iran.* Iran J Microbiol, 2012. 4(3): p. 102-17.
21. Lanata, C.F., et al., *Global causes of diarrheal disease mortality in children <5 years of age: a systematic review.* PLoS One, 2013. 8(9): p. e72788.
22. Ochoa, T.J., et al., *New insights into the epidemiology of enteropathogenic Escherichia coli infection.* Trans R Soc Trop Med Hyg, 2008. 102(9): p. 852-6.
23. Karch, H., et al., *The enemy within us: lessons from the 2011 European Escherichia coli O104:H4 outbreak.* EMBO Mol Med, 2012. 4(9): p. 841-8.
24. Blattner, F.R., et al., *The complete genome sequence of Escherichia coli K-12.* Science, 1997. 277(5331): p. 1453-62.
25. Welch, R.A., et al., *Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli.* Proc Natl Acad Sci U S A, 2002. 99(26): p. 17020-4.
26. Lukjancenko, O., T.M. Wassenaar, and D.W. Ussery, *Comparison of 61 sequenced Escherichia coli genomes.* Microb Ecol, 2010. 60(4): p. 708-20.
27. Zagorec, M., M. Champomier-Vergès, and C. Cailliez-Grimal, *Identification Methods and DNA Fingerprinting: Whole Genome Sequencing*. Second edition. ed. Encyclopedia of food microbiology, ed. C.A. Batt. Vol. Volume 2. 2014, Amsterdam: Academic Press.
28. Mariani-Kurkdjian, P., et al., *Identification of a clone of Escherichia coli O103:H2 as a potential agent of hemolytic-uremic syndrome in France.* J Clin Microbiol, 1993. 31(2): p. 296-301.

29. Ahmed, N., et al., *Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention.* Nat Rev Microbiol, 2008. 6(5): p. 387-94.
30. Blakely, G.W., *Mechanisms of Horizontal Gene Transfer and DNA Recombination*, in *Molecular Medical Microbiology (Second Edition)*, J. Schwartzman, et al., Editors. 2015, Academic Press: Boston. p. 291-302.
31. Lorenz, M.G. and W. Wackernagel, *Bacterial gene transfer by natural genetic transformation in the environment.* Microbiol Rev, 1994. 58(3): p. 563-602.
32. Arutyunov, D. and L.S. Frost, *F conjugation: back to the beginning.* Plasmid, 2013. 70(1): p. 18-32.
33. Fineran, P.C., N.K. Petty, and G.P. Salmond, *Transduction: Host DNA Transfer by Bacteriophages* in *Encyclopedia of microbiology*, M. Schaechter, Editor 2009, Elsevier: Oxford. p. 666-679.
34. Brussow, H., C. Canchaya, and W.D. Hardt, *Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion.* Microbiol Mol Biol Rev, 2004. 68(3): p. 560-602, table of contents.
35. Rodriguez-Beltran, J., et al., *High Recombinant Frequency in Extraintestinal Pathogenic Escherichia coli Strains.* Mol Biol Evol, 2015. 32(7): p. 1708-16.
36. Groth, A.C. and M.P. Calos, *Phage integrases: biology and applications.* J Mol Biol, 2004. 335(3): p. 667-78.
37. Alberts, B., et al., *Site-Specific Recombination*. Molecular Biology of the Cell. 4th edition., ed. A. Uzman. Vol. 4. 2003, New York: Garland Science.
38. CDC, *National STEC Surveillance Overview*, in *Centers for Disease Control and Prevention*2012, US Department of Health and Human Services, CDC: Atlanta, Georgia.
39. RKI, *Infektionsepidemiologisches Jahrbuch meldepflichtiger Krankheiten für 2014*, 2015, Bundesministeriums für Gesundheit, Bundesinstitut Robert Koch-Institut (RKI): Berlin.
40. Besser, R.E., P.M. Griffin, and L. Slutsker, *Escherichia coli O157:H7 gastroenteritis and the hemolytic uremic syndrome: an emerging infectious disease.* Annu Rev Med, 1999. 50: p. 355-67.
41. Keusch, G.T. and M.B. Skirrow, *Enterobacteria, campylobacter, and miscellaneous food-poisoning bacteria*, in *Oxford Textbook of Medicine*, D.A. Warrell, T.M. Cox, and J.D. Firth, Editors. 2005, Oxford University Press: Oxford. p. 489-503.
42. Lowe, R.M., et al., *Escherichia coli O157:H7 strain origin, lineage, and Shiga toxin 2 expression affect colonization of cattle.* Appl Environ Microbiol, 2009. 75(15): p. 5074-81.
43. Wieler, L.H., et al., *Shiga toxin-producing Escherichia coli strains from bovines: association of adhesion with carriage of eae and other genes.* J Clin Microbiol, 1996. 34(12): p. 2980-4.
44. La Ragione, R.M., et al., *Escherichia coli O157:H7 colonization in small domestic ruminants.* FEMS Microbiol Rev, 2009. 33(2): p. 394-410.
45. Barth, S., et al., *Virulence and fitness gene patterns of Shiga toxin-encoding Escherichia coli isolated from pigs with edema disease or diarrhea in Germany.* Berl Munch Tierarztl Wochenschr, 2007. 120(7-8): p. 307-16.
46. Beutin, L., et al., *Prevalence and some properties of verotoxin (Shiga-like toxin)-producing Escherichia coli in seven different species of healthy domestic animals.* J Clin Microbiol, 1993. 31(9): p. 2483-8.
47. Heuvelink, A.E., et al., *Isolation and characterization of verocytotoxin-producing Escherichia coli O157 from slaughter pigs and poultry.* Int J Food Microbiol, 1999. 52(1-2): p. 67-75.
48. Rice, D.H., D.D. Hancock, and T.E. Besser, *Faecal culture of wild animals for Escherichia coli O157:H7.* Vet Rec, 2003. 152(3): p. 82-3.
49. Scaife, H.R., et al., *Wild rabbits (Oryctolagus cuniculus) as potential carriers of verocytotoxin-producing Escherichia coli.* Vet Rec, 2006. 159(6): p. 175-8.
50. Garcia, A., J.G. Fox, and T.E. Besser, *Zoonotic enterohemorrhagic Escherichia coli: A One Health perspective.* ILAR J, 2010. 51(3): p. 221-32.
51. Ogden, I.D., et al., *Long-term survival of Escherichia coli O157 on pasture following an outbreak associated with sheep at a scout camp.* Lett Appl Microbiol, 2002. 34(2): p. 100-4.
52. Chalmers, R.M., H. Aird, and F.J. Bolton, *Waterborne Escherichia coli O157.* Symp Ser Soc Appl Microbiol, 2000(29): p. 124S-132S.
53. Rangel, J.M., et al., *Epidemiology of Escherichia coli O157:H7 outbreaks, United States, 1982-2002.* Emerg Infect Dis, 2005. 11(4): p. 603-9.
54. Farrokh, C., et al., *Review of Shiga-toxin-producing Escherichia coli (STEC) and their significance in dairy production.* Int J Food Microbiol, 2013. 162(2): p. 190-212.
55. Paton, J.C. and A.W. Paton, *Pathogenesis and diagnosis of Shiga toxin-producing Escherichia coli infections.* Clin Microbiol Rev, 1998. 11(3): p. 450-79.

56. Gormley, F.J., et al., *A 17-year review of foodborne outbreaks: describing the continuing decline in England and Wales (1992-2008).* Epidemiol Infect, 2011. 139(5): p. 688-99.
57. Effler, E., et al., *Factors contributing to the emergence of Escherichia coli O157 in Africa.* Emerg Infect Dis, 2001. 7(5): p. 812-9.
58. Xiong, Y., et al., *A novel Escherichia coli O157:H7 clone causing a major hemolytic uremic syndrome outbreak in China.* PLoS One, 2012. 7(4): p. e36144.
59. Akashi, S., et al., *A severe outbreak of haemorrhagic colitis and haemolytic uraemic syndrome associated with Escherichia coli O157:H7 in Japan.* Eur J Pediatr, 1994. 153(9): p. 650-5.
60. Cole, D., et al., *Attributing sporadic and outbreak-associated infections to sources: blending epidemiological data.* Epidemiol Infect, 2014. 142(2): p. 295-302.
61. Karch, H., A. Mellmann, and M. Bielaszewska, *Epidemiology and pathogenesis of enterohaemorrhagic Escherichia coli.* Berl Munch Tierarztl Wochenschr, 2009. 122(11-12): p. 417-24.
62. Mellmann, A., et al., *Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic Escherichia coli.* Emerg Infect Dis, 2008. 14(8): p. 1287-90.
63. Gould, L.H., et al., *Increased recognition of non-O157 Shiga toxin-producing Escherichia coli infections in the United States during 2000-2010: epidemiologic features and comparison with E. coli O157 infections.* Foodborne Pathog Dis, 2013. 10(5): p. 453-60.
64. Lee, J.E., et al., *Phylogenetic analysis of Shiga toxin 1 and Shiga toxin 2 genes associated with disease outbreaks.* BMC Microbiol, 2007. 7: p. 109.
65. Scheutz, F., et al., *Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature.* J Clin Microbiol, 2012. 50(9): p. 2951-63.
66. Jores, J., L. Rumer, and L.H. Wieler, *Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic Escherichia coli.* Int J Med Microbiol, 2004. 294(2-3): p. 103-13.
67. Franzin, F.M. and M.P. Sircili, *Locus of enterocyte effacement: a pathogenicity island involved in the virulence of enteropathogenic and enterohemorragic Escherichia coli subjected to a complex network of gene regulation.* Biomed Res Int, 2015. 2015: p. 534738.
68. Garcia-Angulo, V.A., et al., *A distinct regulatory sequence is essential for the expression of a subset of nle genes in attaching and effacing Escherichia coli.* J Bacteriol, 2012. 194(20): p. 5589-603.
69. Menard, L.P. and J.D. Dubreuil, *Enteroaggregative Escherichia coli heat-stable enterotoxin 1 (EAST1): a new toxin with an old twist.* Crit Rev Microbiol, 2002. 28(1): p. 43-60.
70. McWilliams, B.D. and A.G. Torres, *Enterohemorrhagic Escherichia coli Adhesins.* Microbiol Spectr, 2014. 2(3).
71. Bakhshi, B., S. Fallahzad, and M.R. Pourshafie, *The occurrence of atypical enteropathogenic Escherichia coli strains among children with diarrhea in Iran.* J Infect Chemother, 2013. 19(4): p. 615-20.
72. Croxen, M.A., et al., *Recent advances in understanding enteric pathogenic Escherichia coli.* Clin Microbiol Rev, 2013. 26(4): p. 822-80.
73. Kozub-Witkowski, E., et al., *Serotypes and virutypes of enteropathogenic and enterohaemorrhagic Escherichia coli strains from stool samples of children with diarrhoea in Germany.* J Appl Microbiol, 2008. 104(2): p. 403-10.
74. Bugarel, M., et al., *Virulence gene profiling of enterohemorrhagic (EHEC) and enteropathogenic (EPEC) Escherichia coli strains: a basis for molecular risk assessment of typical and atypical EPEC strains.* BMC Microbiol, 2011. 11: p. 142.
75. Wieler, L.H., et al., *Longitudinal prevalence study of diarrheagenic Escherichia coli in dairy calves.* Berl Munch Tierarztl Wochenschr, 2007. 120(7-8): p. 296-306.
76. Trabulsi, L.R., R. Keller, and T.A. Tardelli Gomes, *Typical and atypical enteropathogenic Escherichia coli.* Emerg Infect Dis, 2002. 8(5): p. 508-13.
77. Bielaszewska, M., et al., *Shiga toxin-mediated hemolytic uremic syndrome: time to change the diagnostic paradigm?* PLoS One, 2007. 2(10): p. e1024.
78. Mellmann, A., et al., *Enterohemorrhagic Escherichia coli in human infection: in vivo evolution of a bacterial pathogen.* Clin Infect Dis, 2005. 41(6): p. 785-92.
79. Bielaszewska, M., et al., *Shiga toxin gene loss and transfer in vitro and in vivo during enterohemorrhagic Escherichia coli O26 infection in humans.* Appl Environ Microbiol, 2007. 73(10): p. 3144-50.
80. Sabat, A.J., et al., *Overview of molecular typing methods for outbreak detection and epidemiological surveillance.* Euro Surveill, 2013. 18(4): p. 20380.
81. Guentzel, M.N., *Escherichia, Klebsiella, Enterobacter, Serratia, Citrobacter, and Proteus*, in *Medical Microbiology*, S. Baron, Editor 1996: Galveston (TX).

82. DebRoy, C., et al., *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, 2016. 11(1): p. e0147434.
83. Prager, R., et al., *Subtyping of pathogenic Escherichia coli strains using flagellar (H)-antigens: serotyping versus fliC polymorphisms.* Int J Med Microbiol, 2003. 292(7-8): p. 477-86.
84. Coombes, B.K., M.W. Gilmour, and C.D. Goodman, *The evolution of virulence in non-o157 shiga toxin-producing Escherichia coli.* Front Microbiol, 2011. 2: p. 90.
85. Maiden, M.C., et al., *Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.* Proc Natl Acad Sci U S A, 1998. 95(6): p. 3140-5.
86. Wirth, T., et al., *Sex and virulence in Escherichia coli: an evolutionary perspective.* Mol Microbiol, 2006. 60(5): p. 1136-51.
87. Feil, E.J., et al., *eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data.* J Bacteriol, 2004. 186(5): p. 1518-30.
88. Cooper, J.E. and E.J. Feil, *Multilocus sequence typing--what is resolved?* Trends Microbiol, 2004. 12(8): p. 373-7.
89. Brown, T.A., *Chapter 5, Mapping Genomes.*, in *Genomes*, T.A. Brown, Editor 2002, Wiley-Liss: Oxford.
90. Vignal, A., et al., *A review on SNP and other types of molecular markers and their use in animal genetics.* Genet Sel Evol, 2002. 34(3): p. 275-305.
91. Sherry, S.T., et al., *dbSNP: the NCBI database of genetic variation.* Nucleic Acids Res, 2001. 29(1): p. 308-11.
92. Bono, J.L., et al., *Phylogeny of Shiga toxin-producing Escherichia coli O157 isolated from cattle and clinically ill humans.* Mol Biol Evol, 2012. 29(8): p. 2047-62.
93. Zhou, Z., et al., *Derivation of Escherichia coli O157:H7 from its O55:H7 precursor.* PLoS One, 2010. 5(1): p. e8700.
94. Peterson, J.W., *Chapter 7, Bacterial Pathogenesis.* 4 ed. Medical Microbiology, ed. S. Baron1996, Galveston, Texas: University of Texas Medical Branch, Department of Microbiology.
95. Feng, P., et al., *Genotypic and phenotypic changes in the emergence of Escherichia coli O157:H7.* J Infect Dis, 1998. 177(6): p. 1750-3.
96. Wu, X.Y., et al., *Comparative analysis of virulence genes, genetic diversity, and phylogeny of commensal and enterotoxigenic Escherichia coli isolates from weaned pigs.* Appl Environ Microbiol, 2007. 73(1): p. 83-91.
97. Lassalle, F., et al., *GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands.* PLoS Genet, 2015. 11(2): p. e1004941.
98. Rivas, L., et al., *Typing and Subtyping Methods for Pathogenic Escherichia coli*, in *Detection and Typing Strategies for Pathogenic Escherichia coli*2015, Springer New York: New York, NY. p. 67-99.
99. Steyert, S.R., et al., *Comparative genomics and stx phage characterization of LEE-negative Shiga toxin-producing Escherichia coli.* Front Cell Infect Microbiol, 2012. 2: p. 133.
100. Whittam, T.S. and R.A. Wilson, *Genetic relationships among pathogenic Escherichia coli of serogroup O157.* Infect Immun, 1988. 56(9): p. 2467-73.
101. Whittam, T.S., I.K. Wachsmuth, and R.A. Wilson, *Genetic evidence of clonal descent of Escherichia coli O157:H7 associated with hemorrhagic colitis and hemolytic uremic syndrome.* J Infect Dis, 1988. 157(6): p. 1124-33.
102. Whittam, T.S., et al., *Clonal relationships among Escherichia coli strains that cause hemorrhagic colitis and infantile diarrhea.* Infect Immun, 1993. 61(5): p. 1619-29.
103. Muller, D., et al., *Identification of unconventional intestinal pathogenic Escherichia coli isolates expressing intermediate virulence factor profiles by using a novel single-step multiplex PCR.* Appl Environ Microbiol, 2007. 73(10): p. 3380-90.
104. Friedrich, A.W., et al., *Shiga toxin 1c-producing Escherichia coli strains: phenotypic and genetic characterization and association with human disease.* J Clin Microbiol, 2003. 41(6): p. 2448-53.
105. von Mentzer, A., et al., *Identification of enterotoxigenic Escherichia coli (ETEC) clades with long-term global distribution.* Nat Genet, 2014. 46(12): p. 1321-6.
106. Edgar, R.C., *Search and clustering orders of magnitude faster than BLAST.* Bioinformatics, 2010. 26(19): p. 2460-1.
107. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. 42(Database issue): p. D206-14.

108. Joensen, K.G., et al., *Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli.* J Clin Microbiol, 2014. 52(5): p. 1501-10.
109. Hazen, T.H., et al., *Draft genome sequences of the diarrheagenic Escherichia coli collection.* J Bacteriol, 2012. 194(11): p. 3026-7.
110. Elliott, S.J., et al., *The locus of enterocyte effacement (LEE)-encoded regulator controls expression of both LEE- and non-LEE-encoded virulence factors in enteropathogenic and enterohemorrhagic Escherichia coli.* Infect Immun, 2000. 68(11): p. 6115-26.
111. Clarke, S.C., et al., *Virulence of enteropathogenic Escherichia coli, a global pathogen.* Clin Microbiol Rev, 2003. 16(3): p. 365-78.
112. Chen, J., S.M. Lee, and Y. Mao, *Protective effect of exopolysaccharide colanic acid of Escherichia coli O157:H7 to osmotic and oxidative stress.* Int J Food Microbiol, 2004. 93(3): p. 281-6.
113. Bielaszewska, M., et al., *Enterohaemorrhagic Escherichia coli O26:H11/H-: a human pathogen in emergence.* Berl Munch Tierarztl Wochenschr, 2007. 120(7-8): p. 279-87.
114. Brooks, J.T., et al., *Non-O157 Shiga toxin-producing Escherichia coli infections in the United States, 1983-2002.* J Infect Dis, 2005. 192(8): p. 1422-9.
115. Taylor, E.V., et al., *Multistate outbreak of Escherichia coli O145 infections associated with romaine lettuce consumption, 2010.* J Food Prot, 2013. 76(6): p. 939-44.
116. Bettelheim, K.A., *Non-O157 verotoxin-producing Escherichia coli: a problem, paradox, and paradigm.* Exp Biol Med (Maywood), 2003. 228(4): p. 333-44.
117. Bielaszewska, M., et al., *Enterohemorrhagic Escherichia coli O26:H11/H-: a new virulent clone emerges in Europe.* Clin Infect Dis, 2013. 56(10): p. 1373-81.
118. Bletz, S., et al., *Evolution of enterohemorrhagic escherichia coli O26 based on single-nucleotide polymorphisms.* Genome Biol Evol, 2013. 5(10): p. 1807-16.
119. Zhang, W., et al., *Structural and functional differences between disease-associated genes of enterohaemorrhagic Escherichia coli O111.* Int J Med Microbiol, 2007. 297(1): p. 17-26.
120. Guth, B.E., et al., *Re-emergence of O103 : H2 Shiga toxin-producing Escherichia coli infections in Sao Paulo, Brazil.* J Med Microbiol, 2005. 54(Pt 8): p. 805-6.
121. Cooper, K.K., et al., *Comparative genomics of enterohemorrhagic Escherichia coli O145:H28 demonstrates a common evolutionary lineage with Escherichia coli O157:H7.* BMC Genomics, 2014. 15: p. 17.
122. Mellmann, A., M. Bielaszewska, and H. Karch, *Intrahost genome alterations in enterohemorrhagic Escherichia coli.* Gastroenterology, 2009. 136(6): p. 1925-38.
123. Wells, J.G., et al., *Isolation of Escherichia coli serotype O157:H7 and other Shiga-like-toxin-producing E. coli from dairy cattle.* J Clin Microbiol, 1991. 29(5): p. 985-9.
124. Schmidt, H., et al., *Non-O157:H7 pathogenic Shiga toxin-producing Escherichia coli: phenotypic and genetic profiling of virulence traits and evidence for clonality.* J Infect Dis, 1999. 179(1): p. 115-23.
125. Ju, W., et al., *Phylogenetic analysis of non-O157 Shiga toxin-producing Escherichia coli strains by whole-genome sequencing.* J Clin Microbiol, 2012. 50(12): p. 4123-7.
126. Ju, W., et al., *Pathogenicity Islands in Shiga Toxin-Producing Escherichia coli O26, O103, and O111 Isolates from Humans and Animals.* Foodborne Pathog Dis, 2014. 11(5): p. 342-5.
127. Schmidt, H., M. Bielaszewska, and H. Karch, *Transduction of enteric Escherichia coli isolates with a derivative of Shiga toxin 2-encoding bacteriophage phi3538 isolated from Escherichia coli O157:H7.* Appl Environ Microbiol, 1999. 65(9): p. 3855-61.
128. Sekse, C., et al., *Is lack of susceptible recipients in the intestinal environment the limiting factor for transduction of Shiga toxin-encoding phages?* J Appl Microbiol, 2008. 105(4): p. 1114-20.
129. Cornick, N.A., et al., *In vivo transduction of an Stx-encoding phage in ruminants.* Appl Environ Microbiol, 2006. 72(7): p. 5086-8.
130. Gamage, S.D., et al., *Nonpathogenic Escherichia coli can contribute to the production of Shiga toxin.* Infect Immun, 2003. 71(6): p. 3107-15.
131. Zweifel, C., N. Cernela, and R. Stephan, *Detection of the emerging Shiga toxin-producing Escherichia coli O26:H11/H- sequence type 29 (ST29) clone in human patients and healthy cattle in Switzerland.* Appl Environ Microbiol, 2013. 79(17): p. 5411-3.
132. Reid, S.D., et al., *Parallel evolution of virulence in pathogenic Escherichia coli.* Nature, 2000. 406(6791): p. 64-7.
133. Bielaszewska, M., et al., *Shiga toxin-negative attaching and effacing Escherichia coli: distinct clinical associations with bacterial phylogeny and virulence traits and inferred in-host pathogen evolution.* Clin Infect Dis, 2008. 47(2): p. 208-17.

134. Marejkova, M., et al., *Enterohemorrhagic Escherichia coli as causes of hemolytic uremic syndrome in the Czech Republic.* PLoS One, 2013. 8(9): p. e73927.
135. Creuzburg, K., et al., *Genetic structure and chromosomal integration site of the cryptic prophage CP-1639 encoding Shiga toxin 1.* Microbiology, 2005. 151(Pt 3): p. 941-50.
136. Wong, C.S., et al., *The risk of the hemolytic-uremic syndrome after antibiotic treatment of Escherichia coli O157:H7 infections.* N Engl J Med, 2000. 342(26): p. 1930-6.
137. McGannon, C.M., C.A. Fuller, and A.A. Weiss, *Different classes of antibiotics differentially influence shiga toxin production.* Antimicrob Agents Chemother, 2010. 54(9): p. 3790-8.
138. Acheson, D.W., et al., *In vivo transduction with shiga toxin 1-encoding phage.* Infect Immun, 1998. 66(9): p. 4496-8.
139. Toth, I., et al., *Transduction of porcine enteropathogenic Escherichia coli with a derivative of a shiga toxin 2-encoding bacteriophage in a porcine ligated ileal loop system.* Appl Environ Microbiol, 2003. 69(12): p. 7242-7.
140. Allison, H.E., et al., *Immunity profiles of wild-type and recombinant shiga-like toxin-encoding bacteriophages and characterization of novel double lysogens.* Infect Immun, 2003. 71(6): p. 3409-18.
141. Ingle, D.J., 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* Microbial Genomics, 2016. Published Ahead of Print: 22 April, 2016.
142. Iguchi, A., S. Iyoda, and M. Ohnishi, *Molecular characterization reveals three distinct clonal groups among clinical shiga toxin-producing Escherichia coli strains of serogroup O103.* J Clin Microbiol, 2012. 50(9): p. 2894-900.

APPENDIX

Table S 1 List of 250 non-O157 strains used for MLST analysis. The characteristics serotype, pathotype, host species the strain originates from and its disease, year and location of isolation, and the laboratory that isolated the strain, are given for each strain, as well as the resulting MLST sequence type (ST) and ST complex (STC). The table continues until page 67.

Strain designation	other designation	ST Complex	ST	Serotype	Pathotype	Host species	Disease	Year	Country	Continent	Source Lab
566	IMT3873	STC29	16	O111:NM	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
1003	1003	STC29	574	O26	STEC	Human	unknown	unknown	Scotland	Europe	SERL, Edinburgh
3000	IMT18537	STC29	29	O26:H11	STEC	Food	none	2009	Germany	Europe	Herbert Schmidt, Institute of Food Science and Biotechnology, University Hohenheim
6549	IMT3540	STC29	21	O26:H11	STEC	Human	Bloody Diarrh	2000	Germany	Europe	IHIT, University Giessen
6555	6555	STC20	346	O103	STEC	Cattle	none	unknown	England	Europe	Department of Veterinary Science, University of Liverpool
152023	152023	STC29	384	O26:NM	STEC	Human	none	2000	Sweden	Europe	HPA, Laboratory of Enteric Pathogens, Colindale
157734	157734	STC20	387	O103	STEC	Cattle	none	unknown	Scotland	Europe	HPA, Laboratory of Enteric Pathogens, Colindale
166206	166206	STC20	386	O103	STEC	Human	none	unknown	Ireland	Europe	HPA, Laboratory of Enteric Pathogens, Colindale
0488/99	HUSEC021	STC32	32	O145:H28	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
05-07246	05-07246	none	343	O103:H25	STEC	Water	none	2005	Germany	Europe	RKI, Wernigerode
05-946	HUSEC001	STC10	43	O111:H10	EHEC	Human	HUS	2005	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
06-00369	IMT16157	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-00384	IMT16158	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-00669	IMT18770	STC32	32	O145:NM	EHEC	Human	unknown	2006	Germany	Europe	RKI, Wernigerode
06-03384	IMT16167	STC32	32	O145:NM	STEC	Human	unknown	2006	Germany	Europe	RKI, Wernigerode
06-03988	IMT16159	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-05011	IMT16168	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-05391	IMT16169	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-05425	06-05425	none	722	O145:H34	STEC	Human	unknown	2006	Germany	Europe	RKI, Wernigerode
06-05613	IMT16170	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-05654	IMT16160	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-05896	IMT16177	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-06306	IMT16171	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-06386	IMT16161	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-06674	IMT16172	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-06766	IMT16162	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-06925	IMT16173	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-07021	IMT16163	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-07698	IMT16174	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-08204	IMT16164	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-08439	IMT18769	STC32	32	O145:NM	EHEC	Human	unknown	2006	Germany	Europe	RKI, Wernigerode
06-08800	IMT16175	STC32	32	O145:NM	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
06-08868	IMT16176	STC32	1096	O145:NM	STEC	Human	unknown	2007	Germany	Europe	RKI, Wernigerode
06EGY4	06EGY4	none	1146	O103:H2	EHEC	Camel	Diarrhea	2006	Egypt	Africa	Iman Shabana, Department of Bacteriology, Mycology and Immunology, Suez Canal University, Ismailia, Egypt
07-01179	07-01179	none	657	O103:H18	EHEC	Human	none	2007	Germany	Europe	RKI, Wernigerode
07-01317	IMT16178	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-02163	IMT16197	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-02614	IMT16198	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-02636	07-02636	STC29	723	O103:H11	EHEC	Human	unknown	2007	Germany	Europe	RKI, Wernigerode
07-02933	IMT16199	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-03671	IMT18756	STC32	32	O145:NM	EHEC	Human	unknown	2007	Germany	Europe	RKI, Wernigerode
07-03830	IMT16179	STC29	29	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-03868	IMT16180	STC29	1097	O26:NM	STEC	Human	none	2007	Germany	Europe	RKI, Wernigerode
07-04094	IMT16165	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-04096	IMT16181	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-04309	IMT16182	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-04392	IMT18766	STC32	32	O145:NM	EHEC	Human	unknown	2007	Germany	Europe	RKI, Wernigerode
07-05275	07-05275	none	812	O103:H31	STEC	Food	unknown	2007	Germany	Europe	RKI, Wernigerode
07-05688	IMT16200	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-05922	IMT16201	STC20	1105	O103:H2	STEC	Human	none	2007	Germany	Europe	RKI, Wernigerode
07-06037-1	IMT16166	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-06408	IMT16202	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-06495	IMT16183	STC29	21	O26:H11	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-06609	IMT16203	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-06702	IMT16204	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-06723	IMT16205	STC20	17	O103:H2	STEC	Human	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-06882	IMT16184	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-06884	IMT16206	STC20	17	O103:H2	STEC	Cattle	unknown	unknown	Germany	Europe	RKI, Wernigerode
07-07334	IMT16185	STC29	21	O26:H11	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-07591	IMT16186	STC29	21	O26:H11	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
0739/03	IMT15915	STC29	29	O26:H11	STEC	Cattle	Diarrhea	2008	Germany	Europe	IHIT, University Giessen
08-00310	08-00310	none	829	O103:H21	STEC	Human	unknown	2008	Germany	Europe	RKI, Wernigerode
08-00549-1	IMT18763	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-01230	IMT18768	STC32	32	O145:NM	EHEC	Human	unknown	2008	Germany	Europe	RKI, Wernigerode
08-03954	IMT18771	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-04631	IMT18757	STC32	32	O145:NM	EHEC	Human	unknown	2008	Germany	Europe	RKI, Wernigerode
08-05123	IMT18751	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-05165	IMT18749	STC20	17	O103:H3	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-05416	IMT18761	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode

Strain designation	other designation	ST Complex	ST	Serotype	Pathotype	Host species	Disease	Year	Country	Continent	Source Lab
08-05616	IMT18759	STC29	1792	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-05621	IMT18760	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-05902	IMT18762	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-06238	IMT18750	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-06257	IMT18745	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-06869	IMT18764	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-07059	IMT18772	STC32	32	O145:NM	EHEC	Human	unknown	2008	Germany	Europe	RKI, Wernigerode
08-07239	IMT18752	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-07340	IMT18765	STC32	32	O145:NM	EHEC	Human	unknown	2008	Germany	Europe	RKI, Wernigerode
08-07469	IMT18758	STC29	1792	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-07655	IMT18755	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
08-07920	IMT18767	STC32	32	O145:NM	EHEC	Human	unknown	2008	Germany	Europe	RKI, Wernigerode
09-00075	IMT18753	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-00077	IMT18748	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-00110	IMT18744	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-01142	IMT18742	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-01527	IMT18747	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-01624	IMT20339	STC32	32	O145:NM	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-01625	IMT20340	STC32	32	O145:H28	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-01641	IMT18743	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-01685	IMT18746	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-02239	IMT18754	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-02629	IMT20341	STC20	386	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03071	IMT20342	STC20	386	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03073	IMT20343	STC20	386	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03259	IMT20344	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03267	IMT20345	STC20	386	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03425	IMT20346	STC32	32	O145:NM	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03673	IMT20347	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03930	IMT20348	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-03957	IMT20349	STC32	32	O145:H28	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-04372	IMT20350	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-04493	IMT20351	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-04614	IMT20352	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-04763	IMT20353	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-04775	IMT20354	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05206	IMT20355	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05247	IMT20356	STC32	32	O145:H28	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05248	IMT20357	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05480	IMT20358	STC20	1789	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05481	IMT20359	STC20	1789	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05573	IMT20360	none	1790	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05631	IMT20361	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05736	IMT20362	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05737	IMT20363	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-05755	IMT20364	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06038-1	IMT20365	STC20	386	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06044	IMT20366	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06045	IMT20367	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06278	IMT20368	STC32	32	O145:H28	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06367	IMT20369	STC32	32	O145:H28	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06633	IMT20370	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06768	IMT20371	STC32	32	O145:H28	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-06991	IMT20372	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
09-07205	IMT20373	STC20	17	O103:H2	EHEC	Human	unknown	2009	Germany	Europe	RKI, Wernigerode
103/00	IMT5031	STC20	17	O103:H2	STEC	Cattle	unknown	2000	Germany	Europe	BfR, Berlin
1530/99	HUSEC018	STC29	29	O26:H11	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
1639/77	IMT9689	STC29	97	O111:NM	STEC	Human	Diarrhea	1977	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
2245/98	HUSEC013	STC29	21	O26:H11	EHEC	Human	HUS	1998	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
248/89	IMT3793	STC29	16	O111:NM	STEC	Cattle	none	1989	unknown	unknown	IHIT, University Giessen
2516/00	HUSEC011	STC29	16	O111:H8	EHEC	Human	HUS	2000	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
2576/97/A	2576/97/A	STC20	17	O103:H2	EHEC	Human	HUS	1997	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
2791/97	HUSEC008	STC20	17	O103:NM	EHEC	Human	HUS	1997	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
2839/98	HUSEC036	none	659	O145:NM	EHEC	Human	HUS	1998	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
2969/99	2969/99	STC20	17	O103:H2	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
3271/00	HUSEC020	STC29	396	O26:H11	EHEC	Human	HUS	2000	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
331/331.1	IMT4406	STC29	113	O26:NM	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
3319/99	HUSEC017	STC29	21	O26:H11	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
3593/00	3593/00	STC29	21	O26:H11	EHEC	Human	Diarrhea	2000	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
413/89-1	IMT477	STC29	113	O26:NM	EHEC	Cattle	Diarrhea	1989	Germany	Europe	IHIT, University Giessen
4392/97	HUSEC033	none	342	O145:H25	EHEC	Human	HUS	1997	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
4417/96	IMT24554	STC29	16	O111:NM	EHEC	Human	HUS	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
4556/99	IMT24564	STC29	16	O111:NM	EHEC	Human	Diarrhea	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
4557/99	HUSEC022	STC32	137	O145:NM	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
4672/99	4672/99	STC32	32	O145:NM	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
4763/23-49	IMT2994	STC20	17	O103:H2	STEC	Food	unknown	unknown	unknown	unknown	IHIT, University Giessen
4797/97	IMT9686	none	33	O103:NM	EHEC	Human	unknown	1997	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
5080/97	HUSEC014	STC29	21	O26:NM	EHEC	Human	HUS	1997	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
5122/99	5122/99	STC32	32	O145:NM	EHEC	Human	HUS	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster

Strain designation	other designation	ST Complex	ST	Serotype	Pathotype	Host species	Disease	Year	Country	Continent	Source Lab
530/98	IMT5029	STC20	17	O103	STEC	Food	none	1998	Germany	Europe	BfR, Berlin
5577/96	5577/96	STC20	17	O103:H2	EHEC	Human	Diarrhea	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
561/379S	IMT3846	STC29	21	O26	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
5714/96	5714/96	STC20	17	O103:H2	EHEC	Human	Diarrhea	1996	Germany	Europe	Jörg Hacker, Research Center for Infectious Diseases, University of Würzburg
5720/96	5720/96	STC29	29	O26:H11	EHEC	Human	HUS	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
6037/96	HUSEC012	STC29	16	O111:NM	EHEC	Human	HUS	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
607/99	IMT5030	STC20	17	O103:H2	STEC	Cattle	unknown	1999	Germany	Europe	BfR, Berlin
718/05	IMT15935	STC20	1099	O103:H2	STEC	Cattle	Diarrhea	2005	Germany	Europe	IHIT, University Giessen
726/99	IMT5032	STC20	17	O103:H2	STEC	Human	unknown	1999	Germany	Europe	BfR, Berlin
7382/96	HUSEC007	STC20	17	O103:H2	EHEC	Human	HUS	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
770/99	IMT5033	STC20	17	O103:H2	STEC	Human	unknown	1999	Germany	Europe	BfR, Berlin
778-5r	IMT19949	STC20	17	O103:H2	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
7795/96	7795/96	STC32	32	O145:NM	EHEC	Human	none	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
804/99	IMT5034	STC20	17	O103:H2	STEC	Human	unknown	1999	Germany	Europe	BfR, Berlin
807/99	IMT5035	STC20	17	O103:H2	STEC	Human	unknown	1999	Germany	Europe	BfR, Berlin
830-1c	IMT20169	STC20	17	O103:H2	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
830-1r	IMT20181	STC20	17	O103:H2	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
830-5c	IMT20170	STC20	17	O103:H2	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
83-574	IMT3741	STC29	21	O26:H11	STEC	Human	HUS	unknown	unknown	unknown	IHIT, University Giessen
AS3837	IMT5066	STC20	17	O103	STEC	Cat	unknown	1997	Germany	Europe	IHIT, University Giessen
C1785	IMT3824	STC29	21	O26:H11	STEC	Cattle	unknown	unknown	unknown	unknown	IHIT, University Giessen
C2023-99A	C2023-99A	STC20	17	O103:H2	STEC	Human	Bloody Diarrh	1999	Denmark	Europe	Flemming Scheutz, Statens Serum Institut, Copenhagen, Denmark
C3426	IMT3813	STC29	21	O26:H11	STEC	Cattle	none	unknown	unknown	unknown	IHIT, University Giessen
CL-37	IMT6616	STC29	16	O111:H8	EHEC	Human	HUS	1982	Canada	North Americ	Mohamed A. Karmali, Health Canada, Guelph, Ontario
DEC10E	IMT3753	STC29	21	O26:H11	EHEC	Cattle	none	1989	USA	North Americ	Thomas S. Whittam, University of Michigan
DEC8A	IMT4320	STC29	16	O111a:NM	EHEC	Human	Diarrhea	1977	USA	North Americ	Thomas S. Whittam, University of Michigan
DEC8B	IMT3800	STC29	294	O111:H8	EHEC	Human	Bloody Diarrh	1986	USA	North Americ	Thomas S. Whittam, University of Michigan
EC1537/00	IMT10304	STC29	21	O26:H11	STEC	Cattle	unknown	1999	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC157/00	IMT10276	STC29	21	O26:NM	STEC	Cattle	unknown	2000	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC1702/00	IMT10305	STC29	21	O26	STEC	Cattle	unknown	2000	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC172/00	IMT10277	STC29	21	O26:NM	EHEC	Cattle	none	2000	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC1721/00	IMT10306	STC29	21	O26:NM	STEC	Cattle	unknown	2000	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC1722/00	IMT10307	STC29	21	O26:NM	EHEC	Cattle	Diarrhea	2000	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC183/99	IMT10279	STC29	21	O26:NM	STEC	Cattle	Diarrhea	1999	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC1840/00	IMT10309	STC29	21	O26:NM	EHEC	Cattle	Diarrhea	2000	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC537/01	IMT10289	STC29	21	O26	EHEC	Cattle	none	2001	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC681/99	IMT10291	STC29	21	O26	STEC	Sheep	none	1999	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC690/02	IMT10292	STC29	21	O26	STEC	Cattle	unknown	2005	United Kingdo	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
GS0354-1	IMT3853	STC29	21	O26:H11	STEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
GS0740-1	IMT5050	STC20	17	O103:H2	STEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
GS0845-1	IMT3870	STC29	21	O26:NM	STEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
IHIT0304	IMT480	STC32	137	O145:H28	STEC	Cattle	Diarrhea	1992	Germany	Europe	IHIT, University Giessen
IHIT1703	IMT485	STC29	16	O111:H2	EHEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
IHIT1706	IMT3749	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
IHIT2087	IMT487	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
IMT15313	IMT15313	STC29	1093	O26:NM	STEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT15990	IMT15990	STC32	32	O145:NM	STEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT15991	IMT15991	STC32	32	O145:NM	STEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT15994	IMT15994	STC32	32	O145:NM	STEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT16014	IMT16014	STC10	48	O103:H43	STEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19609	IMT19609	STC32	32	O145:NM	STEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19636	IMT19636	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19648	IMT19648	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19883	IMT19883	STC155	58	O145:H18	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20180	IMT20180	none	129	O26	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20217	IMT20217	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20231	IMT20231	none	1125	O26:H19	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin

Strain designation	other designation	ST Complex	ST	Serotype	Pathotype	Host species	Disease	Year	Country	Continent	Source Lab
IMT20239	IMT20239	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20259	IMT20259	STC20	376	O103:H2	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20314	IMT20314	STC29	1573	O26:H11	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20337	IMT20337	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT3901	IMT3901	STC20	9	O111:H2	STEC	Cattle	unknown	1985	unknown	unknown	IHIT, University Giessen
IMT4368	IMT4368	STC20	17	O103:H21	STEC	Cattle	unknown	unknown	unknown	unknown	IMT, Veterinary Dep.,Freie Universität Berlin
IMT7831	5822	STC165	165	O111	STEC	Wild rabbit	Enteritis	1998	Gauteng	Africa	Bacteriol. Onderst. Vet. Inst. South Africa
P1933/05-4	IMT15936	STC29	1100	O26:H11	STEC	Cattle	Diarrhea	2005	Germany	Europe	IHIT, University Giessen
P2332/07-1	IMT15977	STC29	29	O26:H11	STEC	Cattle	none	2007	Germany	Europe	IHIT, University Giessen
P3146/08-1	IMT15987	STC29	21	O26:H11	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen
P3146/08-2	IMT15988	STC29	21	O26:H11	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen
P3146/08-3	IMT15989	STC29	21	O26:H11	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen
P355/07	IMT15973	STC29	21	O26:NM	STEC	Cattle	Diarrhea	2007	Germany	Europe	IHIT, University Giessen
P4352/06-1	IMT15967	STC29	21	O26:H11	STEC	Cattle	none	2006	Germany	Europe	IHIT, University Giessen
P4501/05-1	IMT15938	STC29	16	O111:NM	STEC	Cattle	none	2005	Germany	Europe	IHIT, University Giessen
P7278/07	IMT15981	STC29	21	O26:H11	STEC	Cattle	none	2007	Germany	Europe	IHIT, University Giessen
P7404/03-1	IMT15922	STC29	16	O111:NM	STEC	Cattle	Diarrhea	2003	Germany	Europe	IHIT, University Giessen
P8054/02	IMT15911	STC29	21	O26:H11	STEC	Cattle	Diarrhea	2002	Germany	Europe	IHIT, University Giessen
RW0145	IMT3837	STC29	21	O26:H11	STEC	Cattle	Diarrhea	unknown	Switzerland	Europe	IHIT, University Giessen
RW0303	IMT3822	none	1	O103	STEC	unknown	unknown	unknown	Germany	Europe	IHIT, University Giessen
RW0306	IMT3796	STC32	137	O145:H28	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1300	IMT3855	STC29	16	O111:NM	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1372	IMT494	STC20	17	O103:H2	STEC	Camel	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1374	IMT3529	STC20	17	O103:H2	EHEC	Cattle	Bloody Diarrh	1993	Germany	Europe	IHIT, University Giessen
RW1703	IMT3735	STC29	16	O111:H2	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1706	IMT3173	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1993	IMT3909	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
RW2144	IMT3750	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
RW2174	IMT3174	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
RW2198	IMT3746	STC20	17	O103:H2	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
WH 01/02/00	WH 01/02/00	STC29	21	O26:H11	STEC	Cattle	none	1997	Germany	Europe	FLI, Wusterhausen
Wh03/21/011	IMT8227	STC20	17	O103:H2	STEC	Human	unknown	2003	Germany	Europe	FLI, Wusterhausen
WX007869S0	WX007869S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008325S0	WX008325S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008326S0	WX008326S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008349S0	WX008349S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008650S0	WX008650S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008796S0	WX008796S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008809S0	WX008809S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008821S0	WX008821S0	STC29	21	O26:H11	STEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008859S0	WX008859S0	STC10	986	O26:H11	STEC	Cattle	none	unknown	Scotland	Europe	SERL, Edinburgh
x-0712/08	IMT15941	STC20	17	O103:H2	STEC	Cattle	Diarrhea	2008	Germany	Europe	IHIT, University Giessen
x-0713/08	IMT15943	STC20	17	O103:H2	STEC	Cattle	Diarrhea	2008	Germany	Europe	IHIT, University Giessen
x-0719/08	IMT15940	STC20	17	O103:H2	STEC	Cattle	Diarrhea	2008	Germany	Europe	IHIT, University Giessen
x-0770/08	IMT15972	STC20	17	O103:H2	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen

Table S 2 Characteristics (serotype, pathotype, host species the strain originates from and its disease, year and location of isolation, and the laboratory that isolated the strain, ST) of the 41 aEPEC strains of STC29 added for MLST analysis.

Strain designation	other designation	ST Complex	ST	Serotype	Pathotype	Host species	Disease	Year	Country	Continent	Source Lab
3801	IMT3538	STC29	1107	O26	aEPEC	unknown	Diarrhea	1996	unknown	unknown	IHIT, University Giessen
4396	4396	STC29	575	O26	aEPEC	Human	Diarrhea	unknown	Scotland	Europe	SERL, Edinburgh
3866/00	3866/00	STC29	29	O26:H11	aEPEC	Human	Diarrhea	2000	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
549/W146/332S	IMT3857	STC29	21	O26:H11	aEPEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
551/344S	IMT3874	STC29	21	O26:NM	aEPEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
591/W147/338S	IMT3888	STC29	1108	O26:H11	aEPEC	Cattle	Diarrhea	1998	Belgium	Europe	IHIT, University Giessen
6416/87	6416/87	STC29	29	O26:NM	aEPEC	Human	Diarrhea	1987	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
DEC9A	IMT4334	STC29	29	O26:H11	aEPEC	Human	unknown	1961	USA	North America	Thomas S. Whittam, University of Michigan
DEC9B	IMT4333	STC29	29	O26:NM	aEPEC	Human	Diarrhea	1979	USA	North America	Thomas S. Whittam, University of Michigan
DEC9C	IMT4337	STC29	29	O26:NM	aEPEC	Human	unknown	1952	Switzerland	Europe	Thomas S. Whittam, University of Michigan
DEC9E	IMT3766	STC29	21	O26:H11	aEPEC	Human	none	unknown	USA	North America	Thomas S. Whittam, University of Michigan
EC1388/00	IMT10302	STC29	29	O26	aEPEC	Sheep	Nephrosis	2000	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC1725/00	IMT10308	STC29	29	O26:NM	aEPEC	Cattle	Diarrhea	2000	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC196/02	IMT10280	STC29	29	O26:NM	aEPEC	Cattle	Diarrhea	2002	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC225/00	IMT10281	STC29	29	O26:NM	aEPEC	Cattle	Bloody Diarrhea	2000	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC33/67	IMT10273	STC29	21	O26:H11	aEPEC	Cattle	Diarrhea	1967	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC335/98	IMT10283	STC29	29	O26:NM	aEPEC	Sheep	Diarrhea	1998	Sweden	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC38/99	IMT10274	STC29	29	O26:NM	aEPEC	Cattle	Diarrhea	1999	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
EC459/01	IMT10286	STC29	389	O26:NM	aEPEC	Cattle	Mastitis	2001	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
GS0481-1	IMT3861	STC29	21	O26:H11	aEPEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
IMT15301	IMT15301	STC29	29	O26:H11	aEPEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT15302	IMT15302	STC29	29	O26:H11	aEPEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19623	IMT19623	STC29	29	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19947	IMT19947	STC29	21	O26:H11	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19981	IMT19981	STC29	21	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19996	IMT19996	STC29	21	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20165	IMT20165	STC29	21	O26:H11	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20245	IMT20245	STC29	21	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
LH-1	IMT21674	STC29	29	O26:NM	aEPEC	Human	unknown	unknown	Australia	Australia	Gad Frankel, Fac. of Natural Sciences, Imperial College, London
LH-7	IMT21680	STC29	29	O26:NM	aEPEC	Human	unknown	unknown	Australia	Australia	Gad Frankel, Fac. of Natural Sciences, Imperial College, London
LH-8	IMT21681	STC29	29	O26:NM	aEPEC	Human	unknown	unknown	Australia	Australia	Gad Frankel, Fac. of Natural Sciences, Imperial College, London
RW2070	IMT3734	STC29	16	O111:NM	aEPEC	Cattle	Diarrhea	unknown	Germany	Europe	IHIT, University Giessen
WX007861S01K	WX007861S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX008576S01K	WX008576S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX009931S01K	WX009931S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX010765S01K	WX010765S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX011379S01K	WX011379S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2002	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX016290S01K	WX016290S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2003	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX016732S01K	WX016732S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2003	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX017042S01K	WX017042S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2003	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh
WX018178S01K	WX018178S01K	STC29	29	O26:H11	aEPEC	Cattle	none	2003	Scotland	Europe	Michael C. Pearce, Centre for Tropical Veterinary Medicine, University of Edinburgh

Table S 3 Characteristics (ST,serotype, pathotype, host species the strain originates from and its disease, year and location of isolation, and the laboratory that isolated the strain) of the 99 strains of STC29 selected for whole genome sequencing. The table continues until page 71.

Strain designation	other designation	ST Complex	ST	Serotype	Patho-type	Host species	Disease	Year	Country	Continent	Source laboratory
566	IMT3873	STC29	16	O111:NM	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
3000	IMT18537	STC29	29	O26:H11	STEC	Food	none	2009	Germany	Europe	Herbert Schmidt, Institute of Food Science and Biotechnology, University Hohenheim
6549	IMT3540	STC29	21	O26:H11	STEC	Human	Bloody Diarrhea	2000	Germany	Europe	IHIT, University Giessen
03_06688_2		STC29	29	O26:H11	EHEC	Human	unknown	2003	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
06_05398		STC29	591	O26:H11	EHEC	Human	unknown	2006	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
06-00369	IMT16157	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-03988	IMT16159	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-05896	IMT16177	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-06386	IMT16161	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
06-07021	IMT16163	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-04094	IMT16165	STC29	16	O111:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-04096	IMT16181	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-06495	IMT16183	STC29	21	O26:H11	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-06882	IMT16184	STC29	21	O26:NM	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-07334	IMT16185	STC29	21	O26:H11	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
07-07591	IMT16186	STC29	21	O26:H11	STEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
0739/03	IMT15915	STC29	29	O26:H11	STEC	Cattle	Diarrhea	2008	Germany	Europe	IHIT, University Giessen
08-05616	IMT18759	STC29	1792	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-05621	IMT18760	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-05902	IMT18762	STC29	16	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
08-07469	IMT18758	STC29	1792	O111:NM	EHEC	Human	none	2008	Germany	Europe	RKI, Wernigerode
1226_65		STC29	21	O26:H11	EHEC	Human	unknown	1965	Czech Republic	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
12801#76		STC29	21	O26:H11	EHEC	Human	unknown	1976	Czech Republic	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
1639/77	IMT9689	STC29	97	O26:H11	aEPEC	Human	Diarrhea	1977	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
3184_00		STC29	1565	O26:H11	EHEC	Human	unknown	2000	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
331/331.1	IMT4406	STC29	113	O26:NM	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
347#73		STC29	29	O26:H11	aEPEC	Human	unknown	1973	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
413/89-1	IMT4040	STC29	113	O26:NM	EHEC	Cattle	Diarrhea	1989	Germany	Europe	IHIT, University Giessen
4417/96	IMT24554	STC29	16	O111:NM	EHEC	Human	HUS	1996	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
4556/99	IMT24564	STC29	16	O111:NM	EHEC	Human	Diarrhea	1999	Germany	Europe	Helge Karch, Institute for Hygiene, University of Muenster, Münster
549/W146/332S	IMT3857	STC29	21	O26:H11	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
551/344S	IMT3874	STC29	21	O26:NM	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
561/379S	IMT3846	STC29	21	O26	STEC	Cattle	Diarrhea	unknown	Belgium	Europe	IHIT, University Giessen
83-574	IMT3741	STC29	21	O26:H11	EHEC	Human	HUS	unknown	unknown	unknown	IHIT, University Giessen
8574_96		STC29	1705	O26:H11	EHEC	Human	unknown	1996	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
C3426	IMT3813	STC29	21	O26:H11	STEC	Cattle	none	unknown	unknown	unknown	IHIT, University Giessen
CL-37	IMT6616	STC29	16	O111:H8	EHEC	Human	HUS	1982	Canada	North America	Mohamed A. Karmali, Health Canada, Guelph, Ontario
DEC10A	H30	STC29	21	O26:H11	EHEC	Human	Diarrhea	unknown	United Kingdom	Europe	DEC Collection
DEC10B	3047-86	STC29	21	O26:H11	EHEC	Human	Bloody Diarrhea	1986	Australia	Australia	DEC Collection

Strain designation	other designation	ST Complex	ST	Serotype	Patho-type	Host species	Disease	Year	Country	Continent	Source laboratory
DEC10C	1557_77	STC29	21	O26:H11	EHEC	Human	unknown	1977	USA	North America	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
DEC10D		STC29	21	O26:H11	aEPEC	Human	unknown	1952	France	Europe	DEC Collection
DEC8A	IMT4320	STC29	16	O111:NM	EHEC	Human	Diarrhea	1977	USA	North America	DEC Collection
DEC8B	IMT3800	STC29	294	O111:H8	EHEC	Human	Bloody Diarrhea	1986	USA	North America	DEC Collection
DEC8C	86-10049	STC29	21	O111:NM	EHEC	Cattle	scours	1986	USA	North America	DEC Collection
DEC8D	C130-53	STC29	21	O111:H11	aEPEC	Human	Diarrhea	1953	Cuba	Central America	DEC Collection
DEC9A	IMT4334	STC29	29	O26:H11	aEPEC	Human	unknown	1961	USA	North America	DEC Collection
DEC9B	IMT4333	STC29	29	O26:NM	aEPEC	Human	Diarrhea	1979	USA	North America	DEC Collection
DEC9C	IMT4337	STC29	29	O26:NM	aEPEC	Human	unknown	1952	Switzerland	Europe	DEC Collection
DEC9D	C814-67	STC29	29	O26:H11	aEPEC	Human	Diarrhea	1967	Denmark	Europe	DEC Collection
DEC9E	IMT3766	STC29	29	O26:H11	aEPEC	Human	none	unknown	USA	North America	DEC Collection
E10_149		STC29	21	O26:H11	EHEC	Human	unknown	2010	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
EC1840/00	IMT10309	STC29	21	O26:NM	EHEC	Cattle	Diarrhea	2000	United Kingdom	Europe	Muna Anjum, Dep. of Bacteriology, AHVLA, Addlestone
GS0354-1	IMT3853	STC29	21	O26:H11	STEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
GS0481-1	IMT3861	STC29	21	O26:H11	STEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
GS0845-1	IMT3870	STC29	21	O26:NM	STEC	Cattle	none	1995	Germany	Europe	IHIT, University Giessen
IHIT1706	IMT3749	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
IHIT2087	IMT487	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
IMT15301	IMT15301	STC29	29	O26:H11	aEPEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT15302	IMT15302	STC29	29	O26:H11	aEPEC	Cattle	none	2008	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19623	IMT19623	STC29	29	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19636	IMT19636	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19648	IMT19648	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19947	IMT19947	STC29	21	O26:H11	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19981	IMT19981	STC29	21	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT19996	IMT19996	STC29	21	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20165	IMT20165	STC29	21	O26:H11	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20217	IMT20217	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20239	IMT20239	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20245	IMT20245	STC29	21	O26:NM	aEPEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
IMT20337	IMT20337	STC29	21	O26:NM	STEC	Cattle	none	2009	Germany	Europe	IMT, Veterinary Dep.,Freie Universität Berlin
LB228716		STC29	21	O26:H11	EHEC	Human	unknown	2011	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB234771		STC29	21	O26:H11	EHEC	Human	unknown	2011	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB238877		STC29	21	O26:H11	EHEC	Human	unknown	2011	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB267954	FBI-07695	STC29	21	O26:H11	EHEC	Human	unknown	2012	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB275203_1	FBI-07732	STC29	21	O26:H11	EHEC	Human	unknown	2012	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB279967	FBI-07731	STC29	29	O26:H11	EHEC	Human	unknown	2011	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB363084	1740	STC29	21	O26:H11	EHEC	Human	unknown	2014	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB363104	1743	STC29	29	O26:H11	EHEC	Human	unknown	2014	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster

Strain designation	other designation	ST Complex	ST	Serotype	Patho-type	Host species	Disease	Year	Country	Continent	Source laboratory
LB366332i2	1928	STC29	16	O111:H8	EHEC	Human	unknown	2014	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LB369997	2080	STC29	29	O26:H11	EHEC	Human	unknown	2014	Germany	Europe	Alexander Mellmann, Institute for Hygiene, University of Muenster, Münster
LH-1	IMT21674	STC29	29	O26:NM	aEPEC	Human	unknown	unknown	Australia	Australia	Gad Frankel, Fac. of Natural Sciences, Imperial College, London
LH-8	IMT21681	STC29	29	O26:NM	aEPEC	Human	unknown	unknown	Australia	Australia	Gad Frankel, Fac. of Natural Sciences, Imperial College, London
P2332/07-1	IMT15977	STC29	29	O26:H11	STEC	Cattle	none	2007	Germany	Europe	IHIT, University Giessen
P3146/08-1	IMT15987	STC29	21	O26:H11	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen
P3146/08-2	IMT15988	STC29	21	O26:H11	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen
P3146/08-3	IMT15989	STC29	21	O26:H11	STEC	Cattle	none	2008	Germany	Europe	IHIT, University Giessen
P355/07	IMT15973	STC29	21	O26:NM	STEC	Cattle	Diarrhea	2007	Germany	Europe	IHIT, University Giessen
P4352/06-1	IMT15967	STC29	21	O26:H11	STEC	Cattle	none	2006	Germany	Europe	IHIT, University Giessen
P4501/05-1	IMT15938	STC29	16	O111:NM	STEC	Cattle	none	2005	Germany	Europe	IHIT, University Giessen
P7278/07	IMT15981	STC29	21	O26:H11	STEC	Cattle	none	2007	Germany	Europe	IHIT, University Giessen
P7404/03-1	IMT15922	STC29	16	O111:NM	STEC	Cattle	Diarrhea	2003	Germany	Europe	IHIT, University Giessen
P8054/02	IMT15911	STC29	21	O26:H11	STEC	Cattle	Diarrhea	2002	Germany	Europe	IHIT, University Giessen
RW0145	IMT3837	STC29	21	O26:H11	STEC	Cattle	Diarrhea	unknown	Switzerland	Europe	IHIT, University Giessen
RW1300	IMT3855	STC29	16	O111:NM	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1703	IMT3735	STC29	16	O111:H2	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1706	IMT3173	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1993	Germany	Europe	IHIT, University Giessen
RW1993	IMT3909	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
RW2070	IMT3734	STC29	16	O111:NM	aEPEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen
RW2174	IMT3174	STC29	21	O26:H11	STEC	Cattle	Diarrhea	1994	Germany	Europe	IHIT, University Giessen

Table S 4 Presence (indicated with 1) and absence (indicated with 0) of 34 virulence associated genes identified in the genome sequences of 99 strains of STC29 using the 'VirulenceFinder' tool (http://www.genomicepidemiology.org/) [108]. The *stx* genes are given with their subtypes when identified in the respective WGS. The table continues until page 74.

Strain designation	*stx1*	*stx2*	*eae*	*tir*	*nleA*	*nleB*	*nleC*	*iha*	*espA*	*espB*	*espF*	*espI*	*espJ*	*espP*	*efa1*	*ehxA*	*astA*	*cif*	*lpfA*	*epeA*	*toxB*	*ironN*	*etpD*	*iss*	*cba*	*cma*	*celb*	*tccp*	*mchF*	*mchC*	*mchB*	*prfB*	*gad*	*katP*
566	stx1a	0	1	1	1	1	0	1	1	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
3000	0	stx2a	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0	1	1	0
6549	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
03_06688_2	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0	1	1	0
06_05398	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1
06-00369	stx1a	0	1	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
06-03988	stx1a	0	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	1	0	0	0	0	1	1	0	0	0	0	0	1	1	1
06-05896	0	stx2a	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
06-06386	stx1a	0	1	1	1	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	0	0	0	0	1	1	0	0	0	0	1	1	0
06-07021	stx1a	0	1	1	1	0	1	1	1	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
07-04094	stx1a	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
07-04096	0	stx2a	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
07-06495	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
07-06882	stx1a	stx2a	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
07-07334	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
07-07591	stx1a	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
0739/03	0	stx2a	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0
08-05616	stx1a	stx2a	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
08-05621	0	stx2a	1	1	1	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
08-05902	stx1a	stx2a	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
08-07469	stx1a	stx2a	1	1	1	0	1	1	1	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0
1226_65	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1
12801#76	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
1639/77	0	0	1	1	1	1	1	0	1	1	1	0	1	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
3184_00	stx1a	0	1	1	1	1	0	1	1	1	0	1	1	0	1	0	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1	
331/331.1	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	0
347#73	0	0	1	1	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
413/89-1	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	0
4417/96	stx1a	0	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
4556/99	stx1a	0	1	1	1	1	1	1	1	0	1	0	1	0	1	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
549/W146/332S	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
551/344S	stx1a	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1
561/379S	stx1a	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	1
83-574	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0
8574_96	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
C3426	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
CL-37	stx1a	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0
DEC10A	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	0	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
DEC10B	stx1a	0	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1

Strain designation	*stx1*	*stx2*	*eae*	*tir*	*nleA*	*nleB*	*nleC*	*iha*	*espA*	*espB*	*espF*	*espI*	*espJ*	*espP*	*efa1*	*ehxA*	*astA*	*cif*	*lpfA*	*epeA*	*toxB*	*iroN*	*etpD*	*iss*	*cba*	*cma*	*celb*	*tccp*	*mchF*	*mchC*	*mchB*	*prfB*	*gad*	*katP*
DEC10C	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
DEC10D	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
DEC8A	stx1a	0	1	1	1	1	1	1	1	0	0	0	1	0	0	1	1	1	1	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0
DEC8B	stx1a	stx2a	1	1	1	1	1	1	1	0	0	0	1	0	0	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	1	0
DEC8C	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1
DEC8D	0	0	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
DEC9A	0	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
DEC9B	0	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
DEC9C	0	0	1	1	1	1	0	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
DEC9D	0	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
DEC9E	0	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
E10_149	stx1a	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
EC1840/00	stx1a	0	1	1	1	1	0	1	1	1	1	0	1	0	0	0	1	1	1	0	0	0	0	1	1	1	1	1	0	0	0	1	1	0
GS0354-1	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	1	1	1	1	1	1	0
GS0481-1	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	1	1	1	1	1	0
GS0845-1	stx1a	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1
IHIT1706	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
IHIT2087	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	0	1
IMT15301	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0
IMT15302	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0
IMT19623	0	0	1	1	1	1	1	0	1	1	1	0	1	0	0	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
IMT19636	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
IMT19648	0	stx2a	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
IMT19947	0	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0
IMT19981	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
IMT19996	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
IMT20165	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	1
IMT20217	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
IMT20239	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0	0	1	1	1
IMT20245	0	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
IMT20337	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
LB228716	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	1	0	1	0	1	0	0	1	0	0	1	1	1
LB234771	stx1a	0	1	1	1	1	0	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
LB238877	stx1a	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
LB267954	stx1a	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	1	1
LB275203_1	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0	1	0	0	1	0	0	0	1	0	0	0	1	1	1
LB279967	0	stx2a	1	1	1	1	0	0	1	1	1	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	0	1	0	0	0	1	1	0
LB363084	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	0	0	0	0	1	0	0	0	0	1	0	0	1	1	1
LB363104	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	0	0	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0

Strain designation	*stx1*	*stx2*	*eae*	*tir*	*nleA*	*nleB*	*nleC*	*iha*	*espA*	*espB*	*espF*	*espI*	*espJ*	*espP*	*efa1*	*ehxA*	*astA*	*cif*	*lpfA*	*epeA*	*toxB*	*iroN*	*etpD*	*iss*	*cba*	*cma*	*celb*	*tccp*	*mchF*	*mchC*	*mchB*	*prfB*	*gad*	*katP*
LB366332i2	stx1a	stx2a	1	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
LB369997	0	stx2a	1	1	1	1	0	0	1	1	1	0	1	0	0	1	1	1	1	0	0	0	1	1	1	1	0	0	0	0	0	1	1	0
LH-1	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	1	1	0	1	1	0
LH-8	0	0	1	1	1	1	0	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	1	1	0
P2332/07-1	0	stx2a	1	1	1	1	0	1	1	1	1	0	1	0	1	1	1	1	1	0	0	0	1	1	1	1	1	0	0	0	0	1	1	0
P3146/08-1	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	1	0	1	1	0	0	1	1	1
P3146/08-2	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	0	0	0	0	1	0	1	0	1	1	0	0	1	0	1
P3146/08-3	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	1	0	1	0	0	1	0	0	1	0	1
P355/07	stx1a	stx2a	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
P4352/06-1	0	stx2a	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	0	1	1	1	1	0	0	0	0	1	1	1
P4501/05-1	stx1a	stx2a	1	1	1	1	1	1	1	0	1	0	1	0	0	1	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	1	0
P7278/07	stx1a	stx2a	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	0	1	0	1
P7404/03-1	stx1a	0	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
P8054/02	0	stx2a	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
RW0145	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	0
RW1300	stx1a	0	1	1	1	1	1	1	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
RW1703	stx1a	0	1	1	1	1	1	1	0	0	0	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
RW1706	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0	1	0	0	1	0	0	0	0	1	1	1
RW1993	stx1a	0	1	1	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
RW2070	0	0	1	1	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	1	1	0
RW2174	stx1a	0	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0

Table S 5 Information on genes that include ≥5 SNP-sites in the alignment of the MCG of DEC9D and RW2070 that differ in 3,059 SNPs of which 1,933 SNPs (1,174 synonymous SNPs (synSNPs); 759 non-synonymous SNPs (nonsynSNPs)) of 53 genes in total were further analyzed.

Gene name	Product	length [bp]	Identity [%]	syn SNPs	nonsyn SNPs
alkA	3-methyl-adenine DNA glycosylase II, inducible	844	96,7	7	20
amn	Amp nucleosidase	1.454	98,1	25	0
BX75_08170	Type III secretion protein	321	95	13	4
cesD	CesD protein	457	93,9	20	6
cesD2	T3SS chaperone CesD2	408	97,8	7	2
cesT	CesT protein	465	95,4	1	20
cobT	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase	1.080	98,1	14	6
dcd	Deoxycytidine triphosphate deaminase	582	97,9	11	0
dcyD	D-cysteine desulfhydrase	954	98	16	2
ECO9942_15685	T3SS component	429	94,1	28	10
erfK	Probable L,D-transpeptidase ErfK/SrfK	945	98,5	10	4
erfK	L,D-transpeptidase	945	98,5	8	3
escD	Type III secretory protein EscD	1.183	95,5	37	16
escF	T3SS structure protein EscF	222	97,3	6	0
espA	Secretion protein EspA	578	90,5	20	35
espD	Secretion protein EspD	1.143	90,5	77	86
fliD	Flagellar hook-associated protein 2	328	98,2	2	7
fliY	Cysteine transport protein ABC superfamily	800	98,6	9	2
fliZ	Flagella biosynthesis protein FliZ	552	98,4	5	4
gcvR	DNA-binding transcriptional repressor,regulatory protein accessory to GcvA	285	97,2	0	13
gnd	6-phosphogluconate dehydrogenase, decarboxylating	1.368	85	168	35
grlR	Negative regulator GrlR	351	97,3	2	12
hisC	Histidinol-phosphate aminotransferase	1.049	87,5	139	58
hisF	Imidazole glycerol phosphate synthase subunit HisF	777	88,2	79	11
hisG	ATP phosphoribosyltransferase	899	98,2	15	0
ler	LEE encoded regulator	373	96,4	17	3
mpc	T3SS regulator Mpc	357	94,4	25	6

Gene name	Product	length [bp]	Identity [%]	syn SNPs	nonsyn SNPs
mutY	A/G-specific adenine glycosylase	1.830	98,9	19	1
nac	Nitrogen assimilation control protein	918	97,9	18	1
plaP	Putrescine/spermidine ABC transporter	1.285	98,1	14	0
sbcB	Exodeoxyribonuclease I	1.409	98,3	28	7
sdiA	DNA-binding transcriptional activator	723	98,9	4	4
sepZ	SepZ protein	302	81,7	5	71
sfmH	fimbrial-like protein SfmH	1.000	97,6	0	22
shiA	Shikimate and dehydroshikimate transport protein MFS family protein ShiA	1.317	96	14	36
tonB	Energy transducer TonB	720	96,7	0	24
tsx	Nucleoside channel	872	99,5	1	7
udk	Uridine kinase	642	99,1	4	2
wcaD	Colanic acid biosynthesis protein	1.217	87,7	155	66
wcaF	Colanic acid biosynthesis acetyltransferase WcaF	549	81,6	8	91
wza	Lipoprotein required for capsular polysaccharide translocation through the outer membrane	1.056	97,2	23	7
wzb	Low molecular weight protein-tyrosine-phosphatase	440	97,6	8	5
yecC	ABC transporter ATP-binding protein	753	98,9	8	0
yecS	Amino acid ABC transporter permease	669	98,4	11	0
yeeN	Probable transcriptional regulatory protein YeeN	654	98,8	9	3
yeeU_4	Antitoxin of the YeeV-YeeU toxin-antitoxin system	369	97,5	5	3
yeeY	Bacterial regulatory helix-turn-helix , lysR family protein	930	96,8	28	2
yeeZ	Nucleoside-diphosphate-sugar epimerases	786	98,9	7	2
yegH	Fused predicted membrane protein/predicted membrane protein	1.431	99,2	14	3
yegK	Serine/threonine protein phosphatase	681	94,1	22	18
yfhL	4Fe-4S cluster-containing protein	250	98,4	0	5
yhaI	Inner membrane protein YhaI	357	95,5	8	8
yijO	AraC family transcriptional regulator	851	99,5	0	6

Table S 6 Information on genes that include ≥5 SNP-sites in the alignment of the MCG of DEC9D and LH-1 that differ in 1,767 SNPs of which 979 SNPs (830 synonymous SNPs; 149 non-synonymous SNPs) of 62 genes in total were further analyzed.

Gene name	Product	length [bp]	Identity [%]	syn SNPs	nonsyn SNPs
bglJ	2-component transcriptional regulator	678	99,4	4	2
creB	DNA-binding response regulator CreB	690	99	4	3
creD	Inner membrane protein	1.281	99,3	6	3
deoA	Thymidine phosphorylase	1.323	98,9	15	0
deoB	Phosphopentomutase	1.224	98,8	14	1
deoC	Deoxyribose-phosphate aldolase	779	98,6	11	0
deoD	Purine nucleoside phosphorylase DeoD-type	693	98,8	7	1
ECO9942_06936	Transcriptional activator of 4-hydroxyphenylacetate 3-monooxygenase operon, XylS/AraC family	891	98,3	14	1
ECO9942_07006	DNA-binding transcriptional regulator	129	84,5	16	4
ECO9942_17373	Alcohol dehydrogenase	1.020	97,6	16	8
fhuF	Ferric iron reductase	789	97	20	4
fimA	Fimbrial subunit type 1	534	97,1	13	10
fimC	Chaperone FimC	676	99	7	0
fimE	type 1 fimbriae regulatory protein FimE	480	98,8	6	0
fimG	Fimbrial protein FimG	504	98,8	4	2
fimH	Minor fimbrial subunit, D-mannose specific adhesin	903	98,2	13	3
G925_04592	Lipopolysaccharide export system permease lptG	1.083	98,8	13	0
gsiA	Glutathione ABC transporter ATP-binding protein	1.874	99,8	5	0
hisJ	Histidine ABC transporter substrate-binding protein HisJ	781	98,1	15	0
hpaB	4-hydroxyphenylacetate 3-monooxygenase	1.563	98,7	21	0
hpaC	4-hydroxyphenylacetate 3-monooxygenase reductase componen	474	98,5	6	1
hpaD	3,4-dihydroxyphenylacetate 2,3-dioxygenase	807	97,8	17	1
hpaE	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase	1.467	96,9	45	0
hpaH	2-oxo-hepta-3-ene-1,7-dioate hydratase	804	97,1	23	0
hpaX	4-hydroxyphenylacetate permease	1.377	98,9	12	3
hpcE	Homoprotocatechuate catabolism bifunctional isomerase/decarboxylase	1.256	98,9	14	6
iraD	Anti-adapter protein IraD	398	97,9	2	9
kptA	RNA 2'-phosphotransferase	555	97,7	10	3
lplA	Lipoate-protein ligase A	1.017	98,7	10	3
lptF	Lipopolysaccharide export system permease lptF	1.008	99	7	3
mdoB	Phosphoglycerol transferase I	2.253	99,2	18	1

Gene name	Product	length [bp]	Identity [%]	syn SNPs	nonsyn SNPs
mgtA	Magnesium ABC transporter ATPase	2.697	99,4	23	1
phoB	DNA-binding response regulator in two-component regulatory system with PhoR (Or CreC)	1.425	97,9	25	5
prfC	Peptide chain release factor 3	1.533	98,7	27	0
qseD	Cell density-dependent motility repressor	880	99,3	8	1
rob	Right oriC-binding transcriptional activator,AraC family	870	98,7	11	0
serB	3-phosphoserine phosphatase	887	98,6	18	1
smp	Bacterial virulence factor haemolysin family protein	567	97,5	14	0
sms	DNA repair protein radA	1.383	97,5	34	0
thrB	Homoserine kinase	897	98,4	13	1
thrC	threonine synthase	1.158	99,6	7	0
treR	DNA-binding transcriptional repressor	948	98,3	14	3
tsr	Methyl-accepting chemotaxis protein I	1.677	99	15	1
uxuA	Mannonate dehydratase	1.185	99,2	9	0
uxuB	D-mannonate oxidoreductase	1.461	98,9	22	2
uxuR	Fructuronate-inducible hexuronate regulon transcriptional repressor	726	99,2	5	1
yaaH	Conserved inner membrane protein	567	98,8	7	0
yaaJ	Transporter YaaJ	1.329	98,3	11	2
yeeU_3	Antitoxin of the YeeV-YeeU toxin-antitoxin system	369	97,5	6	4
yhaI	Inner membrane protein YhaI	357	95,5	8	8
yjgR	ATPase	1.503	96,4	44	10
yjhA	N-acetylneuraminic acid outer membrane porin	717	99,2	5	1
yjiA	GTP-binding protein	957	96,4	27	7
yjiG	Conserved predicted inner membrane protein	463	99,1	5	1
mdtM	Multidrug resistance protein MdtM	1.233	99	17	2
yjjG	DUMP phosphatase	678	99,4	4	0
yjjI	Glycine radical enzyme, 2C YjjI family	1.551	99,3	10	7
yjjK	ABC transporter ATP-binding protein	1.514	99,6	9	0
yjiM	2-hydroxyglutaryl-CoA dehydratase	1.152	98,7	17	5
yjjP	Inner membrane protein YjjP	771	98,7	12	3
yjjV	DNase	783	98	8	8
yjjX	Non-canonical purine NTP phosphatase	522	98,7	7	3

Table S 7 Information on genes that include ≥5 SNP-sites in the alignment of the MCG of RW2070 and DEC8C that differ in 3,916 SNPs of which 2,271 SNPs (1,619 synonymous SNPs; 652 non-synonymous SNPs) of 115 genes in total were further analyzed.

Gene name	Product	length [bp]	Identity [%]	syn SNPs	nonsyn SNPs
alkA	3-methyl-adenine DNA glycosylase II, inducible	846	98,1	13	11
amn	Amp nucleosidase	1.455	98,1	27	0
asmA	Assembly protein AsmA	1.854	99	17	1
bglJ	2-component transcriptional regulator	678	99	3	3
BX75_08170	Type III secretion protein	321	96,9	11	4
cesD	CesD protein	457	93,9	20	8
cesD2	T3SS chaperone CesD2	408	98,4	7	3
cesT	CesT protein	467	95,3	19	3
cobT	Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase	1.080	98,1	14	6
cobU	Adenosylcobinamide kinase	543	98,2	7	3
creB	DNA-binding response regulator CreB	690	99	3	3
creD	Inner membrane protein	1.353	99,3	7	3
dcd	Deoxycytidine triphosphate deaminase	582	97,9	11	2
dcyD	D-cysteine desulfhydrase	987	98,1	17	2
deoA	Thymidine phosphorylase	1.323	98,8	16	0
deoB	Phosphopentomutase	1.224	99,2	15	0
deoC	Deoxyribose-phosphate aldolase	780	98,6	12	0
deoD	Purine nucleoside phosphorylase DeoD-type	720	99	6	1
EcE24377A_2152	L-cystine ABC transporter, permease protein	669	98,4	11	0
ECO9942_06936	Transcriptional activator of 4-hydroxyphenylacetate 3-monooxygenase operon, XylS/AraC family	891	98,3	14	1
ECO9942_07006	DNA-binding transcriptional regulator	129	84,5	16	4
ECO9942_12027	Transporter	1.431	98,4	20	2
ECO9942_13932	PTS system galactitol-specific transporter subunit IIB	285	97,9	2	4
ECO9942_15685	T3SS component	429	94,1	28	10
ECO9942_17373	Alcohol dehydrogenase	1.020	97,6	16	8
erfK	L,D-transpeptidase	948	98,6	11	4
escD	Type III secretory protein EscD	1.182	97	33	12
escF	T3SS structure protein EscF	222	97,3	6	0
espA	Secretion protein EspA	579	90,5	20	35
espD	Secretion protein EspD	1.143	90,4	74	86
fabG	3-ketoacyl-ACP reductase	762	97,9	13	3
fbaB	Fructose-bisphosphate aldolase	1.053	98,9	12	0
fdnG	Formate dehydrogenase, nitrate-inducible, major subunit	588	74,5	83	67
fepA	Ferrienterobactin receptor	1.484	99,8	5	0
fhuF	Ferric iron reductase	789	97	11	4
fimA	Fimbrial subunit type 1	534	97,1	13	10
fimC	Chaperone FimC	726	99,4	5	2
fimE	Phage integrase family protein	480	98,9	6	0
fimG	Fimbrial protein FimG	504	98,8	4	2
fimH	Minor fimbrial subunit, D-mannose specific adhesin	903	98,2	13	3
fliD	Flagellar hook-associated protein 2	981	98,2	3	8
fliY	Cysteine transport protein ABC superfamily	801	98,6	9	2
fliZ	Flagella biosynthesis protein FliZ	552	98,4	5	4
G925_04592	Lipopolysaccharide export system permease lptG	1.083	98,8	13	0
galS	GalS transcriptional repressor and galactose ultrainduction factor	1.041	99,2	7	2
grlR	Negative regulator GrlR	339	97,4	6	7
hisG	ATP phosphoribosyltransferase	900	99,1	12	0
hpaB	4-hydroxyphenylacetate 3-monooxygenase	1.563	98,6	21	1
hpaC	4-hydroxyphenylacetate 3-monooxygenase reductase componen	474	98,5	6	1
hpaD	3,4-dihydroxyphenylacetate 2,3-dioxygenase	852	97,9	17	1
hpaE	5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase	1.467	96,9	46	1
hpaH	2-oxo-hepta-3-ene-1,7-dioate hydratase	804	97,1	23	0
hpaX	4-hydroxyphenylacetate permease	1.377	99,2	14	2
hpcE	Homoprotocatechuate catabolism bifunctional isomerase/decarboxylase	1.256	98,9	13	6
iraD	Anti-adapter protein IraD	393	96,8	3	8
kptA	RNA 2'-phosphotransferase	555	97,7	6	2
ler	LEE encoded regulator	373	96,4	17	3
lplA	Lipoate-protein ligase A	1.017	98,8	10	2

Gene name	Product	length [bp]	Identity [%]	syn SNPs	nonsyn SNPs
mdoB	Phosphoglycerol transferase I	2.253	99,2	28	1
mdtC	Multidrug resistance protein MdtC	1.026	99,1	12	0
mdtM	Multidrug resistance protein MdtM	1.233	99	17	2
mdtP	Multidrug resistance outer membrane protein MdtP	1.459	99,8	0	5
metG	Methionine--tRNA ligase	2.034	99,8	7	0
mgtA	Magnesium ABC transporter ATPase	2.697	99,1	22	1
mpc	T3SS regulator Mpc	355	94,1	24	6
nac	Nitrogen assimilation control protein	918	97,9	18	0
phoB	DNA-binding response regulator in two-component regulatory system with PhoR (Or CreC)	1.425	98	23	6
plaP	Putrescine/spermidine ABC transporter	1.359	99	21	0
prfC	Peptide chain release factor 3	1.539	98,7	27	1
qseD	Cell density-dependent motility repressor	882	98,9	6	9
rob	Right oriC-binding transcriptional activator,AraC family	870	98,7	11	0
sbcB	Exodeoxyribonuclease I	1.409	97,8	42	4
sdiA	DNA-binding transcriptional activator	723	98,9	4	4
sepZ	SepZ protein	303	82	16	66
serB	3-phosphoserine phosphatase	887	98,6	17	0
shiA	Shikimate and dehydroshikimate transport protein MFS family protein ShiA	1.317	96	20	33
smp	Bacterial virulence factor haemolysin family protein	645	97,7	15	0
sms	DNA repair protein radA	1.383	97,5	4	30
thiM	Hydroxyethylthiazole kinase	789	99,1	4	3
thrB	Homoserine kinase	933	98,5	13	1
thrC	threonine synthase	1.158	99,6	7	0
treR	DNA-binding transcriptional repressor	948	98,3	13	3
tsr	Methyl-accepting chemotaxis protein I	1.677	99	15	2
tsx	Nucleoside channel	877	99,5	1	7
udk	Uridine kinase	642	99,3	5	2
uxuA	Mannonate dehydratase	1.185	99,2	9	0
uxuB	D-mannonate oxidoreductase	1.461	98,8	23	4
uxuR	Fructuronate-inducible hexuronate regulon transcriptional repressor	726	99,2	5	1
yaaH	Conserved inner membrane protein	567	98,8	7	0
yecC	ABC transporter ATP-binding protein	753	98,9	8	0
yeeD	Membrane protein	228	97,7	6	2
yeeN	Probable transcriptional regulatory protein YeeN	717	98,5	11	0
yeeO	MATE efflux family protein	1.457	99,4	9	5
yeeY	Bacterial regulatory helix-turn-helix , lysR family protein	930	98,8	14	3
yegH	Fused predicted membrane protein/predicted membrane protein	1.438	99	17	3
yegK	Serine/threonine protein phosphatase	651	93,9	22	18
yegT	Nucleoside transporter	1.248	98,5	15	4
yegU	ADP-ribosylglycohydrolase	1.005	98,1	17	2
yehA	Predicted fimbrial-like adhesin protein	1.040	98,5	14	9
yehC	Periplasmic chaperone YehC	675	98,5	10	0
yehT	Putative response regulator in two-component system with YehU	720	99,2	6	0
yehY	ABC transporter membrane protein	1.158	99,4	7	3
yeiB	Inner membrane protein	1.083	99,4	6	3
yhdG	tRNA-dihydrouridine synthase	951	98,9	5	5
yjgP	Lipopolysaccharide ABC transporter permease	1.101	99,4	9	1
yjgR	ATPase	1.503	97,6	45	8
yjhA	N-acetylneuraminic acid outer membrane porin nanC	717	99,2	5	1
yjhS	9-O-acetyl-N-acetylneuraminic acid deacetylase nanS	981	98,6	9	11
yjiA	GTP-binding protein	957	96,4	27	7
yjjI	Glycine radical enzyme, 2C YjjI family	1.551	99,3	10	6
yjjK	ABC transporter ATP-binding protein	1.519	99,6	10	0
yjjV	DNase	783	98,2	8	6
yjjX	Non-canonical purine NTP phosphatase	522	98,1	7	3
yohD	Inner membrane protein	579	99	9	0
yohJ	membrane protein YohJ	399	98,5	5	1

Table S 8 List of 63 insertion sites investigated for the presence of mobile genetic elements stating gene name and gene product.

Insertion site	Gene product
argU	tRNA
argW	tRNA
asnT	tRNA
betA	choline dehydrogenase
btuC	vitamin B12-transporter permease
clpA	ATP-binding component of serine protease
cpxP	periplasmic protein combats stress CpxP
cspD	cold shock protein
eutB	ethanolamine ammonia-lyase, large subunit
fimA	major type 1 subunit fimbrin (pilin) FimA
fimC	type I fimbrial chaperone
gcd	glucose dehydrogenase
glyU	tRNA
guaA	GMP synthase
icdA	isocitrate dehydrogenase, specific for NADP+
ileX	tRNA
leuX	tRNA
lysT	tRNA
ompA	outer membrane protein 3a (II*;G;d)
pauD	tRNA
pheU, pheV	tRNA
phoH	PhoB-dependent, ATP-binding pho regulon component
pntB	pyridine nucleotide transhydrogenase, beta subunit
potC	polyamine transporter subunit PotC
ppa	inorganic pyrophosphatase
prfC	peptide chain release factor RF-3
proL	tRNA
ptwF	tRNA
rcnA	membrane protein conferring nickel and cobalt resistance
rspB	putative Zn-dependent NAD(P)-binding oxidoreductase
ryeA	sRNA

Insertion site	Gene product
ryeB	sRNA
sbcB	deoxyribophosphodiesterase
selC	tRNA
serC	3-phosphoserine aminotransferase
serU	tRNA
ssrA	tmRNA
thrW	tRNA
wrbA	trp repressor binding protein
yaeF	predicted lipoprotein
yagQ	hypothetical protein
ybaZ	hypothetical protein
ybhC	predicted pectinesterase
ybjK	transcriptional regulator of csgD and ybiJI
yccA	putative carrier/transport protein
ycfD	hypothetical protein
yciD	outer membrane protein
ydaK	putative transcriptional regulator LYSR-type
ydaO	hypothetical protein
ycdU	putative ABC transporter permease
yecA	hypothetical protein
yecE	hypothetical protein
yecN	hypothetical protein
yeeH	pseudogene
yehV	MerR family transcriptional regulator
yfaT	hypothetical protein
yfhL	4Fe-4S cluster-containing protein
ygjH	siderophore-interacting protein
yicC	putative alpha helix protein
yjbM	conserved predicted protein, C-terminal fragment
ynfA	UPF0060 family inner membrane protein
ynfH	oxidoreductase, membrane subunit

Table S 9 Occupied insertion sites of DEC9A. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argW	Entero_HK620_NC_002730	none
btuC	Entero_P88_NC_026014	none
eutB	cryptic_prophage_Eut/CPZ-55_NC_000913.3	none
fimC	CP_933V_NC_002655	none
leuX	KpLE2_phage-like_element_CU928164	none
pheU, pheV	Entero_mEp460_NC_019716	LEE PAI
potC	ECO26_P05_NC_013361	none
rspB	Entero_lambda_NC_001416	none
ssrA	Salmon_RE_2010_NC_019488	none
yaeF	Entero_HK630_NC_019723	none
ybhC	Escher_TL_2011b_NC_019445	none
ycfD	CP_933C_NC_002655	none
yfaT	ECO26_P17_NC_013361.1	none
ygjH	Entero_P4_NC_001609	genomic_island_GEI3.21_AB426061.1
yjbM	ECO111_P17_NC_013364.1	none

Table S 10 Occupied insertion sites of DEC9B. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argW	ECO26_P18_NC_013361.1	none
btuC	Entero_P88_NC_026014	none
eutB	cryptic_prophage_Eut/CPZ-55_NC_000913.3	none
fimA	CP_933V_NC_002655	none
guaA	Entero_phiV10_NC_007804	none
leuX	none	ECO111_IE07_NC_013364.1
pheU, pheV	Entero_mEp460_NC_019716	LEE PAI
potC	CP_933N_NC_002655	none
proL	RM13514_P06_CP006027.1	none
rspB	Entero_lambda_NC_001416	none
serU	none	ECO26_IE03_NC_013361
ssrA	Salmon_RE_2010_NC_019488	none
wrbA	ECO26_P04_NC_013361.1	none
yaeF	Entero_HK630_NC_019723	ECO26_IE02_NC_013361
ybhC	ECO26_P02_NC_013361.1	none
ydaK	CP_933P_NC_002655	none
yfaT	ECO111_P15_NC_013364.1	none
ygjH	none	genomic_island_GEI3.21_AB426061.1
yjbM	Entero_mEp460_NC_019716	none

Table S 11 Occupied insertion sites of DEC9C. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argW	Entero_Sf6_NC_005344	none
btuC	Entero_P88_NC_026014	none
eutB	cryptic_prophage_Eut/CPZ-55_NC_000913.3	none
fimA	Entero_BP_4795_NC_004813	none
icdA	Entero_YYZ_2008_NC_011356	none
leuX	none	ECO111_IE07_NC_013364.1
pauD	Entero_lambda_NC_001416	none
pheU, pheV	Entero_BP_4795_NC_004813	LEE PAI
pntB	Shigel_SflV_NC_022749	none
rspB	Entero_mEp460_NC_019716	ECO111_IE05_NC_013364
ssrA	Entero_Fels_2_NC_010463	none
yaeF	none	ECO26_IS621_NC_013361
ybhC	Entero_HK629_NC_019711	none
yeeH	none	ECO26_IS03_NC_013361
yfaT	Entero_lambda_NC_001416	none
ygjH	none	genomic_island_GEI3.21_AB426061.1
yjbM	Stx2_converting_I_NC_003525	none

Table S 12 Occupied insertion sites of DEC9D. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argU	Stx2_converting_I_NC_003525	none
argW	Entero_Sf6_NC_005344	none
btuC	Entero_P88_NC_026014	none
cpxP	Yersin_L_413C_NC_004745	none
fimC	Entero_BP_4795_NC_004813	none
guaA	Aeromon_schubertii_CP013067	none
leuX	none	ECO111_IE07_NC_013364.1
pauD	Entero_BP_4795_NC_004813	none
pheU, pheV	Stx2_converting_1717_NC_011357	LEE PAI
potC	Entero_YYZ_2008_NC_011356	none
rspB	Entero_lambda_NC_001416	none
serU	Entero_BP_4795_NC_004813	none
ssrA	Salmon_RE_2010_NC_019488	ECO26_IE05_NC_013361
yaeF	none	ECO26_IS621_NC_013361
ybhC	Entero_BP_4795_NC_004813	none
ydaK	Entero_YYZ_2008_NC_011356	none
yfaT	Stx2_converting_1717_NC_011357	none
yfhL	Aggreg_S1249_NC_013597	none
ygjH	Entero_P4_NC_001609	none
yjbM	Entero_mEp460_NC_019716	none

Table S 13 Occupied insertion sites of DEC9E. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argW	Entero_Sf6_NC_005344	none
btuC	Entero_P88_NC_026014	none
fimC	Entero_BP_4795_NC_004813	none
leuX	none	ECO111_IE07_NC_013364.1
pauD	Entero_lambda_NC_001416	none
pheU, pheV	none	LEE PAI
potC	Entero_lambda_NC_001416	none
rspB	Entero_BP_4795_NC_004813	none
ryeA		none
serU	Entero_lambda_NC_001416	none
ssrA	Salmon_RE_2010_NC_019488	ECO26_IE05_NC_013361
ybhC	Escher_HK75_NC_016160	none
yciD	ECO26_P08_NC_013361	none
ydaK	CP_933C_NC_002655	none
yfaT	Entero_HK630_NC_019723	none
ygjH	none	genomic_island_GEI3.21_AB426061.1
yjbM	Entero_YYZ_2008_NC_011356	none

Table S 14 Occupied insertion sites of LH-8. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argU	Stx2_converting_I_NC_003525	none
argW	Stx2_converting_I_NC_003525	none
btuC	Entero_P88_NC_026014	none
eutB	YPDSF_3775_phage_integrase_CP	none
leuX	none	ECO111_IE07_NC_013364.1
pauD	Stx2_converting_I_NC_003525	none
pheU, pheV	none	LEE PAI
potC	Entero_lambda_NC_001416	none
proL	RM13516_P04_CP006262	none
rspB	ECO111_P09_NC_013364.1	none
ssrA	Entero_Fels_2_NC_010463	ECO111_IE04_NC_013364
ybhC	ECO26_P02_NC_013361	none
ydaK	ECO26_P08_NC_013361	none
yfaT	ECO26_P17_NC_013361.1	none
ygjH	none	genomic_island_GEI3.21_AB426061.1
yjbM	ECO111_P17_NC_013364	none

Table S 15 Occupied insertion sites of IMT19623. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argW	ECO26_P18_NC_013361	none
btuC	Entero_P88_NC_026014	none
cpxP	Yersin_L_413C_NC_004745	none
eutB	YPDSF_3775_phage_integrase_CP000946	none
fimA	CP_933V_NC_002655	none
leuX	none	ECO111_IE07_NC_013364.1
pauD	Entero_lambda_NC_001416	none
pheU, pheV	Entero_mEp237_NC_019704	LEE PAI
potC	Shigel_POCJ13_NC_025434	none
proL	RM13514_P04_CP006262	none
serU	none	ECO26_IE03_NC_013361
ssrA	Salmon_RE_2010_NC_019488	none
ydaK	ECO26_P08_NC_013361	none
yfaT	ECO26_P17_NC_013361.1	none
ygjH	none	genomic_island_GEI3.21_AB426061.1
yjbM	Stx2_converting_1717_NC_011357	none

Table S 16 Occupied insertion sites of RW2070. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argU	ECO111_P01_NC_013364.1	none
asnT	none	IE ECO111_IE03_NC_013364.1
btuC	PHAGE_Entero_P88_NC_026014	none
leuX	none	IE ECO111_IE07_NC_013364.1
pheU, pheV	none	LEE PAI
phoH	none	IE ECO111_IE02_NC_013364.1
potC	ECO111_P05_NC_013364.1	none
proL	ECO111_P14_NC_013364.1	none
ptwF	Entero_mEp460_NC_019716	none
rspB	ECO111_P09_NC_013364.1	none
sbcB	Entero_Sf101_NC_027398	none
serU	ECO111_P12_NC_013364.0	none
ssrA	ECO111_P16_NC_013364.1	none
ybhC	Entero_Sf6_NC_005344	none
yccA	ECO111_P03_NC_013364.1	none
yciD	ECO111_P08_NC_013364.1	none
ydaO	Shigel_POCJ13_NC_025434	none
yfaT	ECO111_P15_NC_013364.1	none
yicC	ECO103_P14_NC_013353.1	none

Table S 17 Occupied insertion sites of LH-1. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argW	Entero_Sf6_NC_005344	none
asnT	none	ECO26_IE03_NC_013361
btuC	Entero_P88_NC_026014	none
glyU	none	ECO26_IE06_NC_013361
leuX	ECO26_P21_NC_013361	none
pauD	Stx2_converting_I_NC_003525	none
pheU, pheV	Entero_lambda_NC_001416	LEE PAI
phoH	none	ECO26_IE02_NC_013361
rspB	ECO111_P09_NC_013364.1	none
selC	Entero_YYZ_2008_NC_011356	none
serU	PHAGE_Pectob_ZF40_NC_019522	none
ssrA	none	ECO26_IE05_NC_013361
yagQ	none	ECO26_IE01_NC_013361
ybhC	ECO26_P02_NC_013361	none
yccA	Salmon_ST64B_NC_004313	none
ycfD	CP_933C_NC_002655	none
yciD	ECO26_P08_NC_013361	none
ydaO	Gifsy_2_NC_010393	none
yehV	Entero_mEp460_NC_019716	none
yfaT	Stx2_converting_I_NC_003525	none
yfhL	Salmon_SEN34_NC_028699	none
yjbM	Entero_mEp460_NC_019716	none

Table S 18 Occupied insertion sites of 0739/03. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argU	ECO26_P01_NC_013361	none
argW	Entero_Sf6_NC_005344	ECO26_IE04_NC_013361
btuC	Entero_P88_NC_026014	ECO26_IE08_NC_013361
cspD	ECO26_P12_NC_013361	none
glyU	none	ECO26_IE06_NC_013361
leuX	ECO26_P21_NC_013361	none
pheU, pheV	none	LEE PAI
phoH	none	ECO26_IE02_NC_013361
pntB	ECO26_P12_NC_013361	none
potC	ECO111_P05_NC_013364.1	none
proL	ECO26_P16_NC_013361	none
rspB	ECO111_P09_NC_013364.1	none
ryeB	ECO26_P14_NC_013361	none
selC	Entero_P4_NC_001600	none
ssrA	ECO26_P19_NC_013361	ECO26_IE05_NC013361
wrbA	Escher_P13374_NC_018846	none
yagQ	Entero_YYZ_2008_NC_011356	none
ybhC	ECO26_P02_NC_013361	none
yccA	ECO26_P03_NC_013361	none
ycfD	ECO26_P07_NC_013361	none
yciD	ECO26_P08_NC_013361	none
ydaO	ECO26_P09_NC_013361	none
yfaT	ECO26_P17_NC_013361	none
yfhL	Salmon_SEN34_NC_028699	none
yjbM	RM13514_P19_CP006027.1	none

Table S 19 Occupied insertion sites of IMT20337. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argU	ECO26_P01_NC_013361	none
argW	Entero_Sf6_NC_005344	ECO26_IE04_NC_013361
glyU	none	ECO26_IE06_NC_013361
leuX	ECO26_P21_NC_013361	none
phoH	none	ECO26_IE02_NC_013361
pntB	ECO26_P07_NC_013361	none
potC	ECO111_P05_NC_013364.1	nleG
proL	ECO26_P16_NC_013361	none
rspB	ECO111_P09_NC_013364.1	none
ryeB	ECO26_P14_NC_013361	none
selC	Entero_P4_NC_001609	none
ssrA	none	ECO26_IE05_NC_013361
yagQ	none	ECO26_IE03_NC_013361
ybhC	ECO26_P02_NC_013361	none
ycfD	ECO26_P07_NC_013361	none
yciD	ECO26_P08_NC_013361	none
ydaO	ECO26_P09_NC_013361	none
yecA	Entero_P88_NC_026014	none
yfaT	ECO26_P17_NC_013361	nleG

Table S 20 Occupied insertion sites of IMT19981. The ncbi reference name and number of the phages and insertion elements (IEs) are listed.

Insertion site	Phage Reference	other IEs
argU	ECO26_P01_NC_013360	none
argW	ECO26_P18_NC_013361	none
glyU	none	ECO26_IE01_NC_013361
leuX	Entero_BP_4795_NC_004813	none
pauD	Entero_BP_4795_NC_004813	none
phoH	none	ECO26_IE02_NC_013361
pntB	Shigel_SflV_NC_022749	none
potC	ECO111_P05_NC_013364.1	none
proL	ECO26_P16_NC_013361	none
rspB	ECO111_P09_NC_013364.1	none
ryeB	ECO26_P14_NC_013361	none
selC	Entero_P4_NC_001609	none
serU	none	ECO26_IE03_NC_013361
ssrA	ECO26_P19_NC_013361	none
yaeF	Entero_BP_4795_NC_004813	none
yagQ	none	ECO26_IE01_NC_013361
ybhC	ECO26_P02_NC_013361	none
yccA	ECO26_P03_NC_013361	none
ycfD	ECO26_P07_NC_013360	none
yciD	ECO26_P08_NC_013361	none
ydaO	Shigel_POCJ13_NC_025434	none
yecA	Entero_P88_NC_026014	none
yfaT	Entero_BP_4795_NC_004813	none

PUBLICATIONS

Inga Eichhorn, Katrin Heidemanns, Torsten Semmler, Bianca Kinnemann, Alexander Mellmann, Dag Harmsen, Muna F. Anjum, Herbert Schmidt, Angelika Fruth, Peter Valentin-Weigand, Jürgen Heesemann, Sebastian Suerbaum, Helge Karch, Lothar H. Wieler: *Highly Virulent Non-O157 Enterohemorrhagic Escherichia coli (EHEC) Serotypes Reflect Similar Phylogenetic Lineages, Providing New Insights into the Evolution of EHEC.* Applied Environmental Microbiology, 2015 Oct; 81(20):7041-7. doi: 10.1128/AEM.01921-15

Torsten Semmler, **Inga Eichhorn**, Astrid Bethe, Rolf Bauerfeind, Derek Pickard, Robert A. Kingsley, Gordon Dougan, Lothar H. Wieler: *Genome Sequence of Porcine Escherichia coli Strain IMT8073, an Atypical Enteropathogenic E. coli Strain Isolated from a Piglet with Diarrhea.* Genome Announcement, 2013 Aug 1; 1(4). doi: 10.1128/genomeA.00573-13.

Lothar H. Wieler, Torsten Semmler, **Inga Eichhorn**, Esther-Maria Antao, Bianca Kinnemann, Lutz Geue, Helge Karch, Sebastian Guenther, Astrid Bethe: *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 Pathogens, 2011 Nov 3; 3(1):17. doi: 10.1186/1757-4749-3-17

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.

SELBSTSTÄNDIGKEITSERKLÄRUNG

Hiermit bestätige ich, dass ich die vorliegende Arbeit selbstständig angefertigt habe. Ich versichere, dass ich ausschließlich die angegebenen Quellen und Hilfen in Anspruch genommen habe.

Inga Eichhorn
Berlin, den 09.06.2016