„Bioinformatics analyses of the *Escherichia coli* toxome"

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy

of Freie Universität Berlin

by

M.Sc. Susanne Fleischmann

from Berlin

2019

From the

Institute of Microbiology and Epizootics

Department of Veterinary Medicine

Freie Universität Berlin

04.2014 – 01.2019

Supervisor: Prof. Dr. Lothar H. Wieler

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

2nd reviewer: Univ.-Prof. Dr. Haike Antelmann

Disputation on: 23.05.2019

Abstract

Escherichia coli (*E. coli*) is both, a facultative commensal of the gut microbiota and an important bacterial cause of human as well as animal diseases. Due to the high genetic plasticity, gene transfer allows those bacteria to colonize different sites in the host acting as pathogens, being responsible for a wide range of infections [1, 67]. For toxins alone, variants have been described with respect to functional changes [44, 70, 132]. Based on of phylogenetic differences, the hypothesis of this thesis is that allelic variants of toxins differ in their biological effect with regard to their toxicity. Therefore, *in silico* analyses were performed to unravel sequence differences for all known toxins encoded in several *E. coli* and *Shigella* genomes.

The gene sequence research of all currently published toxins until June 2016 revealed a total set of 39 toxin genes being present in the genome of *E. coli* and *Shigella* spp. To test the hypothesis, the selected reference toxin sequences were used as templates to screen the NCBI database as well as an internal database using the BLAST algorithm. The internal database consists of 423 whole genome sequences of the “ECore” dataset [140] and 69 Shiga-Toxin-producing *E. coli* whole genome sequences [31]. To identify genetic variations of each toxin gene, the software Geneious was used, providing alignments and phylogenetic trees. After identification of allelic variations, the gene sequences were translated into their protein sequences. Finally, the resulting secondary and three-dimensional protein structures were predicted in order to receive indications of potentially functional changes. Detected allelic protein structures were modelled via the open source platform I-Tasser and visualised by the python based software PyMol.

The results show that genetic variations of all 39 toxin genes are available in both databases. The effector protein EspH and the metalloprotease YghJ were analyzed for the first time by their three-dimensional structures, with regard to possible functional changes for their resulting protein variants. Different variants of the Subtilase Cytotoxin SubAB were predicted and analyzed with regard to a different pathogenic potential and toxicity as was shown for the different Shiga-Toxin variants. Furthermore, the three-dimensional protein variants of the serine protease autotransporter of *Enterobacteriaceae* (SPATEs), indicate that the SPATE members are already allelic variations by name and split into two known functionally different classes [132]. The inclusion of the Shiga-Toxin variants [44, 111] revealed the same results, indicating that the chosen bioinformatics methods are sound and comparable. All the other identified allelic variations of toxins in the genome of *E. coli* and *Shigella* spp. due to *in silico* analyses had no impact on the resulting three-dimensional protein models, indicating a conserved function as in the case of EspF.

This thesis gives a first overview about the occurrence and impact of toxin variants in the genome of *E. coli* and *Shigella* spp. on the basis of bioinformatics analysis. The “toxome” allows the screening of all currently known toxin genes and its variants in the whole genome sequence of the mentioned bacteria with a BLASTn algorithm.

Zusammenfassung

Escherichia coli (*E. coli*) ist sowohl ein fakultativer Kommensal der Darmmikrobiota als auch ein bedeutender bakterieller Verursacher humaner sowie tierischer Erkrankungen. Aufgrund der hohen genetischen Plastizität ermöglicht der Gentransfer diesen Bakterien eine Kolonisierung verschiedenster Orte im Wirt, wo sie als pathogene Erreger ein weites Spektrum an Infektionen auslösen können [1, 67]. Für Toxine wurden Genvarianten beschrieben die funktionelle Veränderungen verursachen [44, 70, 132]. Anhand phylogenetischer Differenzen wurde die Hypothese entwickelt, dass sich Allele eines Toxins in ihrem biologischen Effekt und dadurch in ihrer Toxizität unterscheiden. Daher wurde eine *in silico* Analyse durchgeführt um Sequenzunterschiede aller bekannten Toxine die in verschiedenen *E. coli* und *Shigella* kodiert sind, zu bestimmen.

Die Gensequenzrecherche aller bekannten Toxine bis Juni 2016, welche im Genom von *E. coli* und *Shigella* spp. präsent sind, ergab ein Set von insgesamt 39 Toxingenen. Zur Überprüfung der Hypothese dienten ausgewählte Referenztoxinsequenzen als Vorlage zum Screening der NCBI Datenbank sowie einer internen Datenbank, welche 423 Ganzgenom Sequenzen aus dem "ECore" Datenset [140] und 69 Shiga-Toxin produzierende Ganzgenom sequenzierte *E. coli* [31] beinhaltet, unter Verwendung des BLAST-Algorithmus. Zur Identifikation genetischer Variationen jedes Toxingens wurden Alignments und phylogenetische Bäume mittels der Software Geneious erstellt. Nach der Identifizierung von Allelen wurden die Gensequenzen in ihre resultierenden Proteinsequenzen übersetzt um anschließend ihre sekundäre- und dreidimensionalen Proteinstrukturen zu modellieren, welche Indizien für mögliche funktionelle Veränderungen geben. Die ermittelten Proteinvarianten wurden unter Verwendung der frei verfügbaren Plattform I-Tasser modelliert und mit Hilfe der Python basierten Software Pymol visualisiert.

Die Ergebnisse zeigen, dass genetische Variationen von allen 39 Toxin Genen in beiden Datenbanken vorhanden sind. Das Effektor Protein EspH und die Metalloprotease YghJ wurden zum ersten Mal anhand ihrer dreidimensionalen Proteinstruktur auf funktionelle Veränderungen in ihren Proteinvarianten analysiert. Die unterschiedlichen Proteinvarianten des Subtilase Cytotoxins SubAB wurden bestimmt und hinsichtlich ihres unterschiedlichen Pathogenitätspotentials und ihrer Toxizität, wie es bei den Shiga-Toxin Varianten gezeigt wurde, analysiert. Des Weiteren zeigten die modellierten drei-dimensionalen Proteinvarianten der Serinen Protease Autotransporter der *Enterobacteriaceae* (SPATEs), dass die SPATEs bereits durch ihren Namen als Allele gekennzeichnet sind und sich in zwei funktionell unterschiedliche Klassen aufteilen [132]. Das Einbeziehen der bekannten Shiga-Toxin Varianten [44, 111] ergab unter Einbeziehung veröffentlichter Studien vergleichbare Ergebnisse, was die Zuverlässigkeit und Aussagekraft der gewählten bioinformatischen Methoden nahelegt. Alle anderen Toxinvarianten, welche im Genom von *E. coli* und *Shigella* spp. identifiziert wurden, haben *in silico* keinen Einfluss auf das resultierende dreidimensionale Proteinmodel und implizieren eine ähnliche Funktion wie im Falle von EspF.

Diese Arbeit liefert einen ersten Überblick über das Vorhandensein und den Einfluss von Toxinvarianten im Genom von *E. coli* und *Shigella* spp. auf der Grundlage bioinformatischer Analysen. Das „Toxom“ erlaubt ein Screening aller bekannter Toxingene sowie Toxinvarianten in der Genomsequenz der genannten Bakterien mittels BLASTn Algorithmus.

Content

Abstract I

Zusammenfassung II

Content IV

Tables VI

Figures VII

List of Abbreviations IX

1. Introduction 1

2. Literature 2

2.1 Characteristics of *Escherichia coli* and *Shigella* spp. 2

2.2 Transfer of virulence factors 4

2.3 Toxins as important virulence factors 6

2.4 Genetic variability of toxins 9

2.5 Background of bioinformatics algorithms 12

2.6 Basis of protein structure for 3D modeling 17

2.7 Influence of protein structure on biochemical properties 20

3. Material and methods 23

3.1 Material 23

3.1.1 Toxin reference sequences 23

3.1.2 Internal database ECore – *E. coli* Reference Collection 25

3.1.3 NCBI database - National Center for Biotechnology Information 26

3.2 Methods 27

3.2.1 Nucleotide BLAST (BLASTn) 27

3.2.2 Algorithm for alignment construction 28

3.2.3 Algorithm to create a phylogenetic tree 30

3.2.4 Prediction of secondary structure 30

3.2.5 Three-dimensional structure prediction and visualization 31

4. Results 33

4.1 Toxins of *E. coli* and *Shigella* spp. 33

4.2 Genetic variability of toxins expressed by *E. coli* and *Shigella* spp. 34

4.2.1 Intracellular acting toxins (IATs) 35
4.2.2 Membrane damaging toxins (MDTs) 40
4.2.3 Receptor targeted toxins (RTTs) 41
4.3 Three-dimensional structure prediction of selected protein variants 49
5. Discussion 59
5.1 Toxins of *E. coli* and *Shigella* spp. 59
5.2 Genetic variability of toxins in the genome of *E. coli* and *Shigella* spp. 60
5.2.1 Intracellular acting toxins (IATs) 60
5.2.2 Membrane damaging toxins (MDTs) 62
5.2.3 Receptor targeted toxins (RTTs) 63
5.3 Influence of bioinformatics methods on the results 66
5.3.1 Whole genome sequencing 67
5.3.2 Algorithm for alignment and phylogenetic tree construction 67
5.3.3 Prediction of secondary structures 69
5.3.4 Prediction of three-dimensional structures 70
6. Conclusion 72
7. Danksagung 75
8. Curriculum Vitae 76
9. Erklärung 77
10. References 78
11. Appendix i

Tables

Table 1: Classification of the serine protease autotransporters from Enterobacteriaceae (SPATEs)..... 11
Table 2: Popular web servers for remote homologous / fold recognition [61]..... 19
Table 3: Published toxin genes in the genome of *E. coli* and *Shigella*..... 23
Table 4: Occurrence of allelic variations among intracellular acting toxins (IATs)..... 35
Table 5: Occurrence of allelic variations among membrane damaging toxins (MDTs)..... 40
Table 6: Occurrence of allelic variations among receptor targeted toxins (RTTs)..... 41
Table 7: Toxins of *E. coli* and *Shigella* spp. exhibiting three-dimensional (3D) protein variants 49

Figures

Figure 1: Classification of extraintestinal pathogenic *E. coli* (ExPEC) and intestinal pathogenic *E. coli* (InPEC) [92, 58] 2
Figure 2: Pathogenic schema of intestinal pathogenic *E. coli* (InPEC) [92, 58] 3
Figure 3: Evolution of pathogenic *E. coli* via mobile genetic elements [58, 1] 4
Figure 4: Classification of bacterial toxins by their mechanisms on target cells [own model based on source [2] 8
Figure 5: Concept of the BLASTp algorithm [96, 3] 13
Figure 6: Concept of the BLASTn algorithm [3] 14
Figure 7: Definition of the global and local alignment [96, 129] 15
Figure 8: Hierarchical structures of the protein organization [76, 33] 17
Figure 9: Detailed pairwise alignment. 28
Figure 10: Nucleotide alignment 29
Figure 11: Protein alignment 29
Figure 12: Prediction of secondary structure 31
Figure 13: Predicted 3D protein structure. 32
Figure 14: Tree diagram to represent sequence analysis and phylogenetic distribution. 34
Figure 15: Phylogenetic tree of the SPATEs reference gene sequences 36
Figure 16: *sigA* (SPATE – Serine protease autotransporter of *Shigella flexneri* 2a) sequence analysis and phylogenetic distribution. 36
Figure 17: *subAB* (Subtilase Cytotoxin) sequence analysis and phylogenetic distribution 37
Figure 18: *stx1AB* (Shiga-Toxin 1) sequence analysis and phylogenetic distribution 38
Figure 19: *stx2AB* (Shiga-Toxin 2) sequence analysis and phylogenetic distribution 39
Figure 20: *clyA* / *hlyE* (Cytolysin A / Hemolysin E) sequence analysis and phylogenetic distribution 40
Figure 21: *espF* (T3SS - secreted effector protein) sequence analysis and phylogenetic distribution 43
Figure 22: *espH* (T3SS - secreted effector protein) sequence analysis and phylogenetic distribution 44
Figure 23: *yghJ* (Metalloprotease) sequence analysis and phylogenetic distribution. 48
Figure 24: Three-dimensional protein model of the passenger domain of class-1 (i. e. EspP) and class-2 SPATEs (i. e. Hbp and SepA) 50
Figure 25: Three-dimensional protein model of the SubAB variants 51
Figure 26: Three-dimensional protein model of the EspH variants 53
Figure 27: Three-dimensional protein model of the EspH protein variants in complex with human RhoGEF (1A2B) 55
Figure 28: Three-dimensional protein models of the YghJ metalloprotease variants 56

Figure 29: Three-dimensional protein models of the extracted M60-like pfam13402 domain of the zinc and nickel binding metalloprotease YghJ of *E. coli* 58

List of Abbreviations

3D	Three-dimensional
A/E	Attaching-and-effacing
AA	Amino acid
Acc. No.	Accession number
ADP	Adenosine diphosphate
APEC	Avian pathogenic *E. coli*
astA (EAST)	Heat-stable enterotoxin
BLAST	Basic Local Alignment Search Tool
BLASTn	Protein BLAST
BLASTp	Nucleotide BLAST
BLOSUM	Blocks Substitution Matrix
C1-INH	C1 esterase inhibitor
CASP	Critical Assessment of Structure Prediction
cdtVa-c / CdtVa-c	Cytoletal distending toxin
CFs	Colonization factors
cif / Cif	T3SS - Cycle-inhibiting factor
clbA-Q / ClbA-Q	Colibactin locus
clyA / *ClyA*	Cytolysin A
cnf1 / Cnf1	Cytotoxin necrotizing factor 1
cnf2 / Cnf2	Cytotoxin necrotizing factor 2
DAEC	Diffusely adherent *E. coli*
DNA	Desoxyribonucleic acid
E Value	Expect Value
E.	*Escherichia*
EAEC	Enteroaggregative *E. coli*
eatA / EatA	SPATEs - ETEC autotransporter A
ECore	*E. coli* Reference Collection
efa1 / Efa1	EHEC factor for adherence
EHEC	Enterohaemorrhagic *E. coli*
ehxA / EhxA	EHEC hemolysin
EIEC	Enteroinvasive *E. coli*
epeA / EpeA	SPATEs - EHEC plasmid encoded autotransporter
EPEC	Enteropathogenic *E. coli*
ER	Endoplasmic reticulum
espC / EspC	SPATEs - EPEC secreted protein C
espF / EspF	T3SS effector Protein
espH / EspH	T3SS effector Protein
espI / EspI	SPATEs - *E. coli* secreted protease
espP / EspP	SPATEs - extracellular serine protease (EHEC)
ETEC	Enterotoxigenic *E. coli*
Expec	Extraintestinal pathogenic *E. coli*
FASTA	text-based format representing sequences
Gb3	Globotriaosylceramide
GMP	Guanosine monophosphate
GOR method	Garnier-Osguthorpe-Robson method

GTP	Guanosine triphosphate
hbp / Hbp	SPATEs - hemoglobin binding protein
HGT	Horizontal gene transfer
HKY	Hasegawa, Kishino and Yano
hlyA / HlyA	Hemolysin A
hlyE / HlyE	Hemolysin E
HPI	High pathogenicity island
HUS	Hemolytic uremic syndrom
i-Tasser	Hierarchical protein 3D structure modeling approach
IATs	Intracellular acting toxins
IMD	Imidazol
InPEC	Intestinal pathogenic *E. coli*
ipaB / IpaB	T3SS - invasion plasmid antigen
*ipgB*2 / IpgB2	T3SS effector protein from *Shigella*
ipgD / IpgD	T3SS - Inosital phosphate phosphatase
LEE	Locus of Enterocyte Effacement
leoA / LeoA	Dynamin-like protein
lifA / LifA	Lymphocyte inhibitory factor
LPS	Lipopolysaccharide
LT	Heat-labile
LT (*elt*)	Heat-labile enterotoxin
map / Map	T3SS - mitochondrion-associated protein
MDTs	Membrane damaging toxins
MUC	Mucins
NA	Nucleic acid
NaCl	Sodium chloride
NCBI	National Center for Biotechnology Information
Neu5Gc	N-glycolylneuraminic acid
NGS	Next generation sequencing
NMEC	Neonatal meningitis *E. coli*
NMR	Nuclear Magnetic Resonance Spectroscopy
OEP	Outer membrane efflux protein
OM	Outer membrane
PAI I - IV	Pathogenicity island I – IV of UPEC
PAIs	Pathogenicity islands
PAM	Point accepted mutation
PCR	Polymerase-chain-reaction
PDB	Protein Data Bank
pEAF	Adhesion-factor plasmid
pENT	Enterotoxin-encoding plasmid
pet / Pet	SPATEs - plasmid encoded toxin
PFAM	Protein families
PFTs	Pore-forming toxins
Phyre	Protein Fold Recognition Server
pic / Pic	SPATEs - protease involved in intestinal colonization
pINV	Invasion-factor plasmid
PIR	Protein information Resource

pO157	EHEC strain O157 plasmid
pssA / PssA	SPATEs - protease secreted by STEC
PyMol	Molecular visualization software
RhoGEFs	Rho Guanine Nucleotide Exchange Factors
RKI	Robert Koch-Institute
RNA	Rinonucleic acid
RTTs	Receptor targeted toxins
RTX	Repeats-in-Toxin
S	Score
S.	*Shigella*
sat / Sat	SPATEs - secreted autotransporter toxin
Sec.	Secondary
sepA / SepA	SPATEs - *Shigella* extracellular protein A
set1 (ShET1)	*Shigella* enterotoxin 1
set2 (ShET2)	*Shigella* enterotoxin 2
sigA / SigA	SPATEs – Protease of *Shigella flexneri* 2A
SNP	Single nucleotide polymorphism
sopE / SopE	T3SS effector protein of *Salmonella*
SPATEs	Protease autotransporter of *Enterobacteriaceae*
spp.	Species
ST	Heat-stable
STa (estla)	Heat-stable enterotoxin a
STb (estlb)	Heat-stable enterotoxin b
stcE / StcE	Secreted protease of C1 esterase inhibitor from EHEC
STEC	Shiga-Toxin-producing *E. coli*
stx1 / Stx1	Shiga-Toxin 1
stx2 / Stx2	Shiga-Toxin 2
StxΦ	Shiga-Toxin-encoding bacteriophage
subAB / SubAB	Subtilase Cytotoxin
Swiss-Prot	Swiss Institute of Bioinformatics Protein database
T	Threshold
T3SS	Type three secretion system
tir / Tir	T3SS - translocated intimin receptor
TrEMBL	Translated European Molecular Biology Laboratory
tRNA	Transfer ribonucleic acid
tsh / Tsh	SPATEs - temperature sensitive hemagglutinin
UniProt	Universal Protein Resource
UPEC	Uropathogenic *E. coli*
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
vat / *Vat*	SPATEs - vacuolating autotransporter toxin
virA / VirA	T3SS effector protein
WGS	Whole genome sequence
X-ray	Roentgen crystallography
yghJ / YghJ	Metalloprotease

1. Introduction

Escherichia coli (*E. coli*) is both, a facultative commensal of the gut microbiota and an important bacterial cause of human as well as animal diseases. The expression of virulence genes plays a key role during infection and pathogenesis. The virulence attributes are frequently encoded on mobile genetic elements, such as plasmids, bacteriophages or chromosomal pathogenicity islands. These encoding genes can be mobilized and, via horizontal gene transfer, recombined in different strains resulting in novel combinations of virulence factors, which can also be integrated into the chromosome. Due to the high genetic plasticity of the *E. coli* species, the ability to gain or lose virulence attributes by horizontal gene transfer allows those organisms to colonize different sites and act as pathogens being responsible for a wide range of infections [1, 67]. Bacterial virulence is influenced by specific factors for adherence, invasion, metabolism and toxins, which promote the pathogen's survival and proliferation into the host [163].

However, recent studies have suggested that the pathogenesis of *E. coli* is considerably more complex than previously assumed. Besides the occurrence of virulence factors [19, 54], toxins can form variants (alleles) which are also able to influence pathogenesis. Funk *et al*. (2013) and Nüesch-Inderbinen *et al*. (2014) analyzed the Subtilase Cytotoxin (*subAB)* genes of Shiga-Toxin-producing *E. coli* (STEC) strains with respect to their presence and genetic location. The phylogenetic analyses resulted in four different *subAB* gene variants. The chromosomally located genes showed more genetic diversity than the plasmid-located genes, indicating a different phylogenetic origin [44, 111]. Similar results in genetic differences were also shown for different Shiga-Toxin (Stx) variants, first described in the 1980s [68]. Karve *et al.* (2014) described the two major isoforms Stx1 and Stx2 including Stx2 variants as significantly different in glycolipid receptor binding preferences and consequently in their resulting toxicity [70]. Furthermore, sequence relationships with regard to different enzymatic functions were also described for the serine protease autotransporter of *Enterobacteriaceae* (SPATEs), a toxin family which consists of already known allelic variations [132]. Those phylogenetic differences could reflect a different pathogenic potential and toxicity of proteins expressed from variants of toxin genes.

However, neither the exact mechanism is known nor have systematic analyses been performed to unravel such functional differences on a larger scale. This thesis aimed to investigate on basis of an initial *in silico* analysis the genetic data of all known toxins released in several *E. coli* and *Shigella* genomes, which are closely related organisms [177], to find allelic variations of each toxin gene. With the aid of these allelic variations conclusions can be drawn about potentially different toxic activities of the expressed proteins in respect of their structure.

2. Literature

2.1 Characteristics of *Escherichia coli* and *Shigella* spp.

Escherichia (*E*.) *coli*, a Gram-negative and facultative anaerobe rod, is a main bacterial species of the normal intestinal microbiota of humans and animals. Besides commensal strains, *E. coli* is also a well-known pathogen. Pathogenic strains are able to cause diverse intestinal and extraintestinal diseases due to the expression of different virulence factors. These virulence factors allow *E. coli* strains to infect eukaryotic cells in different ways inducing a wide range of infections. According to virulence mechanisms that are involved in diseases process, *E. coli* strains can be divided into two different groups (Figure 1) [104, 67].

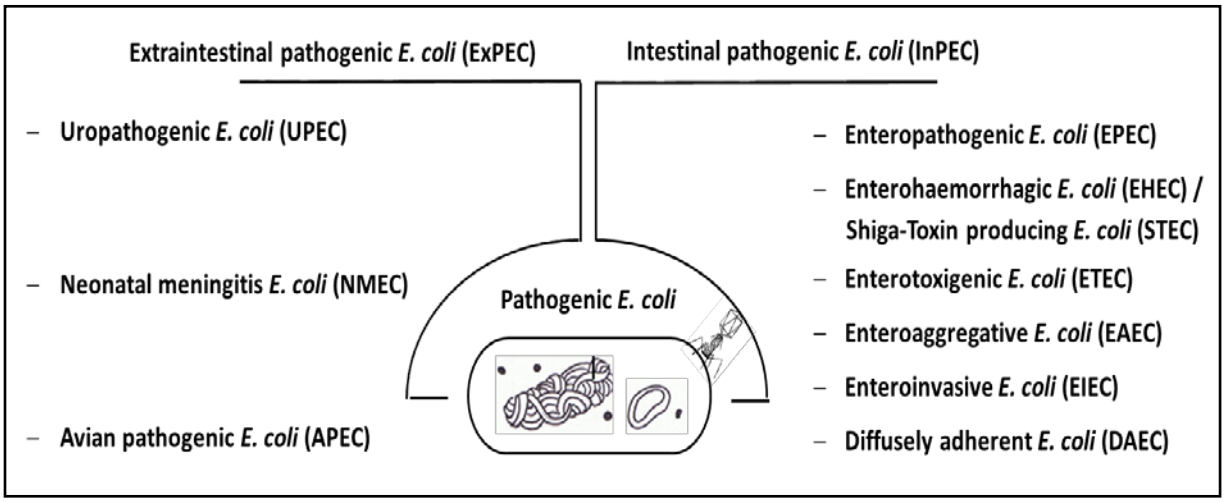

Figure 1: Classification of extraintestinal pathogenic *E. coli* (ExPEC) and intestinal pathogenic *E. coli* (InPEC) [104, 67]

The extraintestinal pathogenic *E. coli* (ExPEC) group is responsible for disease outside the intestinal tract and includes strains causing urinary tract infections (UPEC), neonatal meningitis (NMEC) and avian specific infections (APEC). In contrast to the ExPEC, intestinal pathogenic *E. coli* (InPEC) are associated with diarrheagenic infections and include six well-described pathotypes, like the enteropathogenic *E. coli* (EPEC), enterohaemorrhagic *E. coli* (EHEC), enterotoxigenic *E. coli* (ETEC), enteroaggregative *E. coli* (EAEC), enteroinvasive *E. coli* (EIEC) and diffusely adherent *E. coli* (DAEC). Each of them has different features in its interaction with eukaryotic cells causing infections to the human and animal intestinal tract (Figure 2, p. 3) [104, 67].

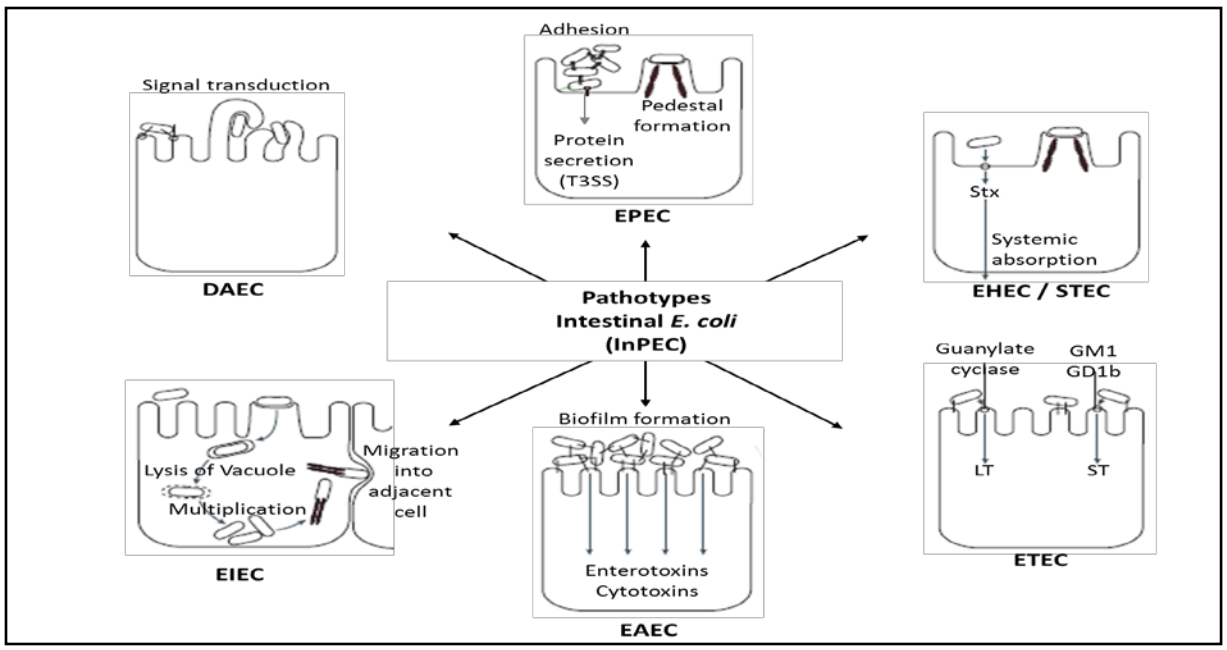

Figure 2: Pathogenic schema of intestinal pathogenic *E. coli* (InPEC) [104, 67] – T3SS: Type three secretion system, LT: Heat-labile enterotoxin, ST: Heat-stable enterotoxin, Stx: Shiga-Toxin

To provide an overview of infections with InPEC, the annual statistics for Germany at the Robert Koch-Institute (RKI) report around ten thousand infections for the year 2014 [128]. After 2014 the RKI reported for Germany even infections with EHEC excluded the hemolytic uremic syndrome (HUS). To give an example of a global prevalence, Shiga-Toxin-producing *E. coli* (STEC - involving the subset of EHEC strains) related illnesses between 1990 and 2012 caused 2.801.000 acute infections and 230 deaths annually, worldwide [91]. This data includes also the well-known EHEC O104:H4 outbreak in Germany 2011 with 4.321 infections and 53 deaths [126, 127]. The annual statistic for Germany for the year 2017 at the RKI report around 597 infections with EHEC excluding HUS [129].

The bacterial genus *Shigella* (*S.*) is the most common cause of acute bloody diarrhea and responsible for a significant proportion of morbidity and mortality associated with diarrheal disease [154, 76, 155]. From 1990 until 2009 in Asia alone around 125 million infections with around 14.000 deaths were associated annually with Shigellosis [8]. Several recent whole-genome-based studies show that *Shigella* species and *E. coli* are closely related organisms. The phylogenetic relationship was constructed from the concatenated alignments of the 2.034 genes in the core genome of *E. coli* and *Shigella* genomes. The results of these studies show, that the four established *Shigella* species (*S. boydii, S. sonnei, S. flexneri* and *S. dysenteriae*) were clustering together with the *E. coli* strains in the phylogenetic trees [177, 176, 123, 143]. Expression experiments of the *Shigella* virulence plasmid in comparison with commensal *E. coli* strains suggests a convergent evolution of the characteristics during horizontal gene transfer (HGT) between *Shigella* species and *E. coli* [120]. These studies suggest that *Shigella* spp. should be considered as a subtype of the species of *E. coli.*

2.2 Transfer of virulence factors

Due to the high genetic plasticity of the *E. coli* species, the ability to gain or lose virulence attributes allows those organisms to colonize different sites and act as pathogens being responsible for a wide range of infections. This horizontal gene transfer (HGT) allows bacteria the uptake of mobile genetic elements from different species and genera to adapt to novel environments and form distinct genotypes or even subspecies. The gene transfer takes place by transformation (uptake of free DNA particles from the environment, extracted by living bacteria or liberated during autolysis), transduction (infection with bacteriophages) or conjugation (DNA transfer of plasmids through cell-cell contact). The virulence attributes are frequently encoded on several mobile genetic elements, such as plasmids, bacteriophages or chromosomal pathogenicity islands (PAIs). These encoding genes can be mobilized and, via HGT, recombined into different strains resulting in novel combinations of virulence factors, which can also be integrated into the chromosomal genome. The most successful combinations of virulence factors constitute the specific pathotypes of *E. coli* (Figure 3). These are capable of inducing specific clinical syndromes in healthy humans and animals after infection [67, 1, 23].

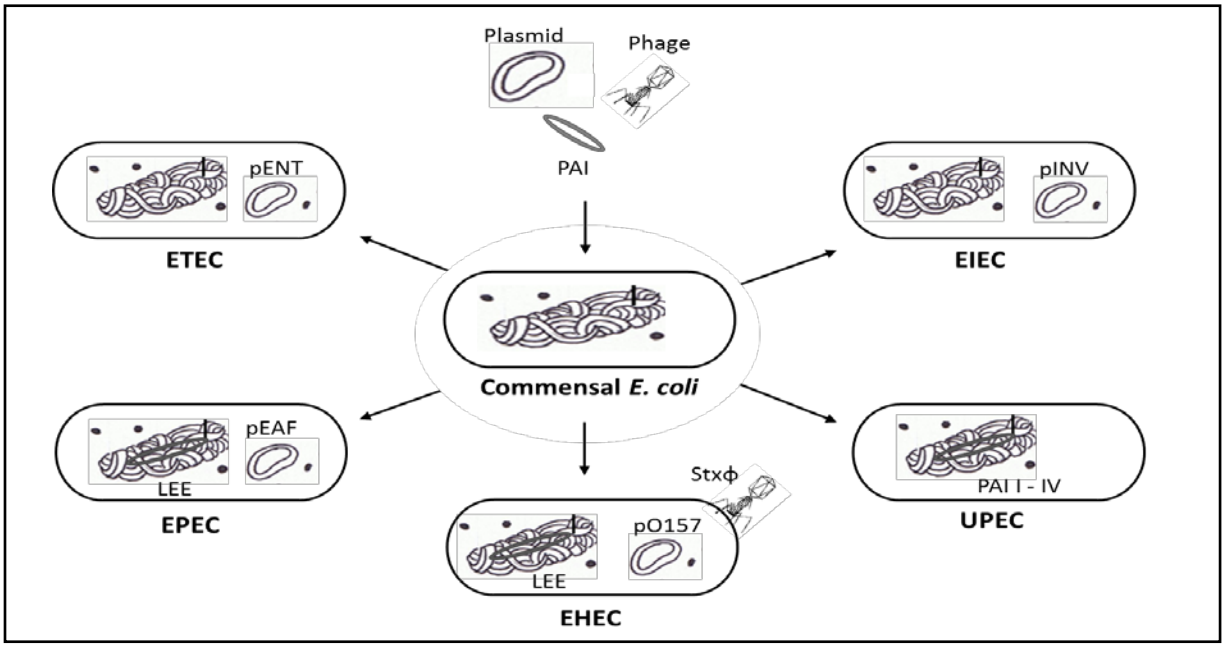

Figure 3: Evolution of pathogenic *E. coli* via mobile genetic elements [67, 1] – PAI: Pathogenicity islands (LEE: Locus of Enterocyte Effacement, PAI I - IV: Pathogenicity island I – IV of UPEC) – Plasmid (pENT: Enterotoxin-encoding plasmid, pEAF: Adhesion-factor plasmid, pO157: EHEC strain O157 plasmid, pINV: Invasion-factor plasmid) - Phage (StxΦ: Shiga-Toxin-encoding bacteriophage)

One main feature of the different pathotypes is the presence of virulence associated plasmids. Plasmids are extrachromosomal DNA elements with the ability to autonomously replicate and can be transferred during conjugation between bacteria. Plasmids are able to carry drug resistance factors, virulence factors and factors to metabolize rare substances. Some of these pathotype-specific plasmids are highly conserved and others are extremely

diverse [27, 65]. For ETEC, the genes that encode the heat-stable (ST) and heat-labile (LT) enterotoxin including colonization factors (CFs) are found on enterotoxin-encoding plasmids (pENT) [24, 46, 160]. These pENTs were also shown to encode antibiotic resistance genes [67]. There are also numerous examples of virulence associated plasmids known, inducing the adhesion-factor plasmid (pEAF) of EPEC [156] and the invasion-factor plasmid (pINV) in EIEC and *Shigella* which encodes effectors like the invasion plasmid antigen (*ipaB*) and the inosital phosphate phosphatase (*ipgD*). These molecules are important for the bacteria to invade host cells [116, 80]. The EAEC plasmid encodes adherence-factors [53] and a set of toxins such as the plasmid encoded toxin (Pet), a serine protease autotransporter of enterobacteriaceae (SPATEs) [32] and a heat-stable enterotoxin (EAST1) [104]. During the infection with EHEC strains the plasmid pO157, which encodes genes for hemolytic activity and adherence, plays an important role and is found in 99 to 100% of clinical O157:H7 isolates from humans [115, 82, 121].

Bacteriophages are viruses that infect bacteria and can genetically modify their host in becoming a part of their genome. The nucleotide sequences of bacteriophages are associated with different G+C contents, oligonucleotide frequencies, tRNA genes (transfer ribonucleic acid) or codon usage from their host's genome. Bacteriophages can constitute as much as 10 - 20% of a bacterium's genome and encode a variety of toxin genes [16, 27]. The production of the Shiga-Toxin (Stx) is the major virulence factor of EHEC / STEC strains, which is encoded on a lambda-like bacteriophage and was one of the steps in the evolution of EHEC from EPEC [124]. Transduction studies using the Stx-encoding bacteriophage f3538(Δstx2::cat) show that Stx can be transmitted via bacteriophages to pathogenic as well as to nonpathogenic *E. coli* lineages and integrated in their genome. This is yet another proof that bacteriophages are able to confer single toxin genes and virulence without any other virulence factors to commensal *E. coli* strains [136].

Coding regions for protein toxins, such as haemolysins, the cytotoxic necrotizing factor (CNF) or SPATEs can be also located on chromosomal PAIs. PAIs are large genomic regions (10-200 kb) harboring one or more genes that are linked to virulence factors and are only present in the genomes of pathogenic strains. The PAI is often flanked by direct repeats, those elements facilitate insertion and deletion of the entire island at a relatively high frequency. Furthermore, PAIs are typically associated with tRNA genes and carry mobility factors, such as integrases, transposases and insertion elements. Bacteriophages use such specific insertion points to integrate into the host genome. Transferred plasmids between bacteria can replicate and, under certain conditions, also integrate on these insertion points into the chromosome. Therefore, PAIs have evolved from bacteriophages and plasmids. By the different G+C contents compared to the core genome, PAIs can be identified [27, 137]. First PAIs were described in a UPEC strain 536, which contains four known islands PAI I-IV.

For example PAI II$_{536}$ inserted at the *leuX* tRNA gene and PAI II$_{J96}$ inserted at the *pheU* gene encode the haemolysin *hlyA*, the cytotoxic necrotizing factor (*cnf1*) and fimbriae [26, 165]. A well- characterized PAI in InPEC is the Locus of Enterocyte Effacement (LEE PAI) from EPEC and EHEC, which encodes for a typ III secretion system, the *efa1* / *lifA* (EHEC factor of adherence) gene and other virulence factors to produce attaching-and-effacing (A/E) lesions. The LEE is inserted at the *selC* tRNA gene in EPEC$_{2348/69}$ and EHEC$_{EDL933}$, which is the same site of the PAI I$_{536}$ of UPEC$_{536}$. The insertion of different PAIs at the same chromosomal region of EPEC / EHEC and UPEC indicates the presence of hot spots in the *E. coli* chromosome, where different PAIs can insert and define different pathotypes [67, 93, 151]. PAIs have also been identified in other pathotypes of *E. coli.* Some PAIs are unique to individual pathotypes as described and allow the bacteria to adapt to specific environments and to cause diverse diseases, whereas other PAIs are present in more than one pathotype [137]. The high pathogenicity island (HPI) originating from pathogenic *Yersinia* spp. was also detected in InPEC as well as ExPEC strains [137, 69, 138]. In EAEC the HPI was also identified as well as the *she* (SHI-1) PAI, which encodes the *Shigella* enterotoxin (ShET1) and the toxin Pet of the SPATEs family, similar to *S. flexneri* [9, 157].

The presence of a broad spectrum of specific virulence attributes is absent in commensal *E. coli* isolates, but the uptake of mobile genetic elements as well as the loss of chromosomal-DNA regions, point mutations or other DNA rearrangements are able to evolve these commensal *E. coli* into a pathogenic *E. coli* lineage. The concerted action of DNA acquisition and gene loss results in a genome-optimization process with respect to certain growth conditions, including host infection and colonization. This gives evidence that horizontal gene transfer plays a major role in the evolution of different bacterial pathotypes [67, 1, 27].

2.3 Toxins as important virulence factors

Bacterial pathogens possess a variety of virulence factors to enter into, replicate within or persist on the host organism. Besides virulence factors for adherence and invasion, intra- and extracellular survival mechanisms, nutrient acquisition, motility, biofilm formation and gene regulation, toxins play an important part as virulence factors, which offer bacterial strategies to interact with mammalian cells [163]. Using diverse mechanisms, bacterial toxins manipulate the host cells' functions to favor conditions for their survival and spread in the host. Therefore, bacterial toxins can be characterized by their specific mechanism of action on eukaryotic cells [33]. Toxins are a product of evolution involving HGT to affect healthy humans and animals in processes of the protein synthesis machinery, actin polymerization, signal transduction pathways, intracellular trafficking of vesicles or immune inflammatory responses [52]. Henkel *et al.*, (2010) described toxins as factors of bacterial pathogens

damaging the host through their own action. Bacterial protein toxins are both, single proteins or oligomeric protein complexes and are organized into distinct domains [59].

Since 2006, the bacterial toxin repertoire consists of around 160 toxins produced by Gram-positive bacteria and 179 toxins expressed by Gram-negative bacteria. Most of the toxins are single-chain polypeptides with a molecular size ranging from 2-3 kDa for *E. coli* to thermostable enterotoxins up to 300 kDa for *Clostridium difficile* toxins A and B. However, many toxins occur as oligomeric multimolecular complexes comprising two or more non-covalently bonded distinct subunits. The *E. coli* heat-labile enterotoxin (LT) forms a heterohexamer, a complex composed of one A-subunit for ADP-ribosylating and a B-subunit with five identical fragments that bind specifically at the surface of intestinal target cells [2]. Stx and SubAB toxins also form such heterohexamers composed of a single A-subunit and five identical fragments in the B-subunits. The B-subunit domain can also form a single fragment such as the B-subunits of the SPATEs. A great number of bacterial toxins consist two functionally different motives. The A part causing the intracellular damage and the B part serving to bind A to appropriate receptors at the cell surface to deliver it into the cytosol. Both motives can be located in two distinct domains of a single-chain polypeptide or in two different proteins in the case of oligomeric toxins. The A and B domains may be linked by disulfide bonds or non-covalent interactions. The A domain of these A-B type toxins encodes catalytic activities such as ADP-ribosyltransferases, adenylcyclases, metalloproteases, RNA N-glycosidases, glycosyltransferases, deaminases, proteases, deoxyribonucleases or phosphatases [2, 59].

Toxins can be classified by their mechanisms of action on target cells (Figure 4, p. 8). The first category are toxins acting ultimately on intracellular targets after crossing the cell membrane (intracellular acting toxins - IATs; Figure 4a, p. 8). To the second category belong all toxins that act strictly at the surface of the cell membrane. This process is divided into two distinct types of toxins with totally different mechanisms. On the one hand the physical damage of the membrane on the target cell (membrane damaging toxins - MDTs; Figure 4b, p. 8) and on the other hand the triggering of intracellular processes during transmembrane signals after binding on appropriate cell receptors (receptor targeted toxins - RTTs; Figure 4c, p. 8) [2]. All currently published toxins expressed by *E. coli* and *Shigella* spp. until June 2016 are classified based on their toxic activity on mammalian cells in Table 3, p. 23 ff.

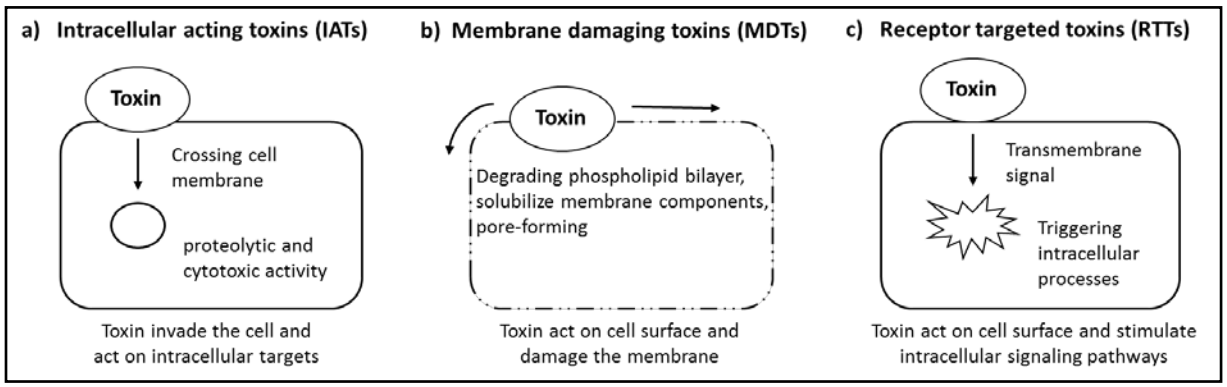

Figure 4: Classification of bacterial toxins by their mechanisms on target cells [own model based on source [2]

The first toxin group that acts on intracellular targets after crossing the cell membrane comprises for example the Stx, SubAB and the SPATEs. These toxins invade the cell and possess proteolytic and cytotoxic functions. The Stx B-subunit binds to Gb3 (Globotriaosylceramide) or Gb4 (Globotetraosylceramide)-receptor within lipid rafts of the target cell membrane that contains cholesterol. The AB_5 complex of the Stx is internalized within an endosome. From the endosome the toxin traffics to the Golgi and moves to the ER where the disulfide bridge keeps the active part of the A-subunit tethered free from the B-subunit. The Stx A-subunit is responsible for inhibition of the protein synthesis of the target cell by cleaving to N-glycosidic bonds of adenine in the ribosomal RNA. This damage induces a stress response in the ER which mediates both pro-inflammatory and pro-apoptotic effects [134, 64, 153]. The SubAB pathway shows likewise an ER stress induced cell death, which is triggered by cleaving the molecular chaperone BiP/GRP78 in the ER, leading to activation of the RNA-dependent protein kinase. SubB has a strong preference for binding to cell surface glycans terminating in the sialic acid N-glycolylneuraminic acid (Neu5Gc) [168]. Trafficking studies of Pet, a SPATEs member, have revealed that Pet is internalized into host cells via clathrin-dependent endocytosis [107]. Inside the cell, Pet moves from the cell surface to endosomes, the Golgi and in the end to the ER. After that Pet moves to the cytosol to reside in close contact with its substrate, an actin-binding protein α-foldrin (spectrin) [106]. Following Pet, which represents the first known SPATE to display enterotoxin activity, many other SPATEs were found to cleave to α-foldrin and trigger similar biological effects [132].

Many members of MDTs have important virulence and pathogenicity factors inducing cell swelling and subsequently cell lysis. Three types of MDTs have been described, based on their ability to damage or disrupt the cell membrane: enzymatically active cytolysins which degrade the phospholipid bilayer, cytolysins which solubilize components by a detergent-like action and pore-forming cytolytic toxins which create channels. The pore-forming toxins (PFTs) are released by the bacteria as monomeric water-soluble proteins and bind to components on the cell surface. On the membrane, monomers can be concentrated and form non-covalently associated oligomers, with amphipathic properties for integrating into the

membrane and forming channels of various sizes depending on the toxin involved [2]. For example the Repeats-in-Toxin (RTX) family including *E. coli* α-hemolysin A (HlyA) possesses a hydrophobic domain that forms cation selective pores in target cell membranes after adsorption on the membrane. The binding to various cells might occur through the recognition of glycosylated membrane components like α-glycophorin, but it is also possible that the calcium-binding domain is responsible for adsorption. The insertion into the cell membrane leads to an irreversible conformational change and produces lesions in the cell membrane. Therefore, HlyA is responsible for a significant calcium influx into cells. The unregulated calcium influx initiates cytoskeletal destruction and cell lysis. The EHEC hemolysin EhxA shows a 61 % identity with HlyA and causes similar effects on target cells, but binds erythrocytes less efficiently and shows no activity against human leukocytes [86]. Cytolysin A (ClyA) shows also a pore-forming activity for erythrocytes in several mammalian species and induces macrophage apoptosis [77, 161].

The properties of the last toxin group that stimulate intracellular signaling pathways exhibit the heat-stable enterotoxins from *E. coli* STa and EAST1. These toxins bind to guanylate cyclase receptors at the surface of intestinal target cells and lead to the activation of the intracellular guanylate cyclase. The activation provokes an elevation of cyclic GMP that stimulates chloride secretion and inhibits NaCl absorption that results in intestinal fluid secretion in the jejunum and ileum. Such a high fluid loss in the jejunum results in a watery diarrhea characteristic for STa of ETEC and EAST1 of EAEC infections [139, 78].

2.4 Genetic variability of toxins

In general toxin families sharing toxic properties and sequence homologies in their active gene region between different bacteria as well as different strains of species or subspecies, which can be a result of horizontal gene transfer. The most common examples of genetic related toxin families are ADP ribosylating toxins (e.g. LT) and other AB toxins (e.g. Stx), pore-forming toxins (e.g. hlyA), enterobacterial autotransporters (e.g. SPATEs) and type III secretion system secreted proteins (e.g. the type III effector proteins espF, espH, virA). The complexity of evolutionary events on the DNA level (e.g. mutations) enables the expression of a variety of toxin variants [27]. Such a polymorphism constitutes the occurrence of several variations of a single gene within a bacterial species. These variants of a specific gene are defined as alleles. An exchange of single nucleotides (single nucleotide polymorphism – SNP) within a gene sequence can lead to a different amino acid during translation, a non-synonymous SNP, and result in a different protein sequence. However, different nucleotide acids in a triplet can also result in the identical amino acid. This is a synonymous SNP, a coding SNP that does not change the protein sequence. Insertions or deletions of single or more nucleotides are also possible and can influence the protein up to the loss of its

functional activity [88,15]. Currently, alleles have been described for toxins of specific pathotypes of *E. coli* among pathogenic *E. coli* and *Shigella* strains [44, 111, 70, 132].

The virulence factor Stx of STEC strains belongs to the AB_5 group of toxins [38, 37]. The active Stx A-subunit is responsible for the inhibition of the protein synthesis of the host cell by cleaving N-glycosidic bonds of ribosomal RNA. The A-subunit is surrounded by a pentamer B-subunit, which binds to receptors on the host cell surface [29, 113; 125]. Stx includes two major isoforms, Stx1 and Stx2, with an amino acid identity of about 60%. The Stx1 isoform can be subtyped into three variants (Stxa, Stxc and Stxd), which show similarities ranging from 95% to 98%. The Stx2 isoform can be subtyped into eighth variants (Stx2a-Stx2h), which show around a 90% amino acid identity to each other. For STEC strains, it is possible that one strain can express one or more Stx variants [6, 70, 135]. Stx subtypes display significant differences in their toxic potential, but the exact reason for toxicity difference is unknown [43]. In previous studies using cell free in-vitro assays, the A-subunit of Stx variants suggested similar enzymatic activities which do not affect the toxicity [56]. In contrast, the B-subunits have shown differences in their receptor binding preferences. Therefore, the B-subunit was hypothesized to influence the cellular toxicity of the Stx variants [70]. Differences in glycolipid binding were linked to host specificity. This is explained by the binding preference to different globotetraosylceramide receptors of cell lines from different hosts. Previous studies show that Stx2a was mostly associated with human diseases and Stx2e was associated with swine diseases [85, 94, 100, 101].

SubAB is also an AB_5 toxin produced by STEC strains and potentially involved in pathogenesis of human disease as an important virulence marker [111, 142]. SubAB has been shown to trigger an endoplasmic reticulum (ER) stress response and stress-induced cell death in a number of animal and human cells by binding to terminal sialic acids of cell membrane receptors [168, 158]. Four allelic variants of SubAB encoding genes have been described. The *subAB*$_{1}$ variant was originally located on the large, conjugative virulence plasmid pO113 [117]. The second *subAB*$_{2\text{-}1}$ variant was located on the pathogenicity island SE-PAI [97, 114] and the third chromosomal variant *subAB*$_{2\text{-}2}$ was encoded on an outer membrane efflux protein (OEP) locus [44]. The fourth allelic variant *subAB*$_{2\text{-}3}$ is associated with a gene predicted to encode a hypothetical protein of yet unknown function [111]. All variants are encoded by the two closely linked genes *sub*A *and sub*B. SubA is the enzymatic active subunit containing the typical catalytic traid. The receptor binding subunit SubB is responsible for the cellular uptake [117]. Phylogenetic analysis of the whole *subAB* gene of all variants show 99% identity to each other. Phylogenetic analysis of the *subA* sequences shows that the plasmid-located are by nearly 100% the most homogeneous ones. A higher genetic diversity of 98% is seen by the phylogenetic analysis of the chromosomal located *subA* genes. Differences between the allele clusters show 91% identity between subA1 to

$subA_{2\text{-}2}$ and 89% between $subA_1$ to $subA_{2\text{-}1}$ [44]. These phylogenetic differences could reflect a different pathogenic potential and toxicity of *subAB*-positive strains similar to different Stx variants [40, 84]. In addition, it is important to reflect whether different variants are associated with several hosts. The alleles $subAB_{2\text{-}1}$ and $subAB_{2\text{-}2}$ are widespread among STEC strains of wild ruminants and sheep. Whereas the variant $subAB_{2\text{-}3}$ could be important in human pathogenesis [111].

The SPATEs constitute a trypsin-like superfamily of virulence factors produced by the type V secretion pathway. They are generally secreted into the external milieu and are highly prevalent among pathogenic *E. coli* and *Shigella* [57, 67, 170]. Such monomeric autotransporters comprise three functionally different domains. The signal peptide targets the protein into the periplasm, the N-terminal passenger domain (α-domain) encodes the protease activity and the pore-forming C-terminal translocator domain (β-domain) targets the protein to the outer membrane (OM) [58, 63, 172]. On basis of phylogenetic analysis SPATEs can be divided into two distinct classes based on the amino acid sequence of the passenger domain [58, 170]. Several studies, discussed in a recent review by Ruiz-Perez and Nataro (2014) and by Andrade *et al.* (2017) suggest a correlation between their phylogenetic clusters and biological functions [4, 132]. Class-1 SPATEs (see Table 1) have a common ability to cause cytopathic effects in cells and display enterotoxin activity, whereas class-2 SPATEs (see Table 1) exhibit a lectin-like activity to degrade a variety of mucins resulting in an advantage for mucosal colonization and immune modulation.

Table 1: Classification of the serine protease autotransporters from Enterobacteriaceae (SPATEs)

Class-1 SPATEs		**Class-2 SPATEs**	
EspC	EPEC secreted protein C	EatA	ETEC autotransporter A
EspP / PssA	Extracellular serine protease (EHEC) / Protease secreted by STEC	EpeA	EHEC plasmid encoded autotransporter
Pet	Plasmid encoded toxin	EspI	*E. coli* secreted protease
Sat	Secreted autotransporter toxin	Hbp / Tsh	Hemoglobin binding protein / Temperature sensitive hemagglutinin
SigA	Protease of *Shigella flexneri* 2A	Pic	Protease involved in intestinal colonization
		SepA	*Shigella* extracellular protein A
		Vat	Vacuolating autotransporter toxin

Identity analysis of the amino acid passenger domains among class-1 and class-2 SPATEs show an identity of 28-34%. Members of the same class share 45-75% amino acid identity in their passenger domains. The high occurrences of allelic variations among each of the SPATE members share similarities from 93 to 97%, suggesting that each SPATE cluster originated from allelic variations of a single parental SPATE [132].

For EAST1 two variants were detected from strain 17-2 (O3:H2) and O-42 (O44:H18). The variant differs from strain 17-2 sequences by one base at codon 21 (ACA → GCA), resulting in a change in the amino acid threonine to alanine. Heterogeneity in the virulence between the strains shows that strain O-42 provoked diarrhea in volunteers, whereas strain 17-2 did not. However, these two strains do not only differ in their EAST1 sequence, but also in the fimbrial antigen they express and in the sequences flanking EAST1. These differences may explain the variation of pathogenicity for those two strains [103, 169].

All these allelic variations could reflect differences in their pathogenic potential. However, at present no systematic analyses of all known toxin genes of *E. coli* and *Shigella* spp. have been performed to find allelic variations with potentially different toxic activities.

2.5 Background of bioinformatics algorithms

For the prediction of properties related to the function of a newly sequenced gene, sequences homologous to a known gene and their resulting proteins or to the whole family of the protein class give first indications. Such a comparison between a set of nucleotide or amino acid sequences in an alignment gives a first hint of functional and evolutionary homologies or differences [108, 3]. The translation of gene sequences into protein sequences follows the genetic code. Proteins are molecules, which are constructs of a total of 20 different amino acids. The order of the amino acids in a protein sequence is the result of the translated nucleotide acids from a gene, three successive nucleotide acids give a triplet and result in one amino acid [109, 145]. Over time, related DNA or amino acid sequences diverge through the accumulation events such as nucleotide or amino acid substitutions, insertions and deletions. A sequence alignment is the first step to determine regions of mutations in a set of homologous sequences [144]. Databases including nucleotide and amino acid sequences like the National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) or the Protein Data Bank archive (PDB) (http://www.rcsb.org) are useful in these analyses for finding homologous sequences and information about the function with regard to the structures of proteins. The basis of a sequence search with the goal to identify homologous sequences is a scoring matrix, which presents optimal scores for a match and negative scores for mismatches [108].

The Basic Local Alignment Search Tool (BLAST) developed by Altschul *et al*., 1990 is the most common local alignment tool which allows the identification of entries in the database that contain similar sequences to a query nucleotide (BLASTn) or protein (BLASTp) sequence. Including a set of algorithms, BLAST enables the finding of a short fragment of a query sequence that aligns with a high score with a fragment of a subject sequence, which is deposited in the database. BLAST uses the local alignment as alignment type using a subset of sequences which were aligned to a subset of other sequences. This local alignment

method is eminently suited to successfully search very large databases and the alignments reveal regions that are highly similar. Therefore, homologous regions between descent related divers sequences can be uncovered [3].

The first step of the BLAST algorithm is to break the query sequence into short segments of a specific length, which are named short words. Theses short word segments were finally used to create an alignment of sequences found in the database. The default length of these words is three amino acids for BLASTp (Figure 5) and eleven nucleotide acids for BLASTn. The successive words are constructed like a sliding window moving over the first letter to select the next three or eleven letters of a word and so on, following the whole query sequence. Subsequently, all words are compared against a sequence in the database. These initial alignments must be greater than a neighborhood score threshold (T), which is derived by using a scoring matrix. For protein alignments, the BLAST algorithm uses as scoring matrix the Blocks Substitution Matrix 62 (BLOSUM 62). This matrix gives positive and negative values for each amino acid identity or substitution between two aligned sequences. In such a matrix all possible pairs of amino acids between the query and the subject sequence receive a value. Therefore mutations of amino acids are included. Using such a matrix, a resulting score (S) for each alignment can be determined [108, 3].

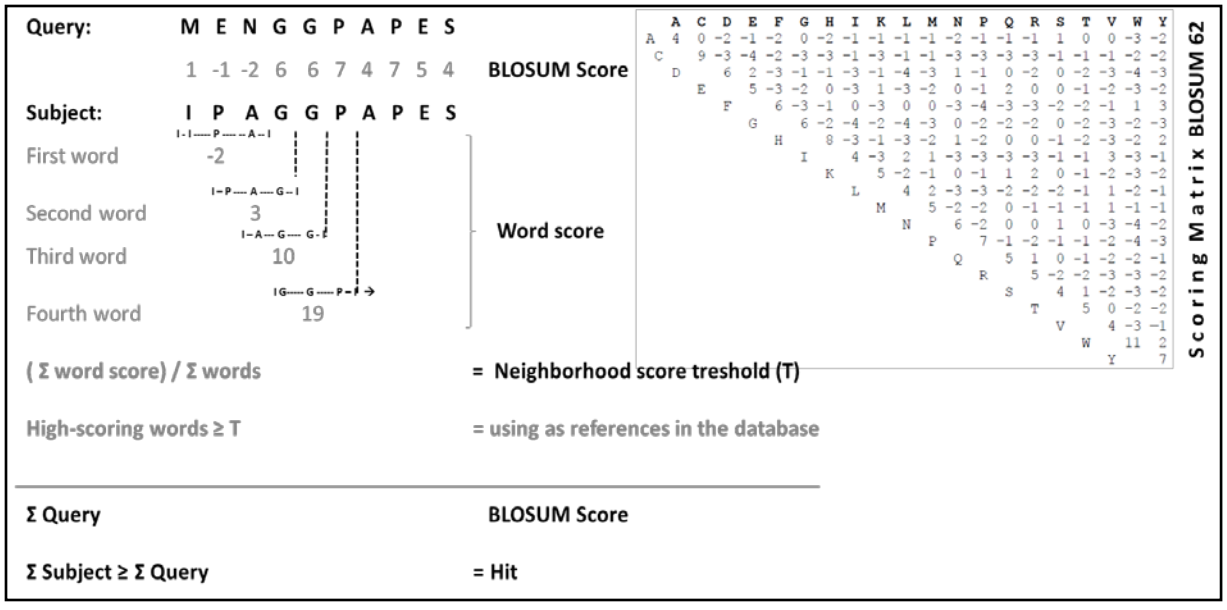

Figure 5: Concept of the BLASTp algorithm [108, 3]

For using a nucleotide query to search for alignments on a nucleotide database a simpler scoring matrix is applied (Figure 6, p. 14). In a nucleotide matrix, each identical match is given the same positive score and all mismatches are given the same negative score [3].

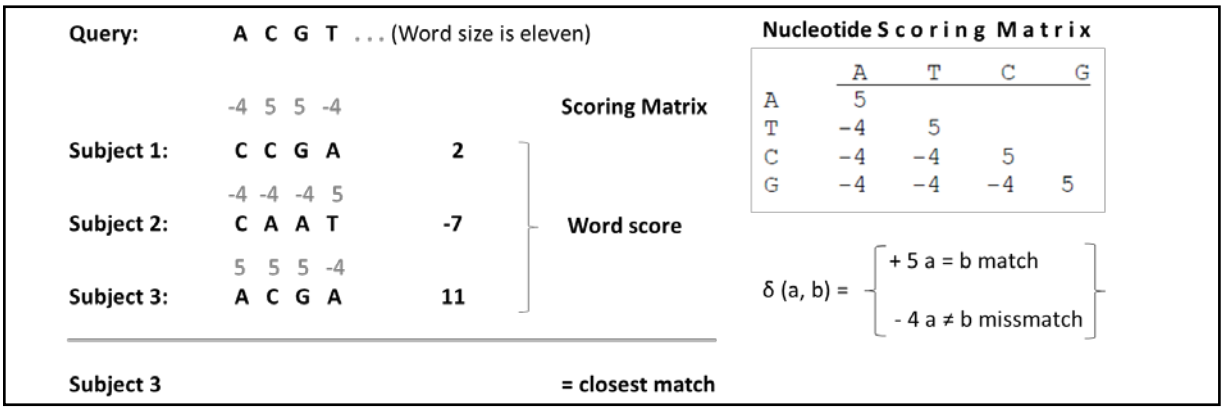

Figure 6: Concept of the BLASTn algorithm [3]

At the end of a BLAST process in NCBI, the program gives a report for each alignment with an Expect (E) Value. The E value reflects the quantity of sequences, which possess an equal sequence length and the database size to result in the same score as the query sequences. Therefore, the lower the E value the more significant is the alignment at a specific probability level and is homologous to the query sequence. The generated report includes also an alignment visualization of the query and the subject sequence. The centerline provides the information if a letter is identical (high similarity), similar but not highly similar (+) and no symbol for non-similarity [105].

Geneious is a popular Java based bioinformatics software platform, which is able to communicate with a number of public molecular biologic databases hosted by NCBI (http://www.ncbi.nlm.nih.gov) and UniProt – Universal Protein Resource (http://www.uniprot.org). UniProt is a comprehensive catalogue of protein data, including protein sequences and functions from Swiss-Prot, TrEMBL and PIR. In Geneious, these databases can be directly screened, the information downloaded and saved in an own database. As a sequence searching tool the BLAST algorithm is used. The results show also a query coverage, E value and identity values for each hit to the query sequence. With the collected sequences of a gene or protein, alignments can be created. These alignments consist of a table with one sequence per row and each column containing homologous residues from the different sequences [14]. To create an alignment in Geneious, the software includes two options, the local and the global alignments (Figure 7, p. 15). The local alignment is appropriate when aligning segments of genomic DNA that may have local regions of similarity embedded in a background of a non-homologous sequence [144]. A global alignment in contrast is an alignment, which assumes homologous sequences along their entire length [108].

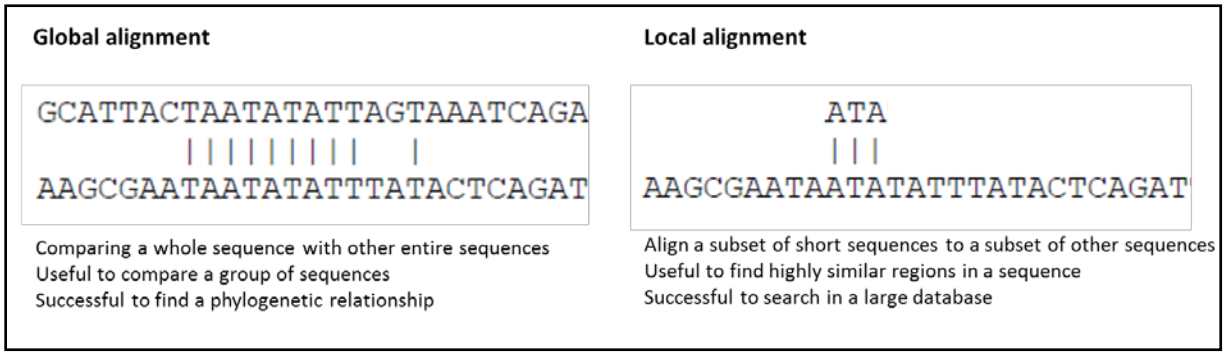

Figure 7: Definition of the global and local alignment [108, 144]

In order to align sequences, a scoring system is required to score matches and mismatches as described for the BLAST algorithm. The scoring system for nucleotide sequences is also a simple match / mismatch weight between the pair of sequences at any given site of comparison. For protein sequences, empirical measurements including the mutation process by comparing a large number of related sequences are necessary. These matrices incorporate the evolutionary preferences for certain substitution over the kind of the query sequence in the form of log-odd scores. In Geneious a number of BLOSUM and PAM (Point accepted mutation) substitution matrices are available to create such a protein alignment [14].

The first step to predict ancestral relationships between the sequences is a multiple sequence alignment, a comparison of multiple related nucleotide or amino acid sequences. Nucleotide sequences in a multiple alignment can be directly translated via Geneious into their protein sequences flanked by their start and stop codons using the genetic code. The multiple sequence alignment is an alignment of more than two biological sequences of similar or different length. The proportion of letters in form of nucleotides or amino acids at each position in such an alignment, are weighted proportionally by using mismatch costs to incorporate differences in letters and gaps. Finally, the multiple alignment can be used as evolutionary distance data to reconstruct phylogenetic trees. The Jukes Cantor, the HKY (Hasegawa, Kishino and Yano) and the Tamura Nei model can be used as genetic distance models to build such a tree. The differences between these models is the equilibrium of the base frequency [14]. The simplest substitution model is the Jukes Cantor model which assumes that all bases have the same equilibrium frequency [66]. The HKY and the Tamura Nei model assume that every base has a different equilibrium base frequency, whereas the Tamura Nei model allows the two types of transitions (A ↔ G and C ↔ T) [55, 149].

Geneious provides two tree building methods, the neighbor-joining method and the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) [14]. The principle of the neighbor-joining method is to find pairs with the minimal total branch length at each stage of clustering. Data of nucleotide or amino acid differences such as substitution matrices are converted into distance data. This distance data is the sum of the branch lengths and the pair that shows

the smallest amount is chosen as a pair of neighbors. Finally, the number of pairs of neighbors in a tree depends on the tree topology. The tree construction starts with a star and splits into an unrooted tree because some pairs are more closely related to each other than other pairs in the tree [133]. In contrast to the neighbor-joining method, UPGMA shall be assumed a molecular clock. This hierarchical clustering method based on the molecular clock hypothesizes a constant rate of evolution. That means a higher mutation rate implying a longer time of evolution in respect of the origin. This method is appropriate when a rooted tree is needed and the rate of mutations determines the distance of the branches in a phylogenetic tree [98].

The secondary structure of protein sequences can be predicted with the Garnier-Osguthorpe-Robson (GOR) method. This method is based on probability parameters derived from empirical studies of known protein structures solved by crystallography. In consequence of this information, propensities of individual amino acids as well as conditional probabilities of amino acids give a neighborhood relationship already formed in the secondary structure used in the formalism of information theory and Bayesian statistics. Features of protein folding such as α-helix, β-strands, turn and coil at each position were predicted based on scoring matrices for each of the four features for all 20 amino acids by following a 17 amino acid window. Therefore the method architecture includes a total of 20x4x17 parameters in one step. The inclusion of homologous sequence information through multiple alignments adds an additional accuracy to the secondary structure prediction. So helices and strands represent the major architectural structures of the conserved core of homologous proteins [49 ,48]. Currently the GOR algorithm combines also evolutionary information with an accuracy of prediction of 73.5% and is therefore the most successful method in this field [141].

To conclude, an alignment is used to find homologous sequences and therefore the first step to determine regions of mutations in these sequences [144]. A common database including genomes, single nucleotide and amino acid sequences is NCBI. For searching homologous sequences, the database uses the BLAST algorithm (http://www.ncbi.nlm.nih.gov). The Geneious software allows the download of genomes or single sequences from the database to compare the public data with internal data. Furthermore, the user can work with a variety of algorithms to search sequences within genomes, to create a sequence alignment and phylogenetic trees [14]. The prediction of properties on the basis of the amino acid sequence is the starting point for understanding possible effects of mutations on the function of a protein [48].

2.6 Basis of protein structure for 3D modeling

The structure of proteins is commonly described by four hierarchical structures of organization (Figure 8), starting with the linear primary structure including the one-letter amino acid code and the linear secondary structure which shows the features of the polypeptide chain (α-helix, β-strand, turn and coil). Hydrogen bonds are the basis of stabilizing these linear secondary structures to form spiral, rodlike α helices, planar β strands and U-shaped turns. The folding of the polypeptide chain into a compact three-dimensional arrangement is represented by the tertiary structure. In contrast to the secondary structure, the tertiary structure is stabilized by hydrophobic interactions and disulfide bonds. For monomeric proteins, which consist of a single polypeptide chain, the tertiary structure is the highest organization level. Multimeric proteins contain two or more polypeptide chains (subunits) and are compared by noncovalent bonds. This organization is comprised in the quaternary structure.

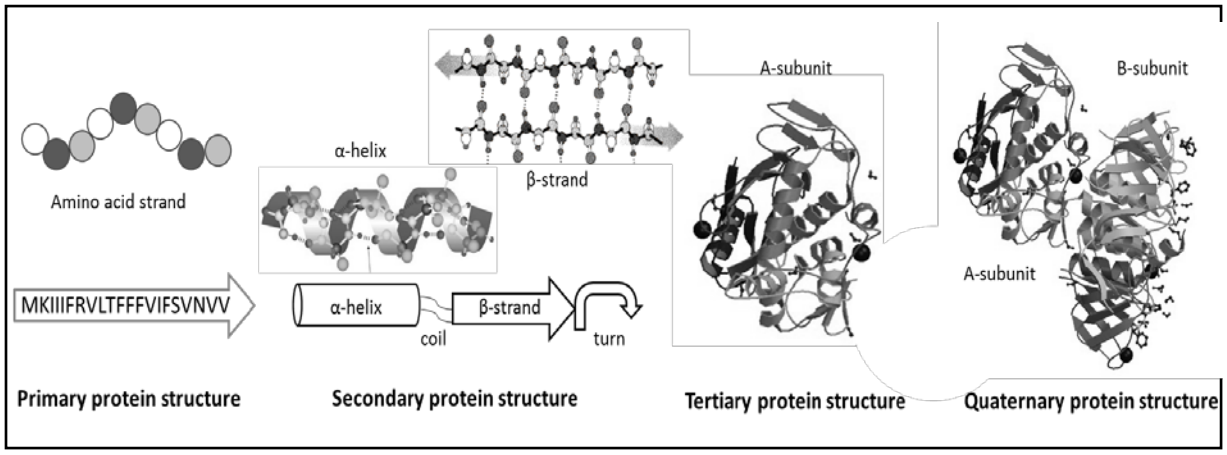

Figure 8: Hierarchical structures of the protein organization [87, 38]

The two regular secondary structures, the α-helix and the β-sheets, are crucial elements of the protein architecture. The α-helix is formed of polypeptides which assume a regular spiral conformation. Each peptide is hydrogen-bonded about its own carbonyl oxygen to the amide hydrogen of the amino acid or residues. This arrangement of bonds forms a polar helix because all hydrogen-bond donors have the same orientation. Every turn of the helix involves 3.6 amino acids. The hydrophobic or hydrophilic properties of the helix are determined by the side chains, as the polar groups of the peptide are involved in hydrogen bonding. The other β-sheet structure consists of laterally packed β-strands. Each β-strand is nearly a full extended polypeptide chain and hydrogen bonded to the others. Like α-helices, β-strands have also a polar definition resulting in the orientation of the peptide bonds. Therefore, the β-sheet arrangement is a very stable and flexible arrangement causing the parallel structures, which can slip over one another. Finally, the α-helix and the β-sheets are linked by turns and coils. Turns are structures, which are stabilized by a hydrogen bond between their end residues and include mainly the amino acids glycine and proline. The lack

of a large side chain of glycine and existing curve in proline allow the polypeptide to fold into a U-shaped turn. In contrast to turns, coils can be formed in many different ways. All these structures were used to pack a protein into a compact arrangement.

Motifs in proteins are regular combinations of these secondary structures and organized into a characteristic three-dimensional structure. Such a motif built for example the combination of one α-helix and two β-strands with an antiparallel orientation to form a zinc ion binding region. This motif is found in most RNA or DNA binding proteins. Presence of the same motifs in different proteins with similar functions indicates that these combinations are very useful in the evolutionary process. Structural and functional regions in proteins are modules of the tertiary structure called domains. A domain is a compactly folded region of various combinations of α-helices, β-sheets, turns and coils. These regions are often physically separated from the other parts of the protein, but connected by the polypeptide chain. Characteristic features are proline and glycine-rich regions, which are conserved in many proteins, arising from homologous structures. Domains define the functional terms based on the activity of a protein and are responsible for its catalytic activity and / or binding ability. Therefore, the domains describe the active center of a protein. Mutations in DNA which encode a protein can change the three-dimensional structure and so the function of a protein. Therefore, the information for folding a protein lies in the nucleotide acid sequence.

Different illustration types of the three-dimensional structure of a protein highlight different information. The simplest way is to trace the course of the backbone atoms with a solid line showing how the polypeptide is packed into the smallest possible volume. The most complex ball-and-stick illustration shows the location of each atom in the protein. The schematic illustration shows how the organization of turns and loops connecting α-helices and β-strands. This type of representation uses shorthand symbols, cylinders for α-helices, arrows for β-strands and a flexible line for all the other parts of the protein. Therefore, the schematic view emphasizes the organization of the secondary structure and various combinations of secondary structures can be easily seen. Information about the protein surface can be provided by the water-accessible model. A water molecule is rolled around the protein structure to clarify lumps, bumps and crevices of the surface. Also regions of positive and negative charges can be painted. Such an illustration shows specific binding interactions to other proteins or ligands [87].

To determine the structure and function of a protein, a number of computational tools of structure prediction using profile-profile matching algorithms have been developed [71]. At present, the three-dimensional PDB archive has grown from experimentally determined protein structures by using X-ray (100.202 entries), NMR (11.145 entries) and electron microscopy (855 entries) as well as 104.417 entries by using the molecular typing for protein

structure prediction [119]. Computational methods originating from simulations of the folding process involving homologous sequences of known three-dimensional structures and using only the protein sequence itself as input data. These template-based homologous models rely on the observation that a number of folds in nature appear to be limited and homologous protein sequences assume similar folding properties. In case that homologous sequences can be found, an alignment can be generated and used directly to build a three-dimensional model [7]. Currently, it is possible to model protein sequences with less than 20% sequence identity to a known protein structure using multiple sequence information. For each segment, known homologous sequences are collected to construct a statistical profile of mutational propensities at each position in the sequence of interest. These short alignments are combined into one large alignment referring to the order of the query sequence [112]. The highest scoring alignments, generated by an E value, are finally used to build a full three-dimensional model of the query sequence. In general, point mutations do not result in a different three-dimensional model, if the point mutation lies in a loop region. Missing or inserted regions caused by insertions or deletions in the alignment are repaired using a loop library. Such a system is typical for a number of freely available structure prediction systems on the web. The user simply pastes their amino-acid sequences into a web page to get a fully downloadable three-dimensional model. Every position along this alignment where a query residue is matched to a template residue shows a score of the profile-profile matching algorithm [71]. Popular web servers for remote homologous / fold recognition are Phyre and I-Tasser (Table 2).

Table 2: Popular web servers for remote homologous / fold recognition [71]

Server	Web address	Confidence	Residues in amino acids
Phyre	http://www.sbg.bio.ic.ac.uk/~phyre2/html	E value	< 1000
I-Tasser	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	E value	10 - 1500
SAM-T06	http://www.soe.ucsc.edu/compbio/SAM_T06/	E value	< 600
PCONS	http://pcons.net/	P value	30 - 800
Robetta	http://robetta.bakerlab.org/	E value	27 - 1000

These homology-based prediction systems use a library of individual structural domains and domain-domain orientations ideally for protein sequences containing multiple domains. The library includes the conserved domain database search service at NCBI or PFAM (Protein families). Therefore, domains are identified and extracted from the query sequence [92, 36]. For optimal prediction, sequences less than 1.000 residues are preferred for use in Phyre. I-Tasser in contrast allows a sequence length up to 1.500 residues and needs therefore more processing time. Additionally, the I-Tasser server possesses the option to build models in the absence of a template, if no consensus to the query sequence can be

found. Such methods divide the query sequence into overlapping fragments of amino acids and the candidate three-dimensional structures were generated using template-based techniques. These structural fragments are then stochastically sampled in respect of empirically derived statistics and assembled to construct a protein conformation [174]. Therefore, the prediction of a three-dimensional protein model is without faults. A combination of different sources, other servers (Table 2, p. 19) or search tools like PDB, are the most reliable way of avoiding false-positive fold recognitions and building the most accurate three-dimensional model [71, 162].

2.7 Influence of protein structure on biochemical properties

Phylogenetic differences in sequence structure of allelic variants have been shown for the toxins stx [70] , subAB [44, 111] and the SPATEs family [132], as described in chapter 2.4 Genetic variability of toxins, p. 9 ff. Those allelic variants could reflect a different pathogenic potential and toxicity on mammalian cells and consequently for the host organism. To analyze changes in enzymatic and toxic activity of the variants, different cell culture or animal models can be used [44, 70, 111]. However, the analysis of biochemical differences between variants of a protein has many possible approaches. First hints for changing biochemical functions reflect the protein structure, which allowed the prediction of active domains. The comparison between active domains of allelic variants reflects possible biochemical changes [48].

One example of changes in biochemical properties is the metal binding preference of zinc-binding motifs on metalloproteases [41]. Metalloproteases represent the largest number of proteolytic enzymes [122]. In general, all metalloproteases contain one or more metal ions, whereby almost all metalloproteases prefer the binding of one or more zinc ions. However, it is also possible that other metal ions like cobalt or manganese can be replaced by zinc. The zinc-binding motif HExxH, which is formed by an α-helix and includes the two histidine (H) residues, is well conserved in many monozinc enzymes at the active site [22]. In the three-dimensional structure of the zinc-binding domain, the two histidine (H) and the glutamic acid (E) residues that cooperate with the zinc ion are engaged in hydrogen bonds with one or two acidic amino acid residues (Glutamatic acid or Asparagine) or other carbonyl oxygen atoms [41]. Through mutational studies, it was proven that the hydrogen bonds between acidic amino acid residues and the zinc ligands (H and G) stabilize the coordination of zinc in the protein [42, 41]. YghJ is a metalloprotease that influences intestinal colonization of ETEC by degrading the major mucins MUC2 and MUC3 in the small intestine by colonizing the cells via oligosaccharides of lipopolysaccharide (LPS) [10, 45, 60, 90]. This metalloprotease was first described in EHEC O157:H7 encoded on the pO157 virulence plasmid, which specifically cleaves the C1 esterase inhibitor (C1-INH) and is named StcE – secreted

protease of C1 esterase inhibitor from EHEC [79]. Genes encoding YghJ are highly conserved in diarrheagenic *E. coli* and exist also in a variety of other enteric pathogens suggesting that this mucin-degrading enzyme may represent a shared feature of these important pathogens to gain nutrients [150]. The active domain of YghJ is the M60-like pfam13402 domain, a canonical HEXXH metalloprotease motif. Biochemical analysis of the domain displayed metal and catalytic glutamate residues dependent proteolytic activity against mammalian mucins. Furthermore, mutations of the predicted catalytic glutamic acid (E) residue of the zinc motif HEXXH to aspartic acid (D) HDXXH dramatically reduce their mucinase activity. The presence of the extended consensus HEXXH suggests that the M60-like domain containing proteins could be considered as glutamic-acid zinc-binding metallopeptidases providing mucin degrading capability [102].

Another example is the glycolipid binding preference of Stx. Stx binds with the B-subunit to the cell surface and this binding is the most important step initiating the toxicity of the Stx holotoxin [164]. Results of a glycolipid receptor interaction study of the different Stx B-subunits of Stx1 and Stx2a-d shows differences in glycolipid binding. The B-subunit of Stx1 displayed a stronger glycolipid binding in contrast to the Stx2 variants, explained by sequence structure differences. Among the Stx2 variants the glycolipid binding affinity is nearly similar because of the fact that the amino acid sequence of the B-subunits are identical except for two to five SNPs. However, the holotoxins of the Stx2 variants show differences in the glycolipid binding preference. This suggests that the A-subunit might take part in receptor recognition, while the C-termini of the A-subunits exhibit genetic differences. The A-subunit of Stx2a contains a basic lysine as a significant difference to Stx2d with an acidic glutamate at the C-terminus [70].

Sequence relationships of SPATE proteins with respect to the amino acid sequence of the passenger domain reveal two distinct functional classes. The functional differences between the cytotoxic and the lecithin-like immunomodulatory SPATEs are caused by differences in a helix-turn-helix motif. The facing motif includes the catalytic triad of the protease. The functional role of the helix-turn-helix motif is still unclear. A possible function might be the accessibility of substrates to the catalytic motif, where substrates are incorporated and cleaved. A fully helix-turn-helix motif harboring two contiguous cysteines with a potential to form a disulfide bond to the catalytic domain is more distinctive of the cytotoxic class. In contrast to the lecithin-like immunomodulatory SPATEs, the helix-turn-helix motif is incomplete and the cysteines are absent [72, 83]. In a mutagenesis study on the cytopathic serine protease Pet, the helix-turn-helix motif was damaged causing proteolytic but no longer cytopathic effects [30].

These examples show that the transcription and translation of allelic variants of toxin genes lead to the expression of proteins with a different tertiary structure. Therefore, these protein variants show changes in their active domains due to changes their catalytic activity, receptor binding capacity or substrate activity. However, systematic analysis to unravel functional differences of allelic variations of all known toxins released in several *E. coli* and *Shigella* genomes have not yet been performed. Therefore, research of all currently published toxins expressed by *E. coli* and *Shigella* is required to identify a subset of allelic variants for each toxin in those genomes. The aim of this thesis is to analyze on basis of an *in silico* analysis the genetic data of all these allelic variants and their resulting structure with regard to their possible functional changes.

3. Material and methods

3.1 Material

To test the hypothesis that allelic variations of toxins lead to functional changes, in this thesis a set of currently all known toxin genes of *E. coli* and *Sigella* spp. until June 2016 was compiled. With these toxin genes we initially analyzed a large set of 423 whole genome sequences of our internal *E. coli* Reference Collection "ECore" [140] and 69 STEC strains [31]. Subsequently, to identify all allelic variants of a toxin gene, we screened the open web database of the National Center for Biotechnology Information (NCBI) for every reference toxin sequence. The resulting genomes as well as plasmids, PAIs and single toxin genes of NCBI were also collected in a data set. By using such a large set we were able to identify all allelic variants of currently known toxins in *E. coli* and *Shigella* spp.

3.1.1 Toxin reference sequences

The resulting literature research of currently all published toxins expressed by *E. coli* and *Shigella* spp. to June 2016, revealed a total set of 39 toxins genes, which are assembled and classified based on their toxic activity on mammalian cells in Table 3. These nucleotide sequences were used as reference sequences to identify allelic variations that are present in *E. coli* and *Shigella* genomes.

Table 3: Published toxin genes in the genome of *E. coli* and *Shigella*

Intracellular acting toxins (IATs)					
Gene designation		**Acc. No.**	**Activity / Effect**	**Patho-type**	**Reference**
eatA	SPATEs - ETEC autotransporter A	AY163491	proteolytic activity, cleaves oligopeptide methoxysuccinyl	ETEC	[Dautin N; Open Access Tox 2010], [Parel SK *et al*; Infect Immun 2004]
epeA	SPATEs - EHEC plasmid encoded autotransporter	AY258503	proteolytic and mucinase activity	EHEC	[Leyton DL *et al*; Infect Immun 2003]
espC	SPATEs - EPEC secreted protein C	AF297061	proteolytic activity, iron secretion	EPEC, STEC	[Kaper JB *et al*; Nat Rev 2004], [Mellies JL; Infect Immun 2001]
espI	SPATEs - *E. coli* secreted protease	AJ278144	proteolytic activity	EPEC, STEC	[Dautin N; Open Access Tox 2010]
espP / pssA	SPATEs - extracellular serine protease (EHEC) / protease secreted by STEC	AF074613 / BA000007	cleaves coagulation factor V, proteolytic and cytotoxic activity	EHEC, STEC	[Kaper JB *et al*; Nat Rev 2004], [Djafari S *et al*; Mol Microbiol 1997]
hbp / tsh	SPATEs - hemoglobin binding protein / temperature sensitive hemagglutinin	AJ223631 / AY545598	hemoglobin interaction, degradation and heme binding	Various (APEC)	[Otto BR *et al*; J Exp Med 1998], [Dautin N; Open Access Tox 2010], [Dozois CM *et al*; Infect Immun 2000]
pet	SPATEs - plasmid encoded toxin	AF056581	proteolytic activity, ion secretion, cytotoxicity	EAEC	[Kaper JB *et al*; Nat Rev 2004]
pic	SPATEs - protease involved in intestinal colonization	CU928145	proteolytic, mucinase and hemagglutinin activity	UPEC, EAEC, EHEC	[Kaper JB *et al*; Nat Rev 2004], [Dautin N; Open Access Tox 2010]

Intracellular acting toxins (IATs)

Gene designation		Acc. No.	Activity / Effect	Patho-type	Reference
sat	SPATEs - secreted autotransporter toxin	AF289092	vacuolation cytotoxin	UPEC	[Kaper JB *et al*; Nat Rev 2004], [Guyer DM *et al*; Infect Immun 2002]
sepA	SPATEs - *Shigella* extracellular protein A	HE610901	proteolytic activity, mucosal atrophy	EIEC	[Kaper JB *et al*; Nat Rev 2004], [Benjelloum-Touimi Z *et al*, Mol Microbiol 1995]
sigA	SPATEs – Protease of *Shigella flexneri* 2A	AY258503	ion secretion, cytotoxicity	EIEC	[Kaper JB *et al*; Nat Rev 2004], [Al-Hasani A *et al*, Infect Immun 2000]
stx1A-B	Shiga-Toxin	AE005674	depurinates rRNA, inhibiting protein synthesis, induce apoptosis	EHEC	[Kaper JB *et al*; Nat Rev 2004], [Karve SS and Weiss AA; PLOSone 2014]
stx2A-B	Shiga-Toxin	AE005174	depurinates rRNA, inhibiting protein synthesis, induce apoptosis	EHEC	[Kaper JB *et al*; Nat Rev 2004], [Karve SS and Weiss AA; PLOSone 2014]
subA-B	Subtilase Cytotoxin	AY258503	proteolytic and cytotoxic activity (appoptosis)	STEC	[Funk J *et al*; BMC Microbiol 2013], [Paton WA and Paton JC; J Tox Rev 2010]
vat	SPATEs - vacuolating autotransporter toxin	AY151282	proteolytic activity	ExPEC	[Dautin N; Open Access Tox 2010]

Membrane-damaging toxins (MDTs)

Gene designation		Acc. No.	Activity / Effect	Patho-type	Reference
clyA / hlyE	Cytolysin A	AY576656	pore-forming, cell lysis	Various	[Ludwig A *et al*; J Bacteriol 2004], [Kaper JB *et al*; Nat Rev 2004]
ehxA	EHEC hemolysin	AF043471	cell lysis	EHEC	[Kaper JB *et al*; Nat Rev 2004]
hlyA	Hemolysin	AP010959	cell permeabilization and lysis	UPEC	[Kaper JB *et al*; Nat Rev 2004]

Receptor targeted toxins (RTTs)

Gene designation		Acc. No.	Activity / Effect	Patho-type	Reference
astA (EAST)	Heat-stable enterotoxin	ECOASTAA	activates guanylate cyclase, resulting in ion secretion	Various	[Kaper JB *et al*; Nat Rev 2004], [de Sousa CP and Durbreuil D; IJMM 2000]
cdtVa-c	Cytoletal distending toxin	AJ508930	DNase activity, blocks mitosis in G2/M phase	Various	[Kaper JB *et al*; Nat Rev 2004], [Janka A *et al*; Infect Immun 2003]
cif	T3SS - Cycle-inhibiting factor	AF497476	blocks mitosis in G2/M phase, induces formation of stress fibres	EPEC, EHEC	[Marchès O *et al*; Mol Microbiol 2003]
clbA-Q	Colibactin locus	AM229678	induces DNA double-strand breaks, cell cycle arrest in G2 phase, megalocytosis	ExPEC	[Putze J *et al*; Infect Immun 2009], [Martin P *et al*; PLOSone 2013]
cnf1 and *cnf2*	Cytotoxin necrotizing factor	X70679 ECOCNF2	permanently activates the regulation of GTPases, promotes new cell activities (transcription, proliferation, survival)	MNEC, UPEC, NTEC	[Boquet P; ELSEVIR Toxicon 2001], [Kaper JB *et al;* Nat Rev 2004]
espF	T3SS effector Protein	FM180568 / AE005174	opens tight functions, induces apoptosis, induce mitochondrial pathway	EPEC, EHEC	[Kaper JB *et al*; Nat Rev 2004], [Hohlmes A *et al*; Infect Immun 2010], [Marchès O *et al*, J Bacteriol 2006]
espH	T3SS effector Protein	FM180568	modulates filopodia and pedestal formation	EPEC, EHEC	[Kaper JB *et al*; Nat Rev 2004], [Tu X *et al*; Mol Microbiol 2003]

Receptor targeted toxins (RTTs)					
Gene designation		**Acc. No.**	**Activity / Effect**	**Patho-type**	**Reference**
ipaB	T3SS - invasion plasmid antigen	AY098990	apoptosis, membrane insertion, lysis of phagocytic vacuole	EIEC	[[Kaper JB et al; Nat Rev 2004], [Gibotti *et al;* J Microbiol 2004]
ipgD	T3SS - Inositol phosphate phosphatase	AF348706	membrane blebbing, increased chloride secretion	EIEC	[Kaper JB *et al*; Nat Rev 2004]
leoA	Dynamin-like protein	FN649414	membrane vesicle associated toxin secretion, GTPase virulence factor	ETEC	[Michie KA *et al*; PLOSone 2014]
lifA / efa1	Lymphocyte inhibitory factor / EHEC factor for adherence	AJ133705 / AF159462	inhibits lymphocyte activation, adhesion, blocks interleucin production	EPEC, EHEC	[Kaper JB *et al*; Nat Rev 2004]
LT (elt)	Heat-labile enterotoxin	CP000795	ADP ribosylates and activates adenylate cyclase resulting in ion secretion	ETEC	[Kaper JB *et al*; Nat Rev 2004]
map	T3SS - mitochondrion-associated protein	FM180568	disrupts mitochondrial membrane potential	EPEC, EHEC	[Kaper JB *et al*; Nat Rev 2004]
set1 (ShET1)	*Shigella* enterotoxin 1	AF348706	ion secretion	EAEC, EIEC	[Kaper JB *et al*; Nat Rev 2004]
set2 (ShET2)	*Shigella* enterotoxin 2	AF348706	ion secretion	EIEC, ETEC	[Kaper JB *et al*; Nat Rev 2004]
STa (estla)	Heat-stable enterotoxin a	FN649417	activates guanylate cyclase resulting in ion secretion	ETEC	[Kaper JB *et al*; Nat Rev 2004]
STb (estlb)	Heat-stable enterotoxin b	FN649418	increases intracellular calcium resulting in ion secretion	ETEC	[Kaper JB *et al*; Nat Rev 2004]
tir	T3SS - translocated intimin receptor	AF070067	inhibits NF-κB activation, nucleation of cytoskeletal proteins, loss of microvilli, GAP-like activity	EPEC, EHEC	[Kaper JB *et al*; Nat Rev 2004], [Ruchaud-Sparangano MH *et al*; PLOSpathogenes 2011]
virA	T3SS effector protein	AF386526	cleaves a-tubulin (intracellular spreading, microtubuli destabilization, membrane ruffing)	EIEC	[Kaper JB *et al*; Nat Rev 2004], [Davis J *et al;* Prot Science 2008]
yghJ	Metalloprotease	AP009048	degrades intestinal mucins (homolog SsIE)	ETEC	[Luo Q *et al;* Infect Immun 2013], [Nesta B *et al;* PLOSpathogens 2013]

3.1.2 Internal database ECore – *E. coli* Reference Collection

The internal database ECore used in this thesis, consists of 423 whole genome sequences of the *E. coli* Reference Collection "ECore" published by Semmler 2018 [140] and contains additional 69 STEC whole genome sequences published by Eichhorn 2015 [31]. This internal ECore database includes 242 internal isolates of the Institute of Microbiology and Epizootics, 193 isolates from other institutes as well as 57 reference genomes collected from NCBI. Within the dataset the prevalent species is *E. coli*, which comprises 471 genomes followed by *Shigella* with twelve genomes. Other species like *E. albertii* (n=5), *E. alvei* (n=3) and *E. fergusonii* (n=1) are also present in the database. The majority of

pathotypes of *E. coli* are ETEC (n=87), STEC (n=69), atypical EPEC (n=36) and EPEC strains (n=27) as well as many strains with unidentified pathotype (n=212). The database involves 248 human isolates and isolates of farm, zoo and companion animals. In addition to the human isolates, the largest group of the animal isolates are those of cattle (n=54) and dogs (n=34). Not all of the isolates are associated with diseases. There are also isolates from healthy humans, healthy animals and environmental isolates. The isolates which were collected in a time range between 1943 to 2011 originate from regions in Europe, Asia, Australia, Africa, North- and South America. Within the appendix, the whole database is listed in Table A 1 (pp. i ff.), including information on every strain with focus on the strain number, species, pathotype, host, possible disease of the host, the source laboratory, year of isolation, region where the strain was isolated and synonymous names of the strain.

3.1.3 NCBI database - National Center for Biotechnology Information

The screening of toxin genes in the open web database NCBI using nucleotide sequences of the reference toxins resulted in a second collection including 86 whole genome sequences, 72 plasmids, 22 PAIs and 276 genes which were not considered within the internal ECore dataset. Within the second dataset the prevalent bacterial species is *E. coli* with 396 entries (including genomes, plasmids, PAIs and genes) followed by *Shigella* (*S. boydii, S. dysenteriae, S. flexneri* and *S. sonnei*) with 56 entries (including genomes, plasmids, PAIs and genes). The prevalent pathotypes of *E. coli* described on NCBI are STEC (n=89), EHEC (n=47), UPEC (n=45) and ETEC (n=44) strains. Equal to the internal ECore database, many strains had unidentified pathotypes (n=142). Also similar to the internal ECore database, the largest group of host entries comprises 263 human isolates. The database involves also data of farm animals (n=79) and food samples (n=16). Around 198 entries are associated with diseases like diarrhea (n=66), urinary tract infection (n=45) and HUS (n=31), but there are also entries listing healthy humans and animals (n=40) as isolation sources. The collected data were published between the years 1917 and 2016. The strains have been isolated in several geographical regions from Europe, Asia, Australia, Africa, to North- and South America. The whole database is listed in Table A 2 (pp. xvi ff.) and attached in the appendix, including information about the accession number in NCBI, species, pathotype, host, disease of the host, the source laboratory, year of isolation, region of isolation and synonymous names of every genome, plasmid, PAI and gene.

3.2 Methods

The nucleotide sequences of the reference toxin genes were used to identify allelic variations of each toxin gene in the NCBI and the internal ECore database. For database searches the BLASTn algorithm was used. An editing of all extracted toxins and their allelic variations were carried out using Geneious version 7.1.2., which provides algorithms for alignment construction, phylogenetic tree creation and secondary structure prediction. The three-dimensional structure of selected allelic variations was modeled using the I-Tasser server for protein structure and function prediction (http://zhanglab.ccmb.med.umich.edu/I-TASSER/), which refer to the protein database UniProt (http://www.uniprot.org) and the molecular visualization system PyMol (https://www.pymol.org).

3.2.1 Nucleotide BLAST (BLASTn)

The NCBI BLASTn algorithm works on the basis of the basic local alignment search tool of Altschul *et* al., 1990 as described in chapter 2.5. The nucleotide level is the starting point in this thesis as the raw genomic data, which forms the basis of the following bioinformatics analysis and interpretations. Therefore, an important part was the screening of the NCBI database for all reference toxin sequences (Table 3, p. 23 ff.) to find all currently published variants of each toxin gene. The BLASTn program allows the matching of query nucleotide sequences against those present sequences recorded in the NCBI nucleotide database.

Each of the toxin reference sequences was uploaded as FASTA file and for sequence search the BLASTn algorithm was used, which allowed the identification of any related sequences. The algorithm parameter settings for sequence search, including general parameters for the search stringency, scoring parameters to choose the scoring matrix and gape penalties as well as filter options, were set as default parameters of BLASTn. The description table shows a summary of the sequences that were identified to be somewhat similar to the input query and provide information about the similarity for every match. This information includes the highest alignment score (Max score), the total alignment score (Total score), the percentage of the query coverage (Query cover), the best Expect value (E value) and the highest percent identity (Ident) of all query-subject alignments in respect of all alignments from that resulting search. Individual sequences can be selected as pairwise query-subject alignment to give a detailed view of SNPs between both sequences (http://blast.ncbi.nlm.nih.gov).

Based on a range going from 85% to 100% identity to the reference sequence, we assumed that these alignments are variants and results for the gene of interest. Alignments with an identity lower than the threshold of 85% cannot be uniquely assigned as gene variants and were excluded from this thesis (see chapter 5.1, p. 59 ff.).

The identified *E. coli* and *Shigella* genomes, plasmids, PAIs and single toxin gene sequences of the database screening were downloaded as FASTA files and collected within the Geneious software to create an individual database as described in chapter 3.1.3, p. 26. Subsequently, the whole genome sequences of the *E. coli* and *Shigella* strains of the internal ECore database (chapter 3.1.2, p. 25 ff.) were also screened for all reference toxin genes using BLASTn within an in-house pipeline (Figure 9). The genomes and their extracted toxin genes were also downloaded as FASTA files and collected with the Geneious software.

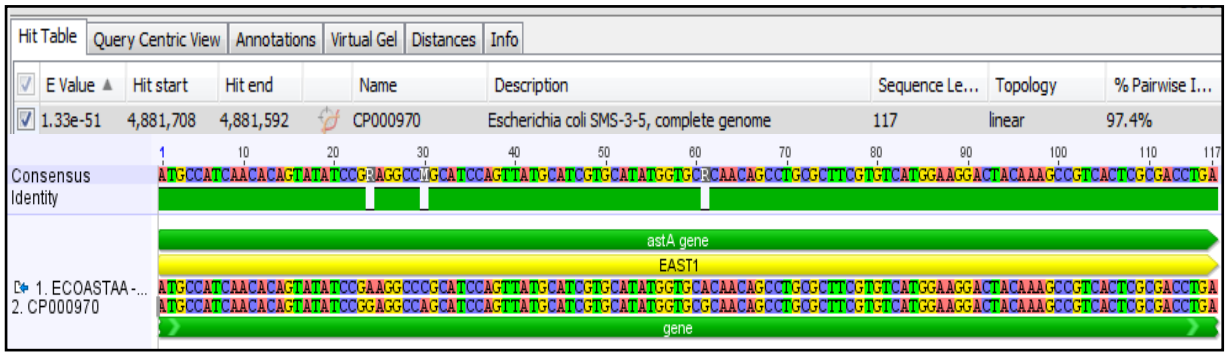

Figure 9: Detailed pairwise alignment of the query heat stable enterotoxin *astA* / EAST sequence (ECOASTAA) and a similar sequence of the *E. coli* genome CP00970 after sequence search in Geneious. The sequence alignment demonstrates the same results as the NCBI screening.

Only toxin genes, which were detected in their full sequence length, are included in this thesis. It might be possible that gene fragments are a result of the assembled sequence that is composed of short overlapping reads (DNA fragments). During sequencing, single genes can be split into fragments that appear in different contigs (parts representing a whole genome). We searched the absent fragments in the whole genome sequence via BLASTn to identify if the gene splits into fragments and assembles all fragments into a whole gene sequence. Toxin genes, which have not been detected in their full sequence length, were not included into this thesis assuming noncoding gene fragment sequences.

3.2.2 Algorithm for alignment construction

The starting point to compare all collected sequences of one toxin gene was the creation of a multiple sequence alignment in order to identify SNPs and finally allelic variations. Multiple sequence alignments in Geneious were done using the progressive pairwise alignment method. This method is the most efficient algorithm when a variety of large sequences needs to be aligned.

To create a pairwise alignment of gene sequences in Geneious, the global alignment was used (as described in chapter 2.5), because whole sequences were compared to find polymorphisms [14, 108]. As scoring system, a simple match / mismatch (5.0 / -4.0) cost matrix between nucleotide acid pairs of the sequences at any given site of comparison was needed to find the optimal alignment of the two sequences (as described in chapter 2.5, p. 12) [14, 3]. The chosen scoring parameters to create a pairwise alignment were the default

parameters of the “Geneious Alignment” (Alignment type: Global alignment with free end gaps, Cost Matrix: 65% similarity (5.0/-4.0)).

To compute a multiple alignment, pairwise alignments for all existing sequences against each other were performed to identify the most related sequences which were aligned. Subsequently, these consensus sequences were aligned with the remaining sequences depending on their similarity to each other. This process is completed when all existing sequences have been combined in one multiple alignment, as exemplary shown in Figure 10 [14, 35, 133].

Figure 10: Nucleotide alignment of four selected *astA* / EAST1 sequences representing SNPs among each other

All nucleotide sequences in a multiple alignment were directly translated into their protein sequences involving their start and stop codons based on the genetic code referring the correct reading frame. Such a multiple alignment of amino acid sequences represents the proportion of the aligned sequences at each position using the scoring matrix BLOSUM 62 to incorporate differences in letters and gaps, as described in chapter 2.5, p. 12 [14]. The chosen parameters to translate a nucleotide alignment are the default parameters in Geneious “Translation Alignment” (Genetic code: Standard, Translation frame: 1, Alignment type: Global alignment with free end gaps, Cost Matrix: Blosum62). The resulting protein alignment is shown in Figure 11.

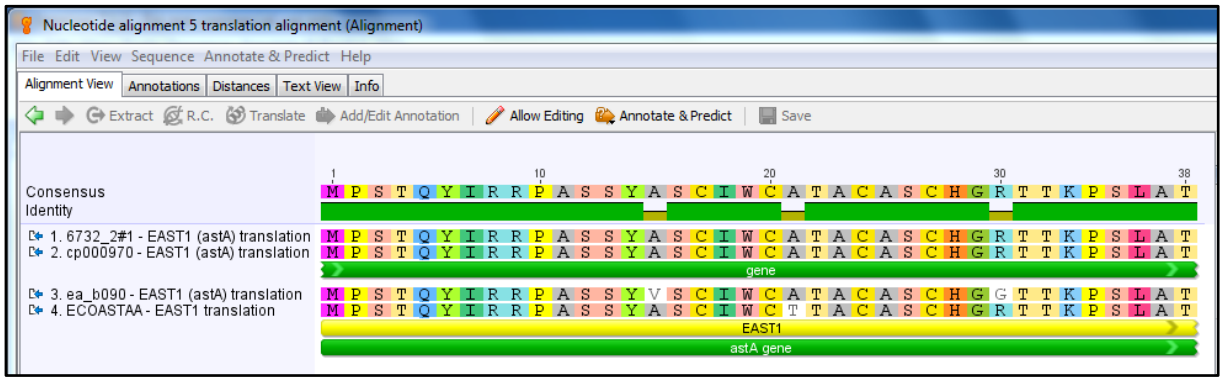

Figure 11: Protein alignment of four selected astA / EAST1 sequences representing changes in their amino acids

In both alignments, a detailed alignment view of the nucleotide or amino acid letters was shown to identify differences in their sequence structure. Another helpful feature was the view of the percentage distance between the sequences, which are summarized in a comparative table. To compare alignments, SNPs among the nucleotide sequences have been defined as synonymous or non-synonymous SNPs with regard to their resulting protein sequences.

3.2.3 Algorithm to create a phylogenetic tree

With the inbuilt tree reconstruction algorithms in Geneious a phylogenetic tree can be directly formed from a multiple sequence alignment. Therefore, multiple alignments can be used as distance data to identify phylogenetic relationships. To build a phylogenetic tree using the Geneious tree builder, different options for the genetic distance model as well as the tree build method were available [14].

As genetic distance model the “Tamura Nei” model was used to build a tree of nucleotide sequences. In contrast to the other models (“Jukes Cantor” and “HKY”), the Tamura Nei model assumes every base has a different equilibrium base frequency. Transitions evolve at a different rate of transversions and allow the two transitions types A ↔ G and C ↔ T [14, 149, 66, 55]. To build a phylogenetic tree of amino acid sequences, Geneious provides the “Jukes Cantor” genetic distance model as a single option. The model assumes that all amino acids have the same equilibrium base frequency of 5% and replacements occur at equal rates [14, 66].

As tree build method, for both alignment types the UPGMA was used. In contrast to the Neighbor-joining method, the UPGMA is based on the assumption of a molecular clock and forms a rooted tree [14, 133, 98]. The method allows an alignment of a variety of sequences and a sequence clustering based on their polymorphisms [98]. Therefore, this approach allows a simplified determination of allelic variations from a multiple gene or protein alignment.

3.2.4 Prediction of secondary structure

The secondary structure of an amino acid sequence can be directly predicted in Geneious using an implementation of the Garnier-Osguthorpe-Robson (GOR) algorithm accessing a server based on the database of Cuff and Barton (1999, 2000) including 513 sequential protein domains, which contain 84.107 residues [14, 141, 20, 21]. The GOR algorithm calculates the α-helices, β-strands, turns and coils probabilities at each residue position and provides an initial structural prediction, which results in the highest probability. The four features of protein folding were predicted based on scoring matrices for each feature for all 20 amino acids by following a 17 amino acid window. Therefore, the method architecture includes a total of 20x4x17 parameters in one step. Regarding heuristic rules,

helices shorter than five residues and strands shorter than two residues were converted into coils. Incorporated multiple sequence alignments increase the evolutionary information content for improved discrimination among secondary structures. Currently the prediction accuracy of the GOR algorithm reached of 73.5% [49 ,48, 141, 75]. Therefore, Geneious enables a first statement about resulting differences in the secondary protein structure of allelic variations of a toxin.

Mutation matrices for transmembrane, extracellular and intracellular proteins have been useful for the interpretation, especially in the absence of known three-dimensional structures. They gave a hint of substitution effects at particular sites in the sequence. The different matrices are published on the website: http://www.russell.embl-heidelberg.de/aas. These matrix scores numerically classify an amino-acid substitution: (i) unpreferred mutations have given negative scores, (ii) preferred substitutions have given positive scores and (iii) neutral substitutions have given zero scores, such as the BLOSUM 62 substitution matrix. Amino acids can share also properties like their size or their polarity and sometimes mutations in the protein structure result in the same secondary structure [11]. An example to predict the secondary structure in Geneious and the outcome that mutations have no influence on the resulting secondary structure is shown in Figure 12.

Figure 12: Prediction of secondary structure of the four selected *astA* / EAST1 sequences

3.2.5 Three-dimensional structure prediction and visualization

If allelic variations have shown differences in their gene and protein sequence as well as in their secondary structure and especially in active domains, the three-dimensional (3D) protein structure was predicted by using the modeling software I-Tasser to give a hint of resulting differences in their functional activity. The I-Tasser server is a hierarchical protein 3D structure modeling approach (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) based on the protein sequence involving a variety of integrated algorithms. A detailed overview in which algorithms were involved in the I-Tasser method has been described by Yang Zhang 2007 and 2008 [174, 175].

The target protein sequence, submitted as FASTA format, was first guided through the PDB structure library with a cut-off of 70% pairwise sequence identity to find possible folds by variants of Profile-Profile alignment. The continuous fragments were then used to reassemble full-length models and unaligned regions were built by *ab initio* modeling. In a second iteration, selected fragments were extracted from the model and screened a second time against the PDB structure library. Finally, the structures were clustered and the lowest energy structure in each cluster was selected. This simulation has been run for separate domains as well as for the full sequence. The final 3D model was automatically generated by assembling the single domains together in a full-length model [175]. The complete results including threading templates, predicted final models, proteins structurally close to the target sequence with their ligand binding and active sites.

The final 3D model can be downloaded as PDB file and visualized by the molecular graphic system PyMOL (https://www.pymol.org/), as shown in Figure 13. All 3D images were generated using PyMOL version 1.3 on Mac OS X. The program provides also the options to create an alignment of 3D protein structures, different visualization models and a separate view for binding ligands.

Figure 13: Predicted 3D protein structure of the translated protein sequence of the reference gene sequence ECOASTAA (*astA* / EAST1) using the I-Tasser server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) and PyMOL (https://www.pymol.org/).

4. Results

4.1 Toxins of *E. coli* and *Shigella* spp.

The focus of this thesis to compile a set of all known toxin genes which are known to be expressed by *E. coli* and *Shigella* spp. from previous studies (references are listed in Table 3, p. 23 ff.), was based on the toxin definition by Henkel *et al.* (2010), that toxins damaging human and animal target cells through their own action [59]. Furthermore, based on their mechanisms of interaction on target cells all toxins resulting from the literature research were classified into three groups:

i) Intracellular acting toxins (IATs),
ii) Membrane damaging toxins (MDTs) and
iii) Receptor targeted toxins (RTTs).

(see chapter 2.3 Toxins as important virulence factors, p. 6 ff.)

The group of IATs consists of the SPATEs family, the Subtilase Cytotoxin SubAB, and the Shiga-Toxins Stx1 and Stx2. These toxins are able to cross the target cell membrane and damage the cell from the inside by cleaving specific intracellular substrates.

The pore-forming cyto- / hemolysins EhxA, ClyA (HlyE) and HlyA, which create channels in the target cell's membrane form the group of MDTs. Through the resulting channels an unregulated calcium influx into the host cell leads to cytoskeletal destruction and finally to cell lysis.

Included in the group of RTTs are the heat-labile enterotoxin (LT) and heat-stable enterotoxins (EAST1, STa and STb), type III effector proteins (Cif, EspF, EspH, IpaB, IpgD, Map, Tir and VirA), cytotoxins (CdtVa-c, Cnf1 and Cnf2), toxins involving colibactin (ClbA-Q), the dynamin like-protein LeoA, the lymphocyte inhibitory factor LifA (Efa1), *Shigella* enterotoxins (ShET1 and ShET2) and the metalloprotease YghJ. These toxins stimulate intracellular signaling pathways after binding to the appropriate cell receptors. The induced signals release a broad and extremely diverse spectrum of effects on target cells, which results in an increased or decreased fluid / cell compound secretion, cell cycle arrest in different phases of mitosis, DNase activity or DNA double-strand breaks, permanent activation (GTPase) or inhibition (NF-κB) of signals and the promotion of new cell activities (apoptosis).

All until June 2016 published 39 toxins expressed by *E. coli* and *Shigella* spp. were collected, classified based on their mechanisms of interacting on target cells and described by their specific effects in Table 3, p. 23 ff.

4.2 Genetic variability of toxins expressed by *E. coli* and *Shigella* spp.

In order to identify allelic variations in the nucleotide sequence of each toxin released by *E. coli* and *Shigella* spp., the first step in this thesis was to compare all the detected toxin gene nucleotide sequences from NCBI and the internal ECore database for each reference toxin gene in an alignment. Subsequent phylogenetic analyses were generated by neighbor-joining method to use the resulting gene clusters to identify allelic variations. After identification of all allelic variations of a toxin gene, the gene sequences were translated into their protein sequences to unveil the genetic alleles, which result in changes in the protein sequences. Finally, the resulting secondary structure and the three-dimensional structure of these protein sequences were predicted, to give an overview of similar or different protein structures in order to receive first indications of potentially functional changes.

The results of these analyses were assembled and graphically displayed in a tree diagram for each toxin, which is exemplarily represented in Figure 14. Such an illustration includes all identified alleles of the nucleic acid (NA) and variants of the resulting amino acid (AA) sequences, the predicted secondary (Sec.) and three-dimensional (3D) protein structures.

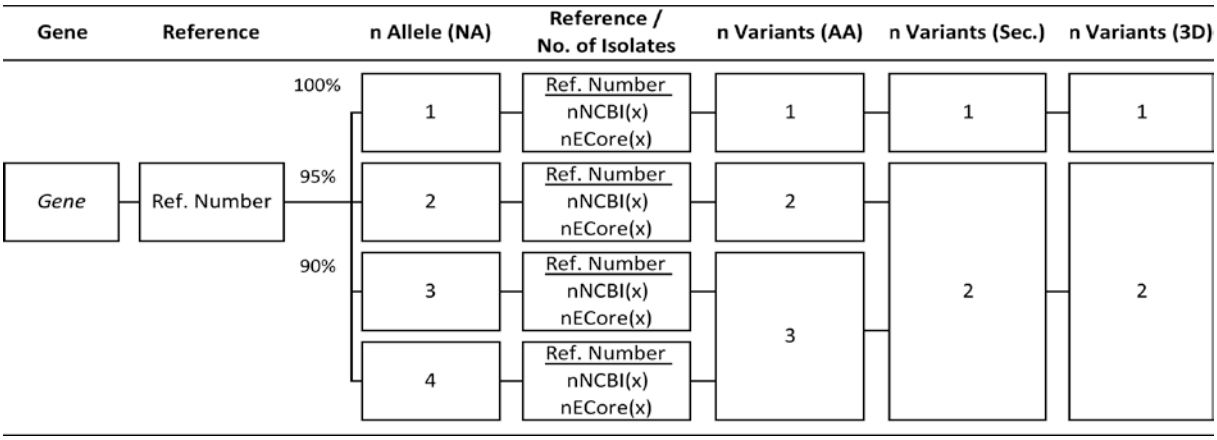

Figure 14: Tree diagram to represent sequence analysis and phylogenetic distribution of allelic variations (n Allele / Variants: number of identified alleles / variants) of a gene (NA: nucleic acid) and its resulting protein (AA: amino acid) sequences, as well as predicted secondary (Sec.) and three-dimensional (3D) structures with respect to the reference sequence (genetic identity in percent).

The example shows four allelic variations in the nucleic acid sequence representing genetic diversities of 90% to 100% identity compared to the chosen reference gene sequence. Each allele was also presented by an underlined reference sequence to give an example for all the collected identical gene sequences from the NCBI and internal ECore database. With regard to the number of isolates identified in both NCBI (nNCBI) and the ECore database (nECore), the frequency of occurrence of an allelic variation was determined. The subsequent columns display either identical or different amino acid sequences derived from the respective nucleotide alleles, likewise the secondary and three-dimensional protein structures derived from the amino acid sequence result either in identical or different structures. The example shows two resulting allelic protein model variants with possible functional differences.

4.2.1 Intracellular acting toxins (IATs)

The bioinformatics results for the identification of allelic variations showed that allelic variations appear for each IAT, including the SPATEs members, SubAB, Stx1 and Stx2. Table 4 shows the number of alleles identified for each toxin relating to their nucleic acid (NA) sequences and variants within their amino acid (AA) sequences as well as their secondary (Sec.) and three-dimensional (3D) protein structure variants.

Table 4: Occurrence of allelic variations among intracellular acting toxins (IATs)

IATs	Gene	n Sequences	n Allele (NA)	n Variants (AA)	n Variants (Sec.)	n Variants (3D)
SPATEs	*sigA*	31	11	6	3	1
	pet	4	3	3	3	1
	sat	45	19	13	5	1
	espP / pssA	40	9	5	4	1
	espC	14	9	5	2	1
	epeA	2	2	2	2	1
	espI	18	10	9	4	1
	eatA	25	15	9	9	1
	sepA	23	14	10	6	1
	pic	48	4	4	4	1
	hbp / tsh	16	13	12	7	1
	vat	92	24	18	6	1
Subtilase cytotoxin	*subAB*	32	25	12	7	4
Shiga Toxin	*stx1AB*	97	19	12	3	3
	stx2AB	104	27	23	10	7

The number of alleles/variants based on sequence/structure differences in the respective alignments. n: Number; NA: nucleic acid sequence, AA: amino acid sequence; Sec.: secondary structure; 3D: three-dimensional structure

The results in Table 4 display allelic variations for each SPATE member within their gene and resulting protein sequences. Nevertheless, the genetic variations in their secondary structures do not influence the resulting protein model and result for each SPATE member in a single three-dimensional protein structure. In contrast to the SPATEs family, the toxins SubAB, Stx1 and Stx2 showed allelic variations, which result in various three-dimensional protein structures.

Due to the close relationship of the SPATE gene members to each other [132] and the multitude of allelic variations for each gene, the next step in this thesis was therefore to identify shared similarities of all published SPATE members using phylogenetic analyses. The resulting phylogenetic tree enables a clear differentiation between the SPATE members due to the correct assignment of allelic variations to the chosen reference sequences (Figure 15, p. 36). The phylogenetic distribution shows also that published genes like *espP* and *pssA* as well as *hbp* and *tsh* describe an identical serine protease autotransporter sequence despite the different gene designations.

In this thesis, the allelic gene sequences of each class-1 SPATE share similarities from 99 to 100% (see Appendix Figure A 1 to Figure A 4, pp. xxvii ff.), while the class-2 SPATEs showed identities of 90 to 100% (see Appendix Figure A 5 to Figure A 11, p. xxix ff.) indicating that class-2 SPATEs are genetically more diverse.

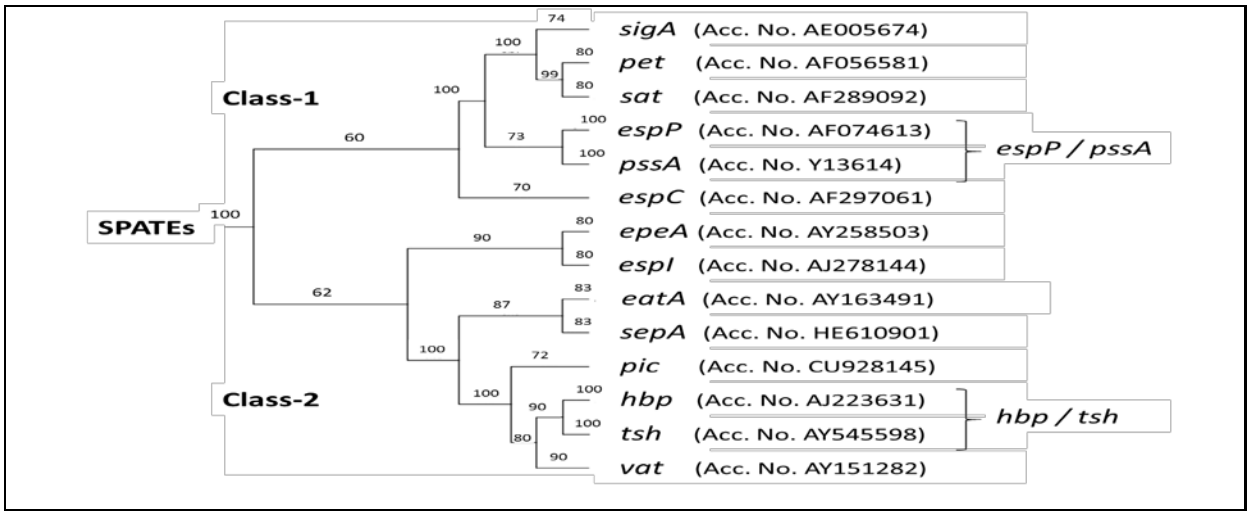

Figure 15: Phylogenetic tree of the SPATEs reference gene sequences showing sequence identity in percent

Figure 16 demonstrates, exemplarily for the SPATEs, the resulting gene analysis of the class-1 SPATE *sigA,* which exhibits eleven gene alleles. These allelic sequences resulted in six protein sequence variants. This is due to synonymous SNPs, which in two cases of nucleotide changes is the result of three alleles translating into one identical amino acid sequence. Due to the use of mutation matrices calculating similar or different properties of amino acids, the protein variants could be combined in three secondary structure variants. Finally, the individual differences in secondary structure were not essential to cause differences in the protein model and therefore result in a single identical three-dimensional protein model. Comparable to the obtained results for *sigA*, the bioinformatics analyses of all the other class-1 (see Appendix Figure A 1 to Figure A 4, pp. xxvii ff.) and class-2 SPATEs (see Appendix Figure A 5 to Figure A 11, pp. xxix ff.) showed comparable results leading to one single three-dimensional protein model for each SPATE.

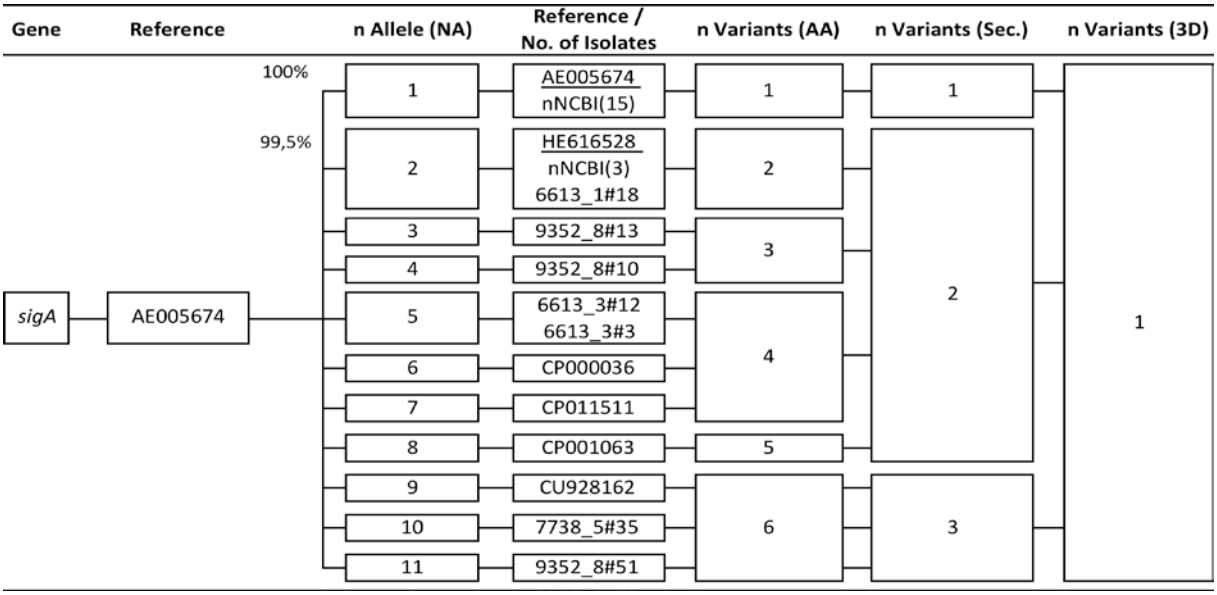

Figure 16: *sigA* (SPATE – Serine protease autotransporter of *Shigella flexneri* 2a) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AE005674 (genetic identity in percent).

The phylogenetic analysis of allelic variations of *subAB* demonstrated a high genetic variability in its gene sequences ranging from 90 to 100% identity, which is also reflected by the diverse variants of the resulting protein sequences as well as secondary and three-dimensional structure variants. The current analysis included, among the identified sequences in NCBI, all allelic *subAB* variants of previous molecular analyses, published from Funk *et al.* 2013. The configuration of clustering according their genomic location of the *subA* units [44] were confirmed in the resulting three-dimensional protein structure for the whole *subAB* gene variants in the present data (Figure 17).

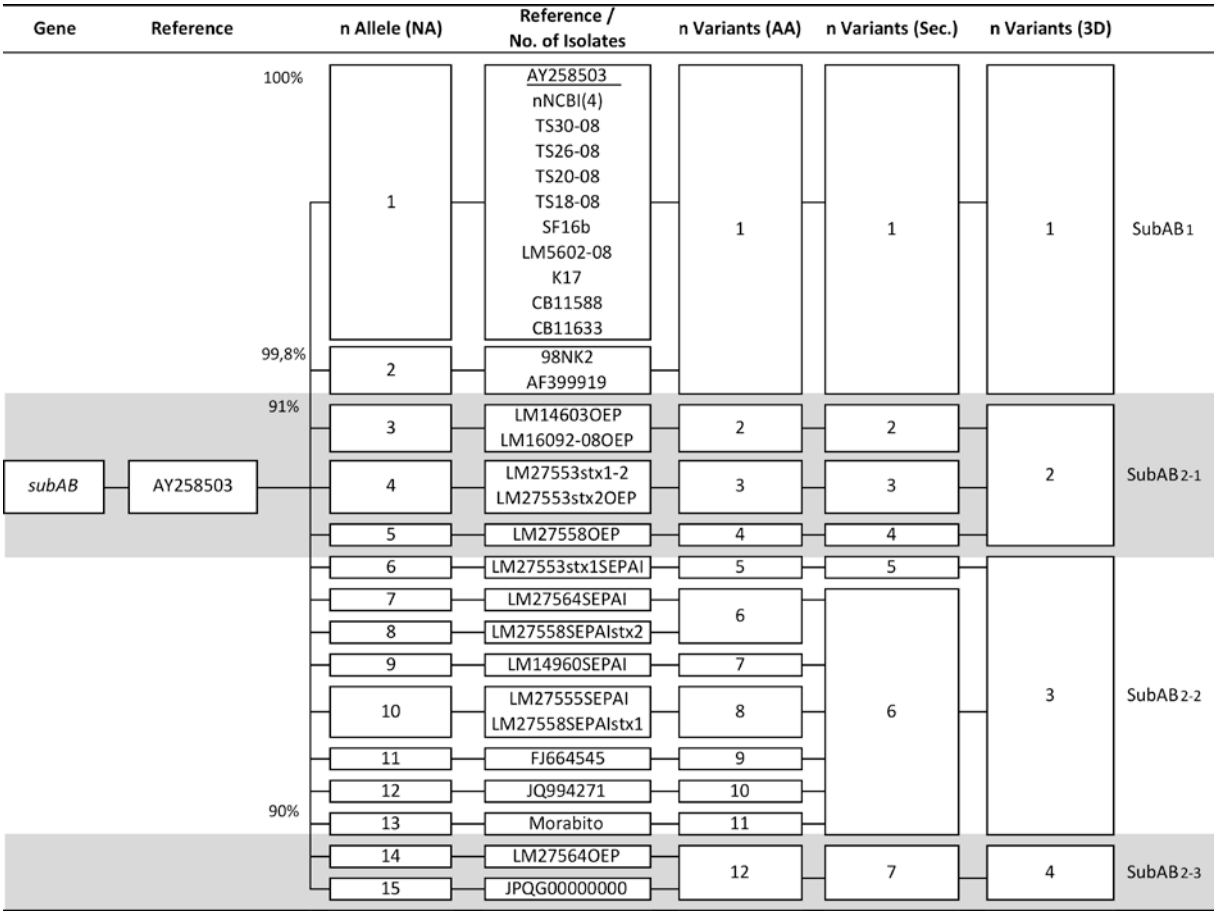

Figure 17: *subAB* (Subtilase Cytotoxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AY258503 (genetic identity in percent). – Analysis including gene sequences published by Funk *et al.* 2013 [44].

The gene sequences (nNA 1 - 2), which were located on a homogeneous virulence plasmid, revealed the three-dimensional variant SubAB$_{1}$. The chromosomally located gene variants that were encoded on the SE-PAI (nNA 3 - 5) resulted in the three-dimensional variant SubAB$_{2\text{-}1}$ while those encoded on the OEP-locus (nNA 6 - 13) resulted in SubAB$_{2\text{-}2}$. The strain LM27564OEP (nNA = 14) of the study by Funk *et al.* 2013 [44], which falls despite its location on the OEP-locus into the SE-PAI cluster, revealed a completely independent three-dimensional variant SubAB$_{2\text{-}3}$ in this thesis. Also included in the SubAB$_{2\text{-}3}$ variant is the *subAB* gene sequence of the human *E. coli* isolate FHI42 (nNA = 15) located on a new *subAB* locus associated with a currently hypothetical protein, published by Nüesch-Inderbinnen *et al.* 2014 [111].

The Shiga-Toxin family could be categorized into the two main subtypes Stx1, including the variants Stx1a, Stx1c and Stx1d being related to the Shiga-Toxin produced by *Shigella dysenteriae* and the immunologically distinct Stx2, including the variants Stx2a to Stx2h. Numerous *E. coli* (STEC / EHEC) and *Shigella* strains have been already analyzed with regard to their expressed Shiga-Toxins variants [5, 70, 67, 73, 95, 135, 167, 171]. In this thesis, the consistent nomenclature for subtyping summarized by Scheutz *et al.* 2012 [135] was used to classify the strains of the internal ECore collection and the according *stx* gene reference sequences were used to supplement this analysis. The *stx1AB* (Figure 18) and *stx2AB* (Figure 19, p. 39) gene sequences, collected from the NCBI and ECore database, were classified according the defined Stx1 and Stx2 subtypes.

Figure 18 shows the 97 collected *stx1AB* gene sequences which resulted in 19 nucleic acid alleles. The sequence similarities ranged from 92% identity for *stx1d* and 97% identity for *stx1c* compared to the reference sequence AE005174 (stx1a). At the end of the analysis, all the collected sequences could be divided into the three predefined Stx1 subtypes.

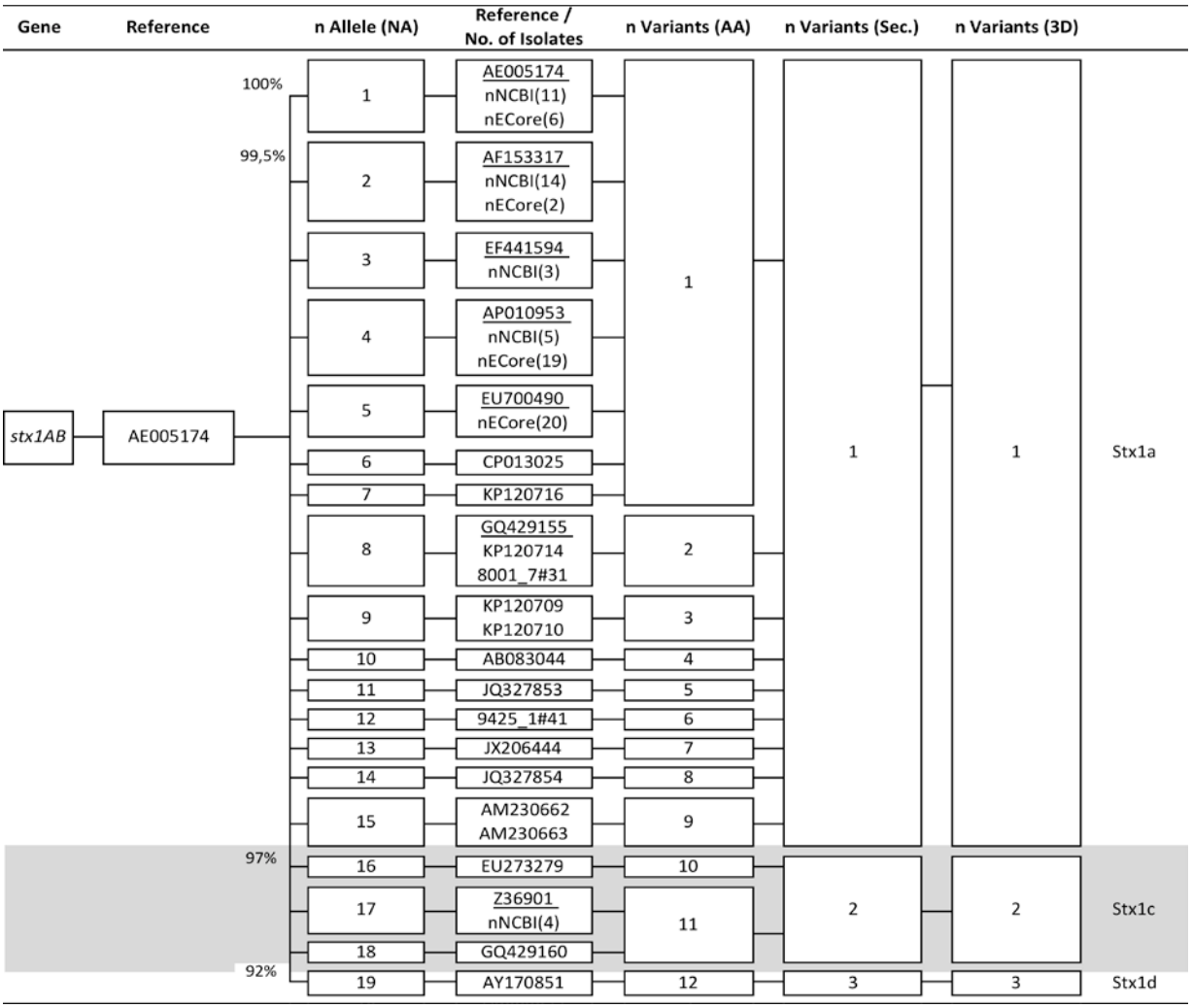

Figure 18: *stx1AB* (Shiga-Toxin 1) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AE00517 (genetic identity in percent).

The 104 collected *stx2AB* sequences, summarized in Figure 19, clustered also into the predefined Stx2 subtypes including Stx2a – Stx2h. Sequence similarities ranged from 71% identity of the most genetically distanced *stx2f* gene sequence AB472687 to 100% identity of the *stx2a* reference gene sequence AE005174.

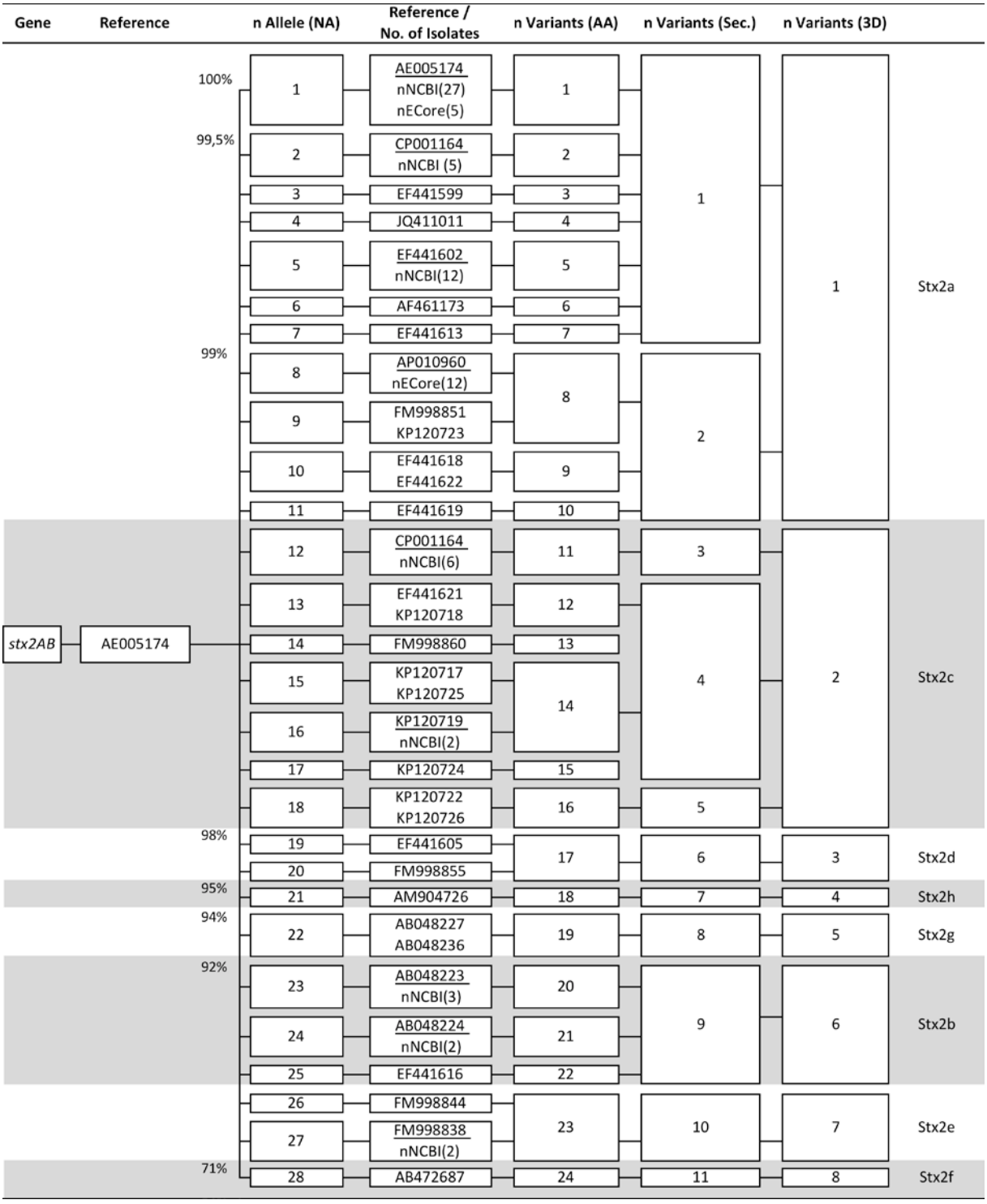

Figure 19: *stx2AB* (Shiga-Toxin 2) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AE005174 (genetic identity in percent).

4.2.2 Membrane damaging toxins (MDTs)

The bioinformatics results of the pore-forming haemolysins of *E. coli* in Table 5 indicated a high occurrence of allelic variations for EhxA and HlyA among their nucleotide (nNA_{EhxA} = 22; $nNA_{Hly\,A}$ = 22), their amino acid sequences (nAA_{EhxA} = 16; $nAA_{Hly\,A}$ = 14) and their resulting secondary structures ($nSec._{EhxA}$ = 8; $nSec._{Hly\,A}$ = 5).

Table 5: Occurrence of allelic variations among membrane damaging toxins (MDTs)

MDTs	Gene	n Sequences	n Allele (NA)	n Variants (AA)	n Variants (Sec.)	n Variants (3D)
Cyto- / Haemolysins	*ehxA*	69	22	16	8	1
	clyA / hlyE	266	5	3	2	1
	hlyA	61	22	14	5	1

The number of alleles/variants based on sequence/structure differences in the respective alignments. n: Number; NA: nucleic acid sequence, AA: amino acid sequence; Sec.: secondary structure; 3D: three-dimensional structure

In contrast to the other MDTs, cytolysin ClyA also known as haemolysin HlyE indicated a lower genetic diversity in its nucleotide (nNA = 5) and protein sequences (nAA = 3) as well as in its secondary structures (nSec. = 2). Figure 20 shows exemplarily for the MDTs the bioinformatics results of the 266 *clyA / hlyE* collected gene sequences which clustered in five gene alleles resulting in a single three-dimensional protein model. The nucleotide sequence similarities amount to around 99.5% to 100% identity compared to the chosen reference gene sequence AY576656.

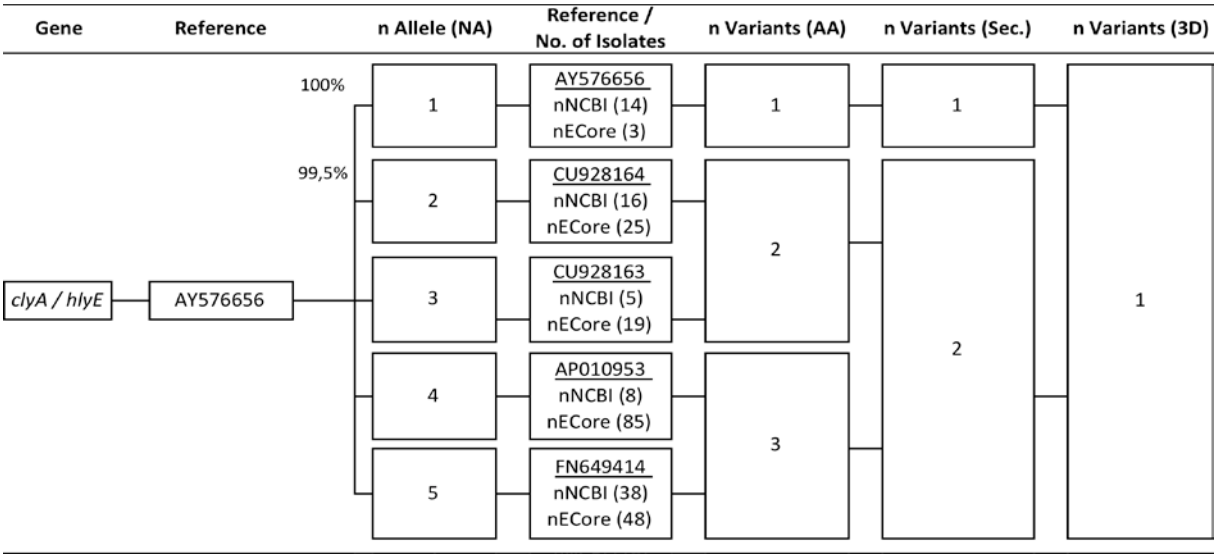

Figure 20: *clyA* / *hlyE* (Cytolysin A / Hemolysin E) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AY576656 (genetic identity in percent).

The gene alleles for *ehxA* shared similarities from 97% to 100% identity (see Appendix Figure A 12, p. xxxiii) and for *hlyA* from 96% to 100% identity (see Appendix Figure A 13, p. xxxiv) referring to their reference gene sequences AF043471 and AP010959. At the end of the analysis, all the collected 69 *ehxA* and 61 *hlyA* gene sequences resulted also in one single EhxA and HlyA three-dimensional protein model like ClyA / HlyE.

4.2.3 Receptor targeted toxins (RTTs)

The bioinformatics results to identify allelic variations among each RTT are summarized in Table 6 and were categorized with respect to their functional activity or location.

Table 6: Occurrence of allelic variations among receptor targeted toxins (RTTs)

RTTs	Gene	n Sequences	n Allele (NA)	n Variants (AA)	n Variants (Sec.)	n Variants (3D)
Enterotoxins	LT (*elt*)	75	22	15	3	1
	EAST1 (*astA*)	99	22	15	2	1
	STa *(estIa)*	35	7	6	2	1
	STb *(estIb)*	23	3	3	1	1
	set1 (ShET1)	23	11	8	2	1
	set2 (ShET2)	25	10	9	2	1
Type III effector proteins	*cif*	95	8	4	1	1
	espF	38	14	11	6	2
	espH	59	14	4	3	3
	map	45	18	13	5	1
	ipaB	23	14	12	1	1
	ipgD	49	29	20	1	1
	tir	32	16	10	2	1
	virA	23	12	8	1	1
Cyto- / Genotoxins	*cdtVa-c*	48	17	8	4	1
	cnf1	36	13	8	1	1
	cnf2	31	11	6	1	1
	clbA-Q	22	10	1	1	1
Dynamin like-proteins	*leoABC*	12	10	10	6	1
Lymphocyte inhibitory factor / EHEC factor for adherence	*lifA / efa1*	72	16	13	2	1
Metalloprotease	*yghJ*	185	25	21	11	7

The number of alleles/variants based on sequence/structure differences in the respective alignments. n: Number; NA: nucleic acid sequence, AA: amino acid sequence; Sec.: secondary structure; 3D: three-dimensional structure

The bioinformatics analyses of diarrhea inducing enterotoxins of *E. coli* and *Shigella* spp. resulted for each enterotoxin in a single three-dimensional protein structure, indicating a highly conserved toxin group (Table 6).

The results of the database gene search for the heat-labile enterotoxin LT (*elt*) yielded 75 gene sequences, which clustered into 22 unique gene alleles. These gene alleles showed sequence homologies of about 99% to 100% identity compared to the reference gene sequence CP000795. The resulting 15 protein sequences are summarized into three secondary structure variants. Finally, the scattered differences in the secondary structure variants were not essential for protein modeling and resulted in one three-dimensional protein model (see Appendix Figure A 14, p. xxxv).

Comparable to the results of the heat-labile enterotoxin LT (*elt*), the sequence analyses of the heat-stabile enterotoxins EASTA (*astA*), Sta (*estla*) and Stb (*estlb*) as well as both *Shigella* enterotoxins *set1* (ShET1) and *set2* (ShET2) resulted in a single three-dimensional model (Table 6, p. 41).

The 99 identified gene sequences of the heat-stable enterotoxin EASTA (*astA),* with sequence similarities of 95% to 100% identity with regard to the reference sequence ECOASTAA, splited into 22 unique gene alleles. The resulting 15 protein sequence variants revealed two secondary structure variants and finally resulted also in one three-dimensional protein model (see Appendix Figure A 15, p. xxxvi).

The phylogenetic results of the identified 35 Sta (*estla*) and 23 Stb (*estlb*) gene sequences indicated the most homogeneous genetic structure of the enterotoxins. The seven gene alleles of Sta (*estla*) with sequence homologies with round about 99% to 100% identity to the reference gene sequence FN649417, yielded in six resulting protein variants, which resulted in two secondary structure variants. The three gene alleles of Stb (*estlb*) with sequence identities of round about 95% to 100% compared to the reference sequence FN649418, resulted in three protein variants and in a single secondary protein structure. As already mentioned, the gene alleles of Sta (*estla*) and Stb (*estlb*) leaded in the end to a consistent three-dimensional protein model for each of the two enterotoxins (see Appendix Figure A 16 for Sta and Figure A 17 for Stb, p. xxxvii).

The gene search for the remaining two enterotoxins revealed 23 gene sequences for *set1* (ShET1) and 25 gene sequences for *set2* (ShET2). These sequences clustered into eleven *set1* (ShET1) gene alleles showing sequence identities in a range from 98% to 100% and for *set2* (ShET2) in ten gene alleles with sequence identities around 99.5% to 100% compared to the reference gene sequence AF348706. The gene clusters resulted in eight protein variants for *set1* (ShET1) and nine protein variants for *set2* (ShET2). Finally, the protein variants of both enterotoxins produced two secondary protein structures, which resulted likewise in one three-dimensional protein model (see Appendix Figure A 18 for *set1* and Figure A 19 for *set2*, p. xxxviii).

The type III effector proteins, which are delivered to infected cells by the type III secretion system (T3SS) of *E. coli* to facilitate unspecific host cell functions, were also analyzed in the toxin group of RTTs. Included into this group were the type III effector proteins Cif, EspF, EspH, Map, IpaB, IpgD, Tir and VirA (Table 6, p. 41).

Figure 21 illustrates the bioinformatics results of the secreted effector protein EspF. The 38 identified gene sequences showed a high sequence diversity of about 85% to 100% identity compared to the reference sequence AE005174, which is also reflected in the high occurrences of allelic variations in the gene (nNA = 14) and the translated protein sequences (nAA = 11). After predicting the secondary structure, the six remaining secondary structure variants leaded to two different three-dimensional protein models. Both models differ in the sequence length of the number of proline-rich repeats in the C-terminal end of the sequence (amino acid residue 74 ascend with regard to the variants). These repeats were identical in size (47 AA) and their sequence was homologous. The *espF* reference gene sequence, originating from EHEC strain EDL933 (Acc. No. AE005174) consists of four identical proline-rich repeats. The sequences that resulted in a second variant of the three-dimensional model, for example the *espF* sequence from the EPEC strain E2348/69 (Acc. No. FM180568), consist of only three proline-rich repeats. However, phylogenetic differences between the gene alleles located in the active N-Terminal domain, which contains the secretion-signal sequence that is sufficient to trigger intracellular signaling pathways, had no influence on the resulting protein model [62].

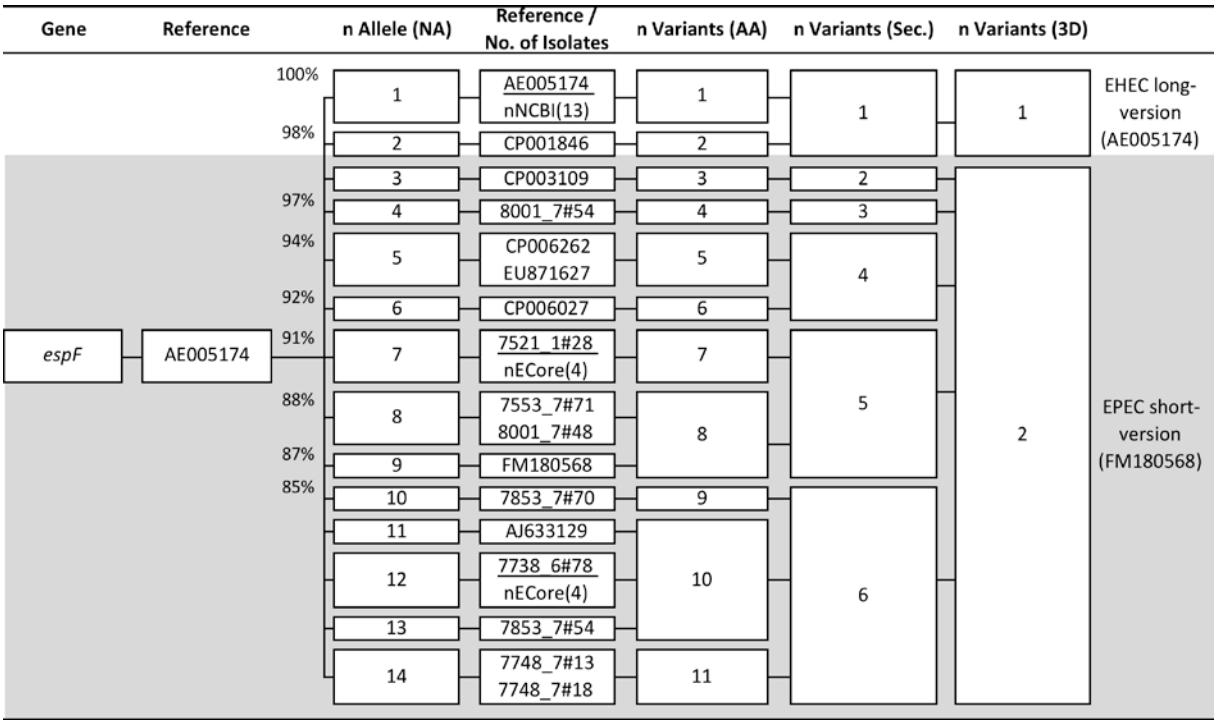

Figure 21: *espF* (T3SS - secreted effector protein) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AE005174 (genetic identity in percent).

Figure 22 summarizes the bioinformatics results of the secreted effector protein EspH. The identified 59 gene sequences, which cluster into 14 gene alleles, showed sequence identities around 95% to 100% compared to the reference sequence FM180568. The resulting four protein sequences yielded in three secondary structure variants, which also remained different from one another in their according predicted three-dimensional protein models. The three models differed in their structural three-dimensional protein arrangement. These three-dimensional protein structure variants of EspH originated from EPEC strain E2348/69 (Acc. No. FM180568), EHEC strain IHIT1190 (Acc. No. FM986651) and EHEC strain 11128 (Acc. No. AP010960). In contrast to the EHEC strains, which exhibited three α-Helices, the EPEC strain showed four α-helices in the predicted three-dimensional model. All of them shared four β-barrel structures but are located at different positions in the model (see chapter 4.3 Three-dimensional structure prediction of selected).

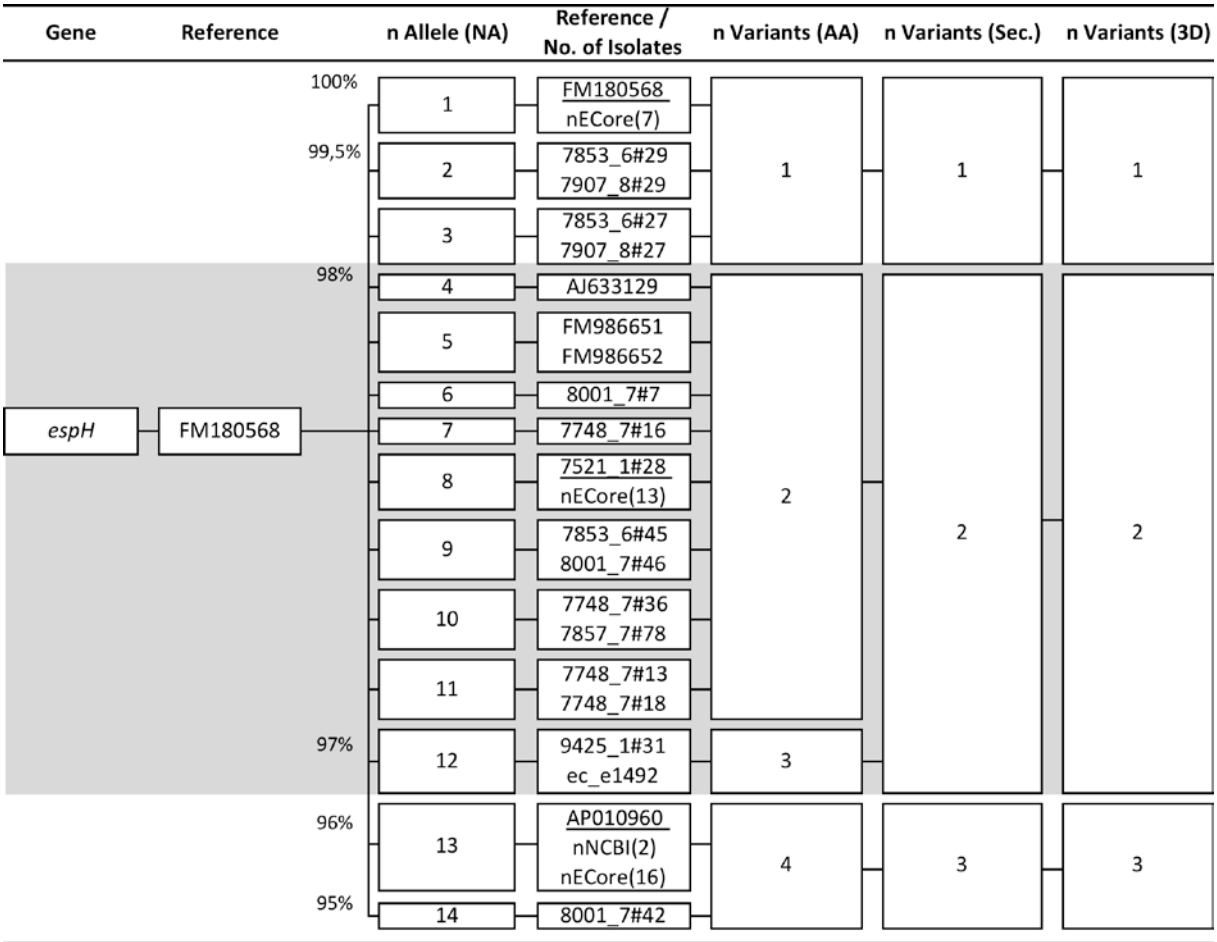

Figure 22: *espH* (T3SS - secreted effector protein) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence FM180568 (genetic identity in percent).

The remaining effector proteins of *E. coli* Cif, IpaB, IpgD, Map, Tir and VirA demonstrated allelic variations in their gene sequences resulting in various protein variants (Table 6, p. 41). Detailed graphs of the analyses of the mentioned effectors are featured in the appendix (pp. xxxix ff.). The bioinformatics analyses of the 23 selected *ipaB* gene sequences, with sequence identities of about 98% to 99% to the reference sequence AY098990 are shown in Figure A 21 (p. xxxix). The collected 49 *ipgD* gene sequences (Figure A 23, p. xli) and 32 *tir* gene sequences (Figure A 24, p. xlii) showed gene sequence identities of about 96% to 100% compared to the chosen reference gene sequence AF348706 (ipgD) and AF070067 (*tir*). The results for the 23 *virA* gene sequences in Figure A 25 (p. xlii) showed a level of low genetic diversity of 99% to 100% identity compared to the reference gene sequence AF386526, as well as the 95 *cif* gene sequences with the reference AF497476 (Figure A 20, p. xxxix). In contrast to the other remaining effector proteins, the 45 gene sequences of *map* showed a higher genetic diversity of 93% to 100% identity in respect of the reference gene sequence FM180568 (Figure A 22, p. xl). Nevertheless, the genetic alleles of *cif*, *ipaB*, *ipgD*, *map, tir* and *virA* showed a high homology regarding their nucleotide sequences and their resulting protein sequences as well as their secondary structures, which all resulted in a single three-dimensional protein model on each respective toxin (Table 6, p. 41).

Cyto- / genotoxins also belong to the group of RTTs inducing DNA breaks which result in irreversible cell cycle arrest or death of the target cells. CdtVa-c, Cnf1, Cnf2 and toxins involving into the colibactinlocus ClbA-Q (Table 6, 41) were included in this group. The bioinformatics analyses resulted in a single three-dimensional protein model for every cyto- / genotoxin, indicating a highly conserved toxin group just like the enterotoxins of *E. coli* and *Shigella* spp. (Table 6, p. 41). A detailed figure of each cyto- / genotoxin is illustrated in the appendix (p. xliii ff.).

The results of the database search for the toxin genes *cdtVa-c* yielded in 48 gene sequences, which clustered into 17 unique gene alleles. These gene alleles show sequence homologies of about 90% to 100% identity compared to the reference gene sequence KF322032. In this thesis, the whole sequence of the CDT holotoxin were analyzed together in one assembly, causing the interplay of the three subunits, composed of the active CdtVb subunit and the two binding CdtVa and CdtVc moieties, being essential for toxic activity [12, 51]. The eight resulting protein sequences resulted in four secondary structure variants. The gene sequences originating from *E. coli* shared a genetic identity of 96% to 100% and translated into the first two secondary structures, while the other two secondary structures resulted from gene sequences originated from *E. albertii*, which shared genetic identities about 90%. Nevertheless, the differences in the four secondary structure variants were not essential for the resulting three-dimensional protein model and resulted for *E. coli* and *E. albertii* in an equal three-dimensional protein model (see Appendix Figure A 26, p. xliii).

The bioinformatics results of the 36 *cnf1* and 31 *cnf2* identified gene sequences indicated a high genetic homology with sequence identities of about 99.5% to 100% compared to the reference gene sequence X70670 (*cnf1*), respectively ECOCNF2 (*cnf2)*. The thirteen gene alleles of *cnf1* translated into eight resulting protein variants, while the eleven gene alleles of *cnf2* resulted in six protein variants. Finally, the protein variants of both cytotoxic necrotizing factors resulted in each case into one secondary protein structure and accordingly into a single three-dimensional protein model for *cnf1* and *cnf2* (see Appendix Figure A 27 for *cnf1* and Figure A 28 for *cnf2,* p. xliv).

The database search for the remaining cyto- / genotoxin genes revealed 22 strains of *E. coli* harboring the pks island which is also known as colibactin locus (55140 bp). In this thesis, all genes involved in peptide-polyketide synthesis and cytopathic effects as well as the accessory proteins required for synthesis of active colibactin, were considered for the analyses (ClbA-Q) [51, 110]. Therefore, all of these genes were analyzed together in a single assembly. The highest genetic diversity of the colibactin loci in comparison the reference gene cluster AM229678 of the ExPEC strain IHE3034 (Acc. No. CP001969) were detected with only 15 SNPs (99.97% identity) within the whole pks locus indicating a highly conserved genomic region. The identified gene clusters resulted in ten gene alleles and leaded for each gene to a single three-dimensional protein model (see Appendix Figure A 29, p. xlv).

The dynamin like-proteins LeoABC expressed from ETEC strains also belong to the group of RTTs (Table 6, p. 41). A detailed illustration of LeoABC is given in the appendix (Figure A 30, p. xlv). Previous bioinformatics analyses of *leoA* and its two upstream located genes *leoB* and *leoC* suggested that LeoA functions in concert with the both dynamin-like proteins LeoB and LeoC [99]. Therefore, in this thesis the dynamin like-proteins LeoABC were analyzed together in one assembly. The bioinformatics analyses of the twelve collected LeoABC genes revealed ten unique gene alleles with sequence identities of about 98% to 100% compared to the reference gene sequence FN649414. These gene alleles resulted also in ten protein variants that accounted for six secondary structure variants. Finally, the scattered differences in the secondary structures were not essential for protein modeling and resulted in a single three-dimensional protein model.

The lymphocyte inhibitory factor LifA (EHEC factor for adherence - Efa1), which is typically produced by attaching and effacing *E. coli* that cause diarrhea and inhibits lymphocyte activation and pro-inflammatory cytokine synthesis [17], were also analyzed within the group of RTTs (Table 6, p. 41). The database search on the gene *lifA* also known as *efa1* revealed 72 gene sequences with sequence homologies of about 99% to 100% identity compared to the reference gene sequence AJ133705. These gene sequences revealed 16 gene alleles and translated into thirteen protein sequence variants. The protein variants could be summarized into two secondary protein structure variants that result in the end in a consistent three-dimensional protein model (see Appendix Figure A 31, p. xlvi).

The metalloprotease YghJ, the last analyzed toxin in the group of RTTs (Table 6, p. 41), influences intestinal colonization by degrading the major mucins in the small intestine [90]. In this thesis, 185 *yghJ* gene sequences have been collected from the NCBI and ECore database using the reference gene sequence CP000243 of the UPEC strain UTI89. The collected gene sequences showed sequence identities among 85% to 100% compared to the chosen reference and resulted in 25 allelic gene variations (Figure 23, p. 48). The 25 gene alleles translated into 21 protein sequence variations and further, after prediction, into eleven secondary structure variants. Finally, the prediction of the three-dimensional protein structure revealed seven protein model variants, which may differ from in their protein folding compactness and their metal binding preference (see chapter 4.3 Three-dimensional structure prediction of selected). The most frequently predicted protein model with 43% (n = 80), resulted from all the 185 detected *yghJ* genes in the database, was the three-dimensional protein variant number 5, including among others the *yghJ* gene of K-12 substrain W3110 (Acc. No. AP009048) as well as the EHEC strain 11128 (Acc. No. AP010960). The diverse predicted protein model variants were identified for all known pathotypes of *E. coli* and include also strains of *E. albertii.*

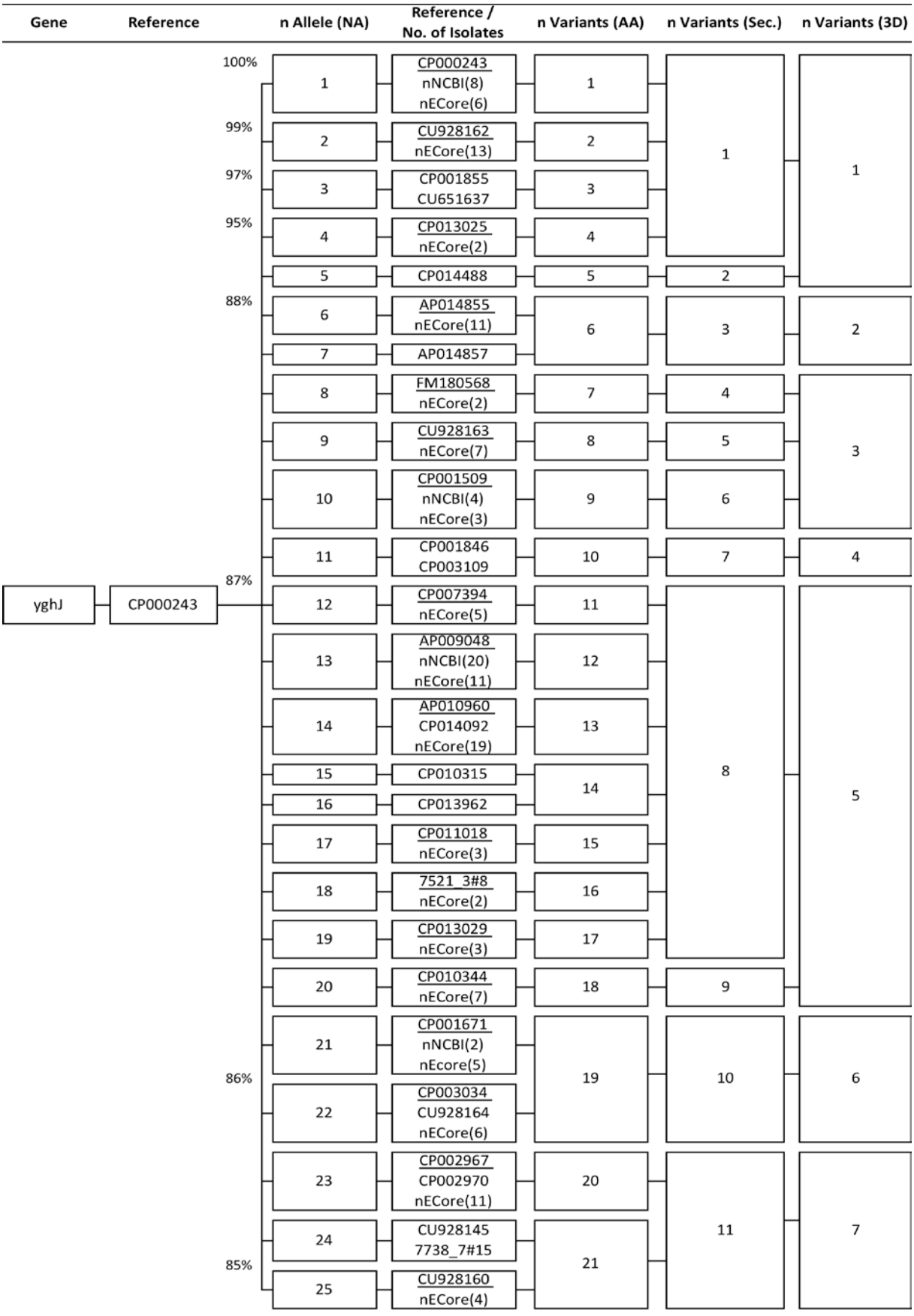

Figure 23: *yghJ* (Metalloprotease) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify allelic variants (n: number of alleles / variants) with respect to the reference sequence CU928161 (genetic identity in percent).

4.3 Three-dimensional structure prediction of selected protein variants

The bioinformatics analysis of the currently published 39 toxins expressed by *E. coli* and *Shigella* spp. did reveal allelic variations for each toxin in their gene and translated protein sequences (Table 4 for IATs, p. 35; Table 5 for MDTs, p. 40; Table 6 for RTTs, p. 41) identified by screening the ECore and NCBI databases. Furthermore, seven toxins also displayed variants in their predicted three-dimensional protein structure and therefore indicating the possibility of functional changes of the toxin. The toxins with identified variants of their three-dimensional protein models are summarized in Table 7. Table 7 consists of the twelve individual three-dimensional protein structures of the SPATEs identified in this thesis as protein variants that could be combined as variants of one SPATE toxin. Furthermore, Table 7 includes several three-dimensional protein model variants identified for *subAB*, *stx1AB* and *stx2AB* from group of IATs, as well as *espF*, *espH* and *yghJ* from group of RTTs.

Table 7: Toxins of *E. coli* and *Shigella* spp. exhibiting three-dimensional (3D) protein variants

Toxin-Group	Gene	n Variants (3D)
IATs	SPATEs	12
	subAB	4
	stx1AB	3
	stx2AB	7
MDTs	/	/
RTTs	*espF*	2
	espH	3
	yghJ	7

Previous analyses of Shiga-Toxins Stx1 and Stx2 [70, 40, 13] have already demonstrated functional differences between their allelic variations. The EspF variants have shown only differences in the number of identical proline-rich repeats, which had no influence on the resulting protein function [62]. Therefore, this chapter focuses on the results of the twelve SPATE members as protein variants of one SPATE toxin and possible functional differences of the variants of the Subtialse Cytotoxin SubAB, the effector protein EspH and the metalloprotease YghJ.

The bioinformatics results of the SPATEs have shown that genetic variations of each SPATE member resulted in an identical three-dimensional protein model, e.g. SigA (Figure 16, p. 36). The low genetic diversity among the SPATE members, as well as their close relationship to each other (Figure 15, p. 36), indicated that each SPATE member already was an allelic variant. Thus, the twelve allelic SPATE variants could be divided with respect to their three-dimensional protein model into two already known functional different classes: the cytotoxic class-1 SPATEs and the lectin-like immunomodulators class-2 SPATEs (Figure 24, p. 50) defined by F. Ruiz-Perez and J. P. Nataro 2014 [132]. Thus, no additional class could be identified in this thesis.

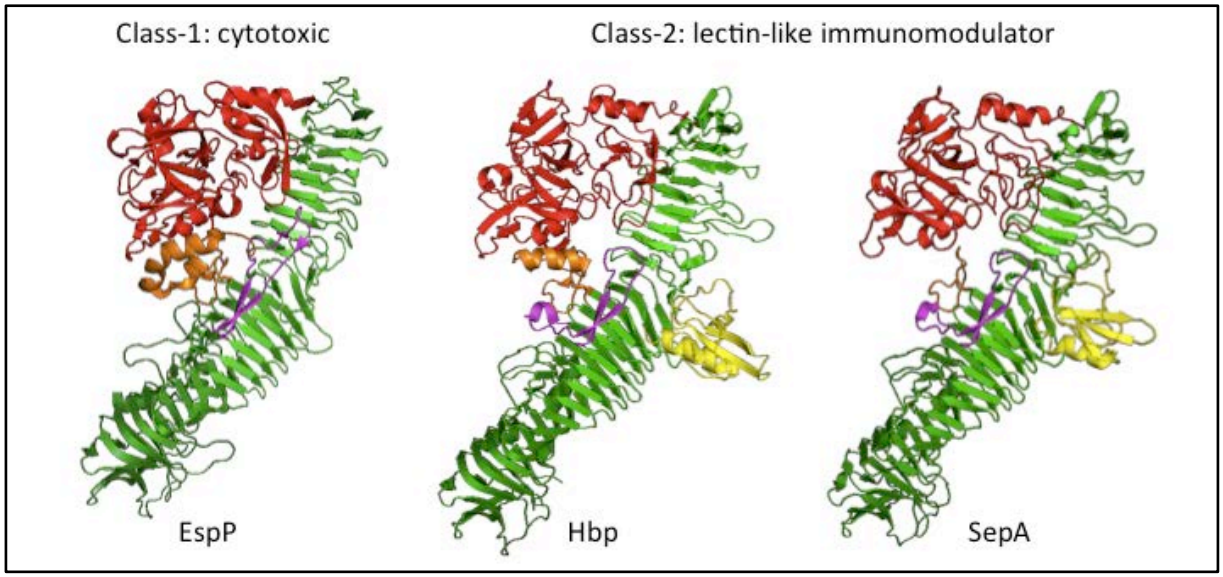

Figure 24: Three-dimensional protein model of the passenger domain of class-1 (i. e. EspP) and class-2 SPATEs (i. e. Hbp and SepA). EspP sequence originating from Acc. No. AF074613, Hbp sequence originating from Acc. No. AJ223631 and SepA sequence originating from Acc. No. HE610901 were modeled after protein sequence prediction with Geneious using the open sources server I-Tasser and the visualization software PyMol. Helices and strands colored red for the protease domain and orange for its facing domain, yellow for the chitinase-like domain and violet for a domain with yet unknown function. The parallel β-strands helix is colored in green.

Figure 24 demonstrates the differences in the predicted three-dimensional protein structures of the passenger domains of class-1 SPATE, i.e. EspP and class-2 SPATEs, i.e. Hbp and SepA. The red colored helices and strands show the protease domain, which contains the catalytic triad of the protein and its closely related orange domain, which is responsible for substrate access [132]. The folding complexity of the protease domain in red and their adjacent orange domain of class-1 SPATEs was much more pronounced than for class-2 SPATEs. Solely the orange domain reflected a helix-turn-helix motif in the model for EspP and one helix in the protein model of Hbp or a motif including only turns in the protein model of SepA. In contrast to the enzyme active domain and its substrate-binding domain, the class-2 SPATEs included an additional β-strand in the violet domain with a yet unknown function in the protein models. Furthermore, the yellow chitinase-like domain was completely missing in the protein model of EspP, a typical characteristic for all class-1 SPATEs. Common to all SPATEs protein models was the parallel β-strands helix being typical for all autotransporters. It was colored in green and nearly identical among the analyzed structures.

Classifying SPATEs could be mainly reduced to identification of the orange domain that restricts the accessibility of substrates to the protease domain. The orange helix-turn-helix domain of EspP (class-1 SPATEs) harbored two contiguous cysteines with the potential to form a disulfide bond. Therefore, this compact disulfide bond-containing domain is able to bind only small proteins, primarily intracellular substrates. In contrast, the decreased orange domain of class-2 SPATEs is able to bind mainly huge extracellular substrates such as mucins, resulting in a lower impendence of substrate access to the protease domain [132].

The bioinformatics results for SubAB have revealed four three-dimensional protein model variants that are shown in Figure 25. The originated chromosomal located *subAB* gene sequences of the SubAB model $SubAB_{2-1}$ from STEC strain LM14960 (located on SE-PAI), $SubAB_{2-2}$ from STEC strain $LM27553_{stx1}$ (located on OEP locus) and $SubAB_{2-3}$ from STEC strain LM27564 (located in a region between $subAB_{2-1}$ and $subAB_{2-2}$) are published by Funk *et al.* 2013 in a molecular analysis of Subtilase Cytotoxin genes of food-born Shiga-Toxin-producing *E. coli*. The originated plasmid located *subAB* gene sequence, which resulted in the three-dimensional model variant $SubAB_1$, were used from *E. coli* strain K17 (Acc. No. HG324027) [44].

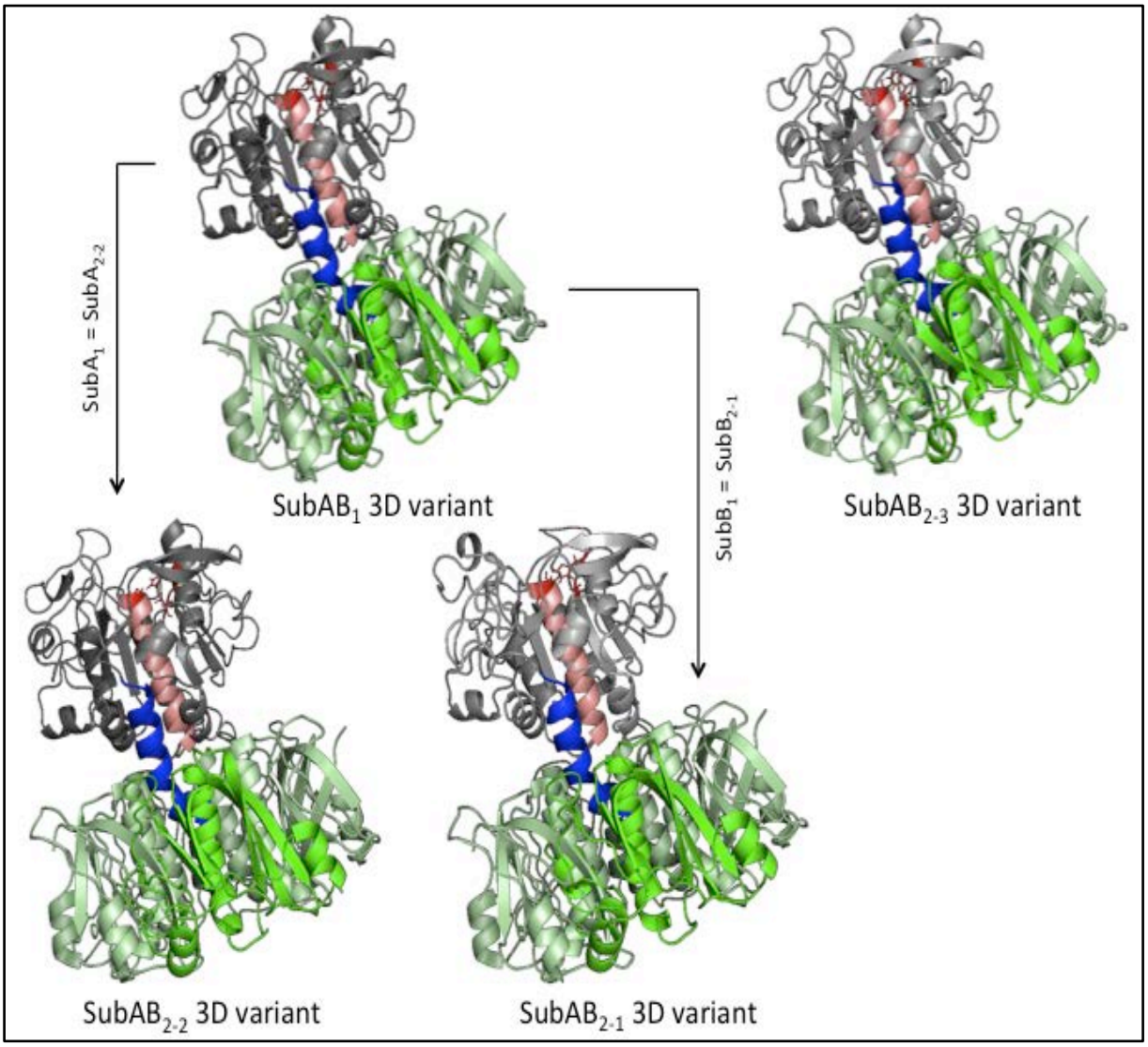

Figure 25: Three-dimensional protein model of the SubAB variants. $SubAB_1$ 3D variant sequence originating from *E. coli* strain K17, $SubAB_{2-1}$ 3D variant sequence originating from STEC strain LM14960, $SubAB_{2-2}$ 3D variant sequence originating from STEC strain $LM27553_{stx1}$ and $SubAB_{2-3}$ 3D variant sequence originating from STEC strain LM27564 [44] were modeled following the protein sequence prediction with Geneious using the open sources server I-Tasser and the visualization software PyMol. The A-subunit is shown in grey, whereas the α-helix A1 is shown in salmon, the α-helix A2 is shown in blue and the catalytic residues as red sticks (D-52, H-89, S-272). Four protomer of the pentameric B-subunit (SubB) are colored in pale green and one protomer in green to emphasize the differences between the 3D variants in the B-subunit.

All four three-dimensional SubAB variants, given in Figure 25 (p. 51), exhibited the typical protein architecture of an AB_5 Toxin comprises an A-subunit responsible for the catalytic activity shown in grey and a B-subunit that binds to the cell surface receptor shown in green. The catalytic triad of SubA responsible for the proteolytic activity was shown in red comprising the residues Asp-52, His-89 and Ser-272 (red sticks). Furthermore, SubA was composed of two α-helices, A1 colored in salmon and A2 colored in blue, which are linked by disulfide bonds. The α-helix A2 reached into the center of the pentameric B-subunit and binds over a hydrogen bond network both subunits together to form the SubAB holotoxin [81]. Four protomer of the pentameric B-subunit were colored in pale green and one protomer in green to emphasize the differences between the 3D variants in the B-subunit.

All four SubAB variants haboured no mutations in their amino acids composition (Asp-52, His-89 and Ser-272) of their catalytic triads (red sticks) and an identical spatial arrangement in their three-dimensional protein structures, which suggests that the structural basis of the proteolytic domain was conserved. Furthermore, the composition of the α-helices A1 (colored in salmon) and A2 (colored in blue) in all four SubAB variants showed also an identical spatial arrangement to bind via hydrogen bonds both subunits together, indicating functional SubAB variants. Therefore, the SubA subunits of all four present SubAB variants modeled in this thesis indicated no evidence of differences in toxicity.

In contrast to SubA, the B-subunits of the present SubAB variants could be responsible for varying cellular toxicity due to differences in their receptor binding domain. Regarding variant $SubB_1$ respectively variant $SubB_{2\text{-}1}$ in contrast to the variants $SubB_{2\text{-}2}$ and $SubB_{2\text{-}3}$, the protomer (colored in green) showed on the left side an additional short α-helix compared to the other both variants. The adjacent α-helix showed three loops in the $SubB_1$ variant, two loops in $SubB_{2\text{-}2}$ variant and only one loop in $SubB_{2\text{-}3}$ variant. These differences could be referred to a different glycolipid binding preference and therefore to a different host cell specificity as was shown for the different Shiga-Toxin variants [70, 40, 84]. Concerning these differences, the remaining structure of the three-dimensional SubB variants was nearly identical.

It should also be mentioned that the three-dimensional protein variant $SubAB_1$, the gene sequence of which is plasmid located, consists of an identical three-dimensional SubA model as variant $SubAB_{2\text{-}2}$ and an identical three-dimensional SubB model as variant $SubAB_{2\text{-}1}$. Both variants, $SubAB_{2\text{-}1}$ and $SubAB_{2\text{-}2}$, originate from chromosomal located gene sequences as described above. In contrast to these variants, the $SubAB_{2\text{-}3}$ variant shared no identical subunits with the other variants and clusters as the only food-born STEC strain together with the human STEC isolate FHI42 (see Figure 17, p. 37) published by Nüesch-Inderbinnen *et al.* 2014 [111].

The bioinformatics results for EspH have revealed three three-dimensional protein model variants that are summarized in Figure 26. Common to all three protein variants was the combination of four α-helices that surround the β-strands, whereby the spatial arrangements among the α-helices differ. Furthermore, the EspH variant 1 based on the sequence originating from EPEC strain E2348/69 (Acc. No. FM180568) exhibits four parallel β-strands in contrast to the other two variants that showed three parallel β-strands in variant 2 based on the sequence originating from EHEC strain IHIT1190 (Acc. No. FM986651) and three β-strands arranged in a curve-like manner in variant 3 based on the sequence originating from EHEC strain 11128 (Acc. No. AP010960).

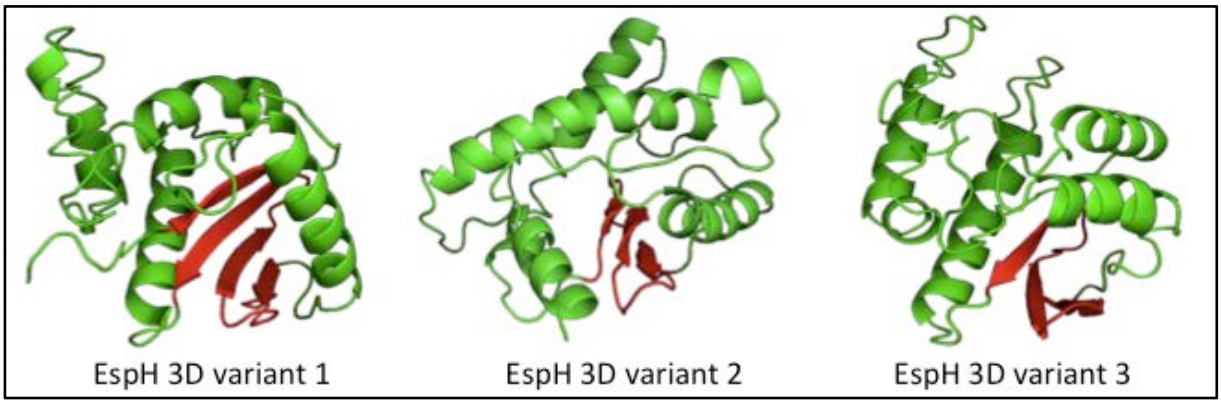

Figure 26: Three-dimensional protein model of the EspH variants. EspH 3D variant 1 sequence originating from Acc. No. FM180568 (EPEC strain E2348/69), EspH 3D variant 2 sequence originating from Acc. No. FM986651 (EHEC strain IHIT1190) and EspH 3D variant 3 sequence originating from Acc. No. AP010960 (EHEC strain 11128) were modeled following the protein sequence prediction with Geneious using the open sources server I-Tasser and the visualization software PyMol. α-helices colored in green and β-strands in red.

It is known that EspH disrupts the actin cytoskeletal structure in the host cell by modulating the assembly kinetics and morphology of actin pedestals by inhibiting a mammalian GTPase of the small Rho GTPase family, which are master regulators of actin cytoskeleton rearrangements. EPEC and EHEC strains are able to inactivate the Ras homolog GTPase (Rho) of the host cell by expressing EspH. EspH binds directly to the conserved PH-DH domain of the cellular Rho Guanine Nucleotide Exchange Factors (RhoGEFs) and blocks therefore the Rho GTPase activation of the host cell inducing cytotoxic effects. At the same time bacterial RhoGEFs that are insensitive to EspH stimulate the GTPase activation again by mimicking mammalian RhoGEFs to allow cell adhesion [28, 146, 166]. Furthermore, J. Cherfils *et al.* 2013 described that an effector protein of *Shigella* (IpgB2) harboring a comparable spatial arrangement in its three-dimensional protein structure to EspF, representing three parallel ß-strads surrounded by six a-helices, also binds to the PH-DH domain of the cellular RhoGEFs to block the mammalian Rho GTPase activation. IpgB2 can be described on basis of its conserved tryptophan and glutamic acid residue, separated by three variable amino acids (^{62}WxxxE^{66}) as a bacterial effector of the RhoGEFs family, which are structurally related to the SopE effector of *Salmonella* [18].

With regard to the structural difference of IpgB2 in comparison to EspH, which binds the same RhoGEFs over the PH-DH domain, it may be possible that all three allelic variations of EspH were able to bind the mammalian RhoGEFs. In this thesis, such an EspH - RhoGEF complex was confirmed using structural modeling for all three EspH variations (Figure 27, p. 55) with the modeler software PyMol and the add-on Autodoc, which predicts the highest probability of a substrate or ligand binding site. In contrast to the three-dimensional model, EspH showed currently no sequence homologies in its protein sequence to any other effector protein [28]. In this thesis, in all three protein variations of EspH a conserved motif including a tryptophan and glutamic acid residue separated by six conserved amino acids ^{124}WFPRTAL^{131}E was detected. Regarding the three-dimensional structure, represented in Figure 27 A (p. 55), the conserved ^{124}WFPRTAL^{131}E motif, shown by red, had no influence on the catalytic activity of EspH and was located at the end of a α-helix structure opposite the predicted catalytic loop, which is colored in blue. The catalytic loop was inserted between a α-helix and a β-strand structure located between the 20th to 37th amino acid residue. The binding surface of EspH (green) and the mammalian RhoGEF (gray) was formed by the interaction of the mentioned α-helix-catalytic loop-β-strand motif of EspH and the loop-helix region colored in yellow as well as the β-strands region of the RhoGEFs, which results in the PH-DH domain. Figure 27 B (p. 55) demonstrates the PH-DH domain of the RhoGEF, forming a pocket binding the GDP - Mg^{2+} complex on the catalytic loop of EspH (blue). GDP is shown in magenta-colored sticks and the GDP-coordinating Mg^{2+} as an orange sphere. In this catalytic pocket Mg^{2+} was displaced while a water molecule is inserted whereby the bonded GDP replaced by GTP. Quite similar processes were described for the effector protein IpgB2 of *Shigella* [74] and for other RhoGEFs [146].

Furthermore, the expression of EspH could be attributed to influence a different formation of actin pedestals between EPEC and EHEC strains. EHEC strains induced the formation of flat and faint actin structures, whereby EPEC strains form elongated pedestals [159]. Therefore, it is possible that the EspH protein variants 2 and 3 of EHEC strains (Figure 27, p. 55) bind less efficiently to the PH-DH domain of the RhoGEFs on basis of their differences in their three-dimensional order. Figure 27 B (p. 55) highlights the possibility of the catalytic loops of both EHEC strains that interact less efficiently with the GDP – Mg2+ complex in the catalytic pocket due to the spatial distance of the catalytic loop on Mg^{2+} compared with the EspH protein variant 1 (EPEC). These spatial differences putatively resulted in the formation of different actin pedestals between EHEC and EPEC strains.

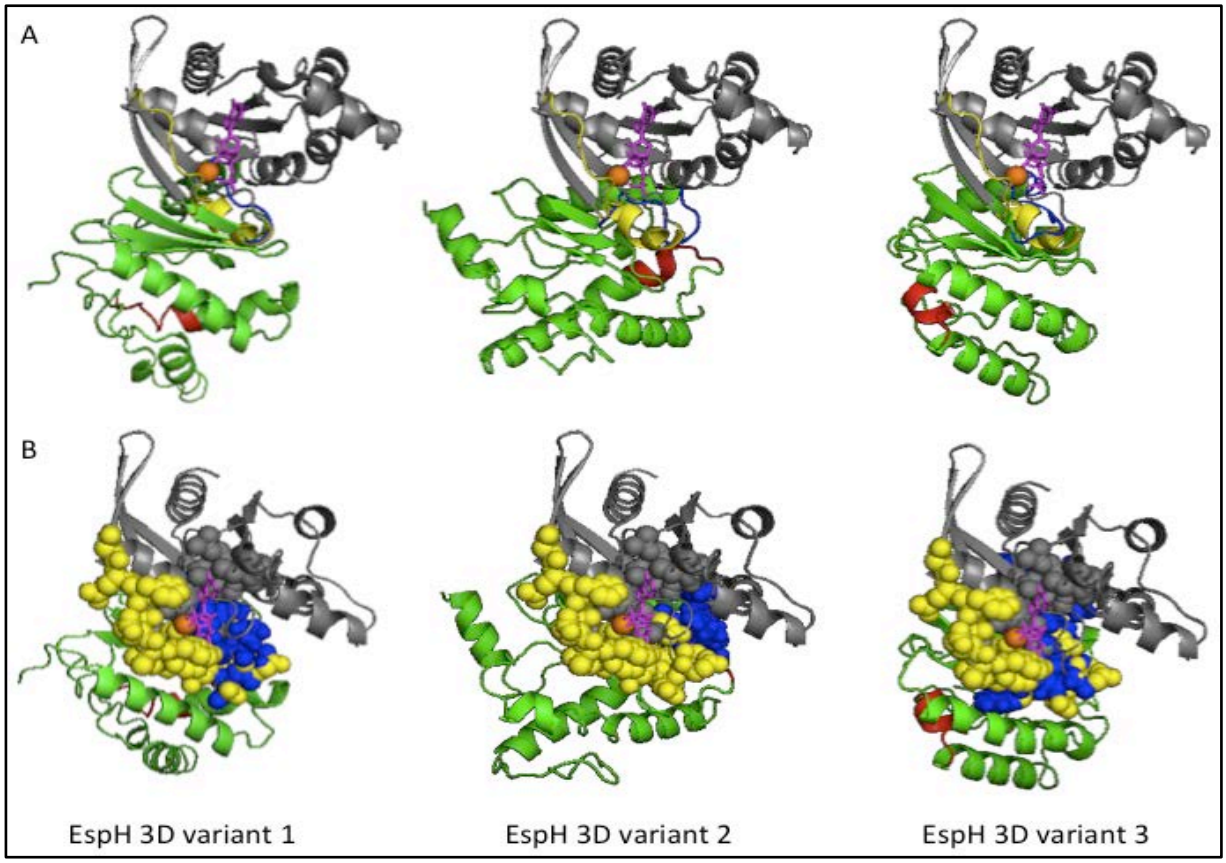

Figure 27: Three-dimensional protein model of the EspH protein variants in complex with human RhoGEF (1A2B). EspH 3D variant 1 represent the EPEC strain E2348/69 (Acc. No. FM180568), EspH 3D variant 2 represent the EHEC strain IHIT1190 (Acc. No. FM986651) and EspH 3D variant 3 represent the EHEC strain 11128 (Acc. No. AP010960). A: EspH-Rho-GDP-Mg^{2+} complex. EspH is shown in green, its conserved region in red and RhoGEF in grey. GDP is shown in magenta-colored sticks and the GDP-coordinating Mg^{2+} as an orange sphere. The catalytic region of EspH is shown as a blue loop and the binding regions of RhoGEF in yellow. B: Catalytic pocket in spheres. The PH-DH region (yellow and grey) binds GDP-Mg^{2+} (magenta - orange) on the catalytic loop of EspH (blue).

The last toxin considered in this chapter is the metalloprotease YghJ. Its bioinformatics results represented a high amount of allelic variants, which can be summarized into seven three-dimensional protein model variants (Figure 28, p. 56). These predicted protein models were very complex and correspond with a multi-domain protein. One of these domains of all seven protein variants was highly similar to a protein, which has an already known three-dimensional structure and function: the M60-like pfam13402 domain. Proteins that contain this domain have the metalloprotease function to degrade host glycoprotein and bind metal-ions [102]. Furthermore, all seven allelic YghJ protein variations exhibited the conserved sequence motif $HE_{VG}H$ in their M60-like pfam13402 domain, which belongs to the $HE_{XX}H$ motif, and is known to be involved in the catalytic activity of metalloproteases of different bacterial species. The two histidine (H) residues are ligands of the metal-ion, whereas the glutamic acid (E) residue represents the catalytic amino acid. Along this motif, an additional conserved glutamic acid seems to be also integrated into the metal binding as a third metal ion ligand, located thirteen residues across the $HE_{XX}H$ motif [102]. Such a third ligand was also present in the protein sequences of all seven YghJ variants ($HE_{VG}H_{NAAETPLxVPGAT}\underline{E}$).

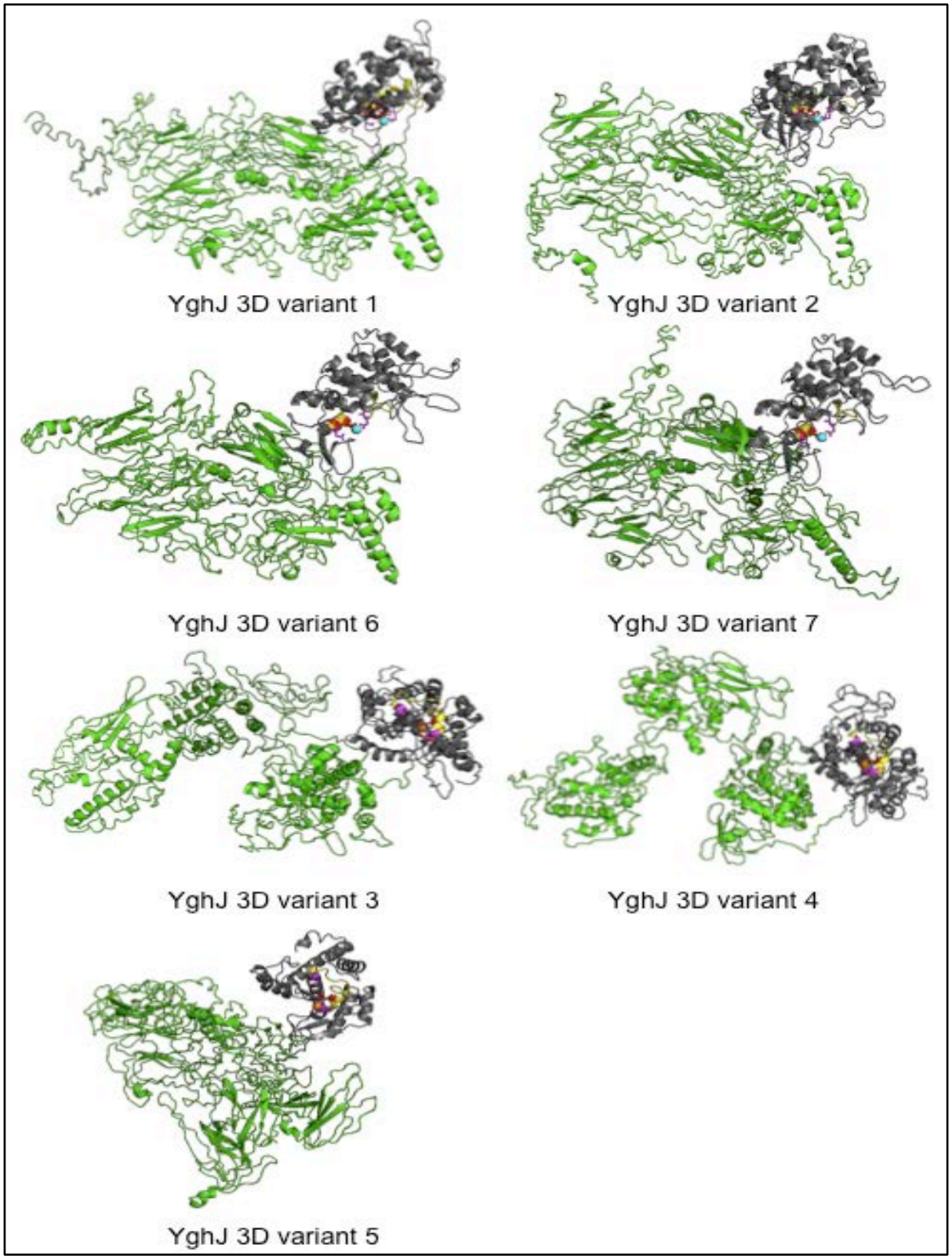

Figure 28: Three-dimensional protein models of the YghJ metalloprotease variants. The protein models were modeled with the open sources server I-Tasser and the visualization software PyMol after protein sequence prediction using Geneious. The YghJ variant 1 (Acc. No. CP000243), variant 2 (Acc. No. AP014855), variant 6 and variant 7 representing nickel binding metalloproteases on basis.of the protein 4FCA chain A of *Bacillus anthracis*. The YghJ variant 3 (Acc. No. CP01509), variant 4 (Acc. No. CP001846) and variant 5 (Acc. No. AP009048) representing zinc binding metalloproteases on basis of the protein POCK95 of *E. coli* K12. YghJ is shown in green and its M60-like pfam13402 domain in grey including the conserved catalytic $HE_{XX}H_{X(13)}H$ motif in yellow (H in red and E by magenta sticks). The binding metal-ion is shown for Zn^{2+} in an orange and for Ni^{2+} in a blue sphere.

Figure 28 (p. 56) demonstrates, that the seven predicted three-dimensional protein structures of YghJ differed not only in their protein folding compactness but also in their metal binding preference. The M60 like pfam13402 domain of the YghJ 3D variant 1 of UPEC strain UTI89 (Acc. No. CP000243), which is shown in grey color, confirmed a three-dimensional alignment with the nickel binding M60 peptidase 4FCA (chain A) of *Bacillus anthracis*. In this protein model nickel is complexed by two histidine residues H_{281} and H_{285} as well as the glutamate residue E_{299} coordinating across the $HE_{XX}H$ motif as the third ligand. The catalytic amino acid represents the glutamate E_{282}. Such a construction was also given in the catalytic $HE_{VG}H_{NAAETPLxVPGAT}E$ motif of the UPEC strain UTI89 (YghJ 3D variant 1 in Figure 28, p. 56), which is represented in yellow (H in red and E in magenta sticks). The three-dimensional structure of the catalytic motif in the M60 like pfam13402 domain of YghJ 3D variant 2 of *E. albertii* strain NIAH_Bird_3 (Acc. No. AP014855), YghJ 3D variant 6 of NMEC strain CE10 (Acc. No. CP003034) and YghJ 3D variant 7 of EAEC strain 55989 (Acc. No. CU928145) exhibited the same affinity to the nickel binding M60 peptidase of *Bacillus anthracis* as the reference strain UTI89 (Figure 28, p. 56). In this thesis, such a M60 like pfam13402 domain of *E. coli,* including a nickel binding catalytic motif was predicted with i-Tasser and modeled with PyMol with regard to the M60 peptidase 4FCA (chain A) of *Bacillus anthracis* to illustrate the probability of YghJ-metalloproteases being active in the presence of nickel (Figure 29, p. 58). Furthermore, the M60 protease 4FCA (chain A) is able to bind linked imidazole (IMD) in addition to nickel, which suggests that the YghJ metalloproteases variants are also active with a Ni^{2+}- IMD ligand. Nevertheless, the described metalloproteases may also function in the presence of zinc, because the catalytic motif of the M60 like pfam13402 domain of *E. coli* K12 strain (POCK95) prefers to bind zinc only over with the two histidine residues, which were also present in the nickel binding size (Figure 29, p. 58). The binding preferences were predicted in this thesis with the PyMol add-on Autodoc. The catalytic amino acid in the zinc-binding motif was the same glutamic residue as in the nickel-binding motif ($HE_{XX}\underline{E}$). The three-dimensional structure of the catalytic motif in the M60 like pfam13402 domain of the YghJ 3D variant 3 of *E. coli* strain BL21(DE3) (Acc. No. CP001509), YghJ 3D variant 4 of EPEC strain CB9615 (Acc. No. CP001846) and YghJ 3D variant 5 of *E. coli* K12 sub-strain W3110 (Acc. No. AP009048) exhibited these structural affinities of the zinc binding M60 like pfam13402 domain POCK95 of the *E. coli* K12 strain (Figure 28, p. 56).

With regard to the whole YghJ protein, the three-dimensional structure of the YghJ 3D variant 3 and 4 were distorted in separate bundles in contrast to the other YghJ models (Figure 28, p. 56). One of these separate bundles represented the M60 like pfam13402 domain in grey. Such a defective protein folding could lead to a nonfunctional metalloprotease. The YghJ 3D variant 5 showed the highest protein folding compactness of all the selected YghJ variants in Figure 28 (p. 56).

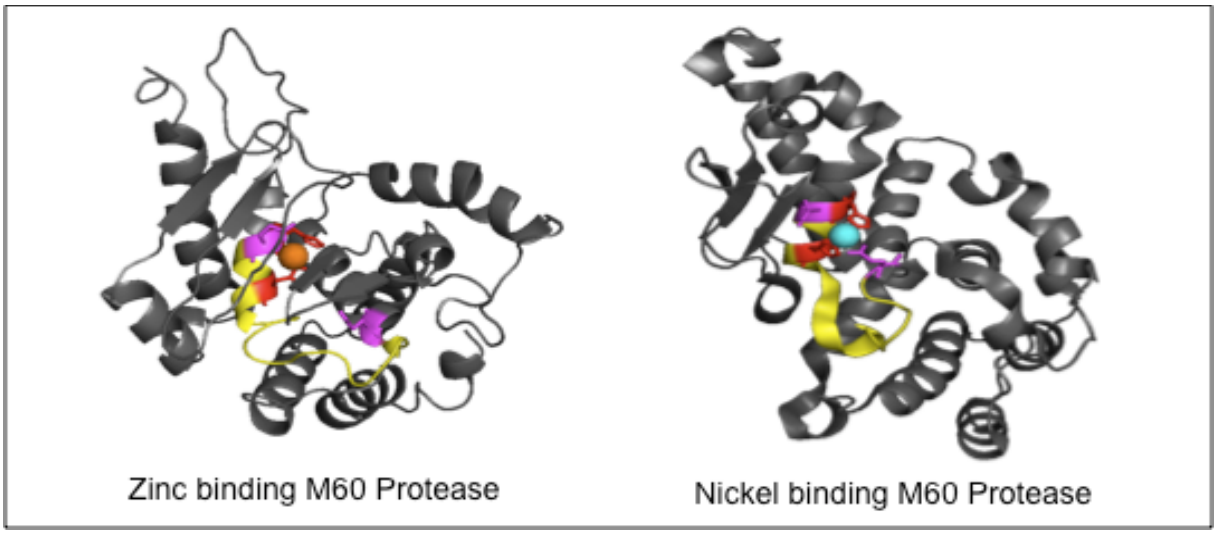

Figure 29: Three-dimensional protein models of the extracted M60-like pfam13402 domain of the zinc and nickel binding metalloprotease YghJ of *E. coli*. The protein models were predicted on basis of the protein POCK95 of *E. coli* K12 for zinc and chain A of protein 4FCA of *Bacillus anthracis* for nickel binding. The M60-like pfam13402 domain of *E. coli* is colored in grey and its conserved catalytic motif $HE_{XX}H_{X(13)}H$ in yellow. The histidin residues (H) are shown in red and the glutamate residues (E) by magenta sticks. The binding metal-ion is shown for Zn^{2+} by an orange and for Ni^{2+} by a blue sphere.

5. Discussion

5.1 Toxins of *E. coli* and *Shigella* spp.

Based on the hypothesis, that the transcription and translation of toxin gene variants leads to the expression of protein variants which show functional changes for example in their catalytic activity, receptor binding capacity or substrate activity, an *in silico* analysis of all toxin gene sequences expressed by *E. coli* and *Shigella* spp. published until June 2016 was required to identify a subset of allelic variants for each toxin present in those genomes. The focus of this thesis was to compile such a set of all known toxin genes on the general paradigm that toxins damage human and animal target cells through their own actions [59]. Therefore, 39 currently published toxins which are expressed by *E. coli* and *Shigella* spp., were collected by their nucleotide acid sequences and classified according to their mechanisms of interaction on target cells into three toxin groups: intracellular acting toxins (IATs), membrane damaging toxins (MDTs) and receptor targeted toxins (RTTs) (see Databases Table 3, pp. 23 ff.).

These 39 reference sequences were used to screen the NCBI and ECore databases to identify allelic variations for each toxin gene. In a range of 85 to 100% of nucleotide sequence identity in the respective alignment we assumed that these genes were variants of the gene of interest. Gene alignments with a lower sequence identity than a threshold of 85% to the reference cannot be uniquely assigned as an allele, since it could also encode for another gene. In this thesis, the present threshold of 85% nucleotide sequence identity was empirically determined, based on the results of the database screening. A general sequence identity threshold in order to define if nucleotide sequences are alleles or have only similarities in their nucleotide sequences cannot be determined for each gene in a same significance. It depends if a gene is highly conserved, embedded in a conserved region of genome or if a gene is embedded in a region with a high genetic plasticity see hot spots. For example, the haemolysin genes *ehxA*, *clyA* (*hlyE*) and *hlyA* were genetically highly conserved toxin genes showing in their alleles high sequence identities to their reference sequences; 97 to 100% for *ehxA* (see Appendix Figure A 12, p. xxxiii), 99,5% to 100% for *clyA* (*hlyE*) (see Figure 20, p. 40) and 96% to 100% for *hlyA* (see Appendix Figure A 13, p. xxxiv). Therefore, a threshold of 96% would be sufficient to find allelic variations of the haemolysin genes causing their low genetic distances. In this thesis, all collected allelic gene sequences of the mentioned haemolysin genes resulted finally in a single three-dimensional protein variant demonstrating that the few SNPs had no impact on the protein level. In contrast to the haemolysins, the most diverse toxin gene in this thesis was the Shiga-Toxin gene *stx2* showing for its allele *stx2f* a nucleotide sequence identities of 71% to the *stx2a* reference sequence (see Figure 19, p. 39). All the other analyzed toxin genes exhibited

allelic gene variants above the chosen threshold of 85% sequence identity and alignments being below the chosen threshold encoded for another gene or belonging to other species / organisms. Based on those facts, the chosen threshold encompasses all alleles of a toxin gene, regardless of whether their genetic structure is conserved or variable. In this context, a new classification of the *stx2f* allele, previous identified by Scheutz *et al*. 2012, as a third Shiga-Toxin subtype Stx3 can be discussed causing its high genetic differences to the reference sequence *stx2a*.

In this thesis, a systematic bioinformatics analysis unravels functional differences of allelic variations of all known toxins encoded in several *E. coli* and *Shigella* genomes has been performed. A first overview, which alleles of toxin genes can result in different protein variants due to transcription and translation, will be presented holding possible functional changes that might result in different toxicities on mammalian cells. A similar systematic approach can be further used to define functional changes of as yet unknown or unpublished toxins as well as for other virulence attributes. To determine the virulence of pathogenic strains, the interaction of all present virulence factors and their variants as well as their transcription and activation has to be considered [163]. This thesis gives a contribution on how to deal with large datasets, whereby bioinformatics tools are helpful to predict *in silico* functional changes of toxin variants and shows an overview of the genetic variability of toxins identified in the genome of *E. coli* and *Shigella* spp.

5.2 Genetic variability of toxins in the genome of *E. coli* and *Shigella* spp.

The bioinformatics results of this thesis have shown that all considered 39 toxin genes in the genome of *E. coli* and *Shigella* spp. (see Databases Table 3, pp. 23 ff.) did reveal allelic variations for each toxin for their gene and translated protein sequences (see Results for IATs Table 4, p. 35; for MDTs Table 5, p. 40; and for RTTs Table 6, p. 41) identified by screening the ECore and NCBI databases.

5.2.1 Intracellular acting toxins (IATs)

The group of intracellular acting toxins (IATs) has been demonstrated to represent toxins with a high genetic variability including the members of the serine protease autotransporters of *Enterobacteriaceae* (SPATEs), the Subtilase Cytotoxin SubAB*,* and the Shiga-Toxins Stx1 and Stx2 (see Results Table 4, p. 35). Therefore, the identified allelic gene variations of these toxins had also influence on their translated protein sequences as well as on their predicted three-dimensional protein models, confirming the hypothesis that genetic variations could reflect functional changes in their toxicity and therefore in a different pathogenic potential on mammalian cells.

In this thesis, all the collected nucleotide sequences for each of the twelve SPATE members expressed by *E. coli* and *Shigella* spp (see Results Figure 15, p. 36) resulted in a single three-dimensional protein model for each SPATE. These twelve SPATE variants were identified as twelve different gene alleles being distinguishable with respect to their three-dimensional protein structure into two already published functionally different protein variants reviewed by Ruiz-Perez and Nataro 2014. The cytotoxic class-1 SPATEs target small intracellular substrates whereas the lectin-like immunomodulator class-2 SPATEs bind mainly huge extracellular substrates due to the configuration of a domain being responsible for substrate access and closely linked to the protease domain [132]. In contrast to class-2 SPATEs (i.e. Hbp and SepA), the chitinase-like domain in the protein model for class-1 SPATEs (i.e. EspP) was completely missing in this thesis and can be used for classifying SPATEs encoded by *E. coli* and *Shigella* spp. in their functional group.

The in this thesis identified 25 *subAB* alleles, including the published *subAB* gene sequences from Funk *et al.* 2013 [44] and Nüesch-Inderbinnen *et al.* 2014 [111], revealed four three-dimensional protein variants confirming the cluster configuration according their genomic location published by Funk *et al.* 2013, including the SubAB variants $SubAB_1$, $SubAB_{2\text{-}1}$ and $SubAB_{2\text{-}2}$ [44]. Furthermore the *subAB* allele of the human *E. coli* isolate FHI42 published by Nüesch-Inderbinnen *et al.* 2014 [111] resulted in a completely independent three-dimensional protein variant $SubAB_{2\text{-}3}$ clustering with strain LM27564OEP published by Funk *et al.* 2013 [44]. The strain LM27564OEP clusters in the study by Funk *et al.* 2013 [44] despite its gene sequence differences as $subAB_{2\text{-}2}$ allele together with the $subAB_{2\text{-}1}$ alleles. The fact, that sequence differences in the *subAB* gene reveal different three-dimensional protein variants and associate with different genomic locations, emphasizes the presumption of Funk *et al.* 2013 that these allelic differences could reflect a different pathogenic potential and toxicity on host cells as was shown for the Shiga-Toxin variants [40, 70, 84]. Both, the Subtilase Cytotoxin and the Shiga-Toxin belong to the AB_5-toxin type as described in chapter 2.4 (pp. 9 ff.). The toxins consist of a catalytic A-subunit that induces cellular dysfunctions and a B-pentamer that recognizes host glycans. The A-subunit interacts over a α-helix named A2 via a hydrogen bond network to the B-pentamer [81, 95]. Previous studies described that the A-subunit of the Shiga-toxin variants display similar activities [56] and the B-subunit has been shown to display differences in receptor recognition by their glycolipid binding preference to immediate differences in cellular toxicity and host specificity [47, 56, 70, 85, 94, 100, 101]. In this thesis, the predicted three-dimensional protein structures of the four SubAB protein variants suggested similar enzyme activities. The differences in the A-subunits due to this analysis do not influence the protease activity whereas the varieties in the B-subunits could reflect differences in the receptor binding preference. Therefore, it might be possible that the receptor binding affinity of the B-subunits of the Subtilase Cytotoxin

variants on glycolipids is likewise specialized to different host cell surfaces like the B-subunits of the Shiga-toxin variants. Mutations in the SubB-subunits are mainly responsible for reduced or increased cytotoxic effects on target cells, due to binding of the holotoxin to the cell surface by the SubB-subunit as first interaction step to trigger subsequent intracellular pathways [81].

The in this thesis identified Shiga-Toxin gene sequences for *stx1* (nNA = 19) and *stx2* (nNA = 27) were consistent with the subtyping by Scheutz *et al.* 2012 [135], including the Stx1 protein variants Stx1a, Stx1c and Stx1d as well as the Stx2 protein variants, including the variants Stx2a to Stx2h. For the global cluster calculation in the study of Scheutz *et al.* 2012, 85 *stx1* and 311 *stx2* toxin sequences were analyzed by using the neighbor-joining method followed by the unweight pair group method (UPGMA) to create phylogenetic trees [135]; the same algorithm parameter was chosen for this thesis. Here, 97 *stx1* and 104 *stx2* gene sequences identified from the ECore and NCBI databases were analyzed to achieve the same clustering in the predefined Stx1 and Stx2 protein variants, confirming the results of Scheutz *et al.* 2012. [135]. Each protein variant resulted in a different three-dimensional protein model, confirming their functional differences [70, 40, 84].

5.2.2 Membrane damaging toxins (MDTs)

The pore-forming cyto- / hemolysins EhxA, ClyA (HlyE) and HlyA represented a group of highly structural conserved toxins. The gene alleles for *clyA* / *hlyE* share similarities in their gene sequences from 99.5% to100% identity (see Figure 20, p. 40), for *ehxA* from 97% to 100% identity (see Appendix Figure A 12, p. xxxiii) and for *hlyA* from 96% to 100% identity (see Appendix Figure A 13, p. xxxiv) referring to their reference gene sequences. All the allelic *clyA* / *hlyE*, *ehxA* and *hlyA* gene sequences resulted for each MDT in a specific single three-dimensional ClyA (HlyE), EhxA and HlyA protein model. Therefore, the identified allelic variations of each MDT had no influence on their translated protein sequences as well as on their predicted protein models and hence imply no functional changes in their toxicity on mammalian cells. That means, that the expression of one of these three MDT proteins by *E. coli* or *Shigella* spp. regardless their allelic gene variant pursue the same target to disrupt the host cell membrane.

For the toxin gene *clyA* / *hlyE* similar results were shown by A. Ludwig *et al*. 2004. The respective GenBank gene sequences were included in this thesis. These 79 *clyA* gene sequences corresponded to a sequence identity of >99% whether encoded by the same *E. coli* pathogroup or even by different pathogroups. The predicted amino acid sequences for *clyA* were either identical or contained between one and three amino acid exchanges [89]. In this thesis, 266 *clyA* / *hlyE* gene sequences of different *E. coli* pathogroups were analyzed

leading to the same result. Additionally, it was shown that all these selected gene sequences resulted in one single protein model.

Pore-forming toxins (PFTs) are used by several bacterial species to disrupt the membrane of target cells. PFTs are produced as non-membrane-bound, soluble monomers, which assemble into the host cell membrane to an oligomeric pore complex [131]. The genetic organization of ClyA / HlyE and EhxA is comparable to HlyA in the genome of *E. coli* and *Shigella* spp. [86]. Such a simple and effective mechanism to obtain nutrients from target cells plays an important role for bacteria and can explain why those genes are highly conserved without any structurally changes.

In a study which aimed to identify conserved genomic regions in the genome of *E. coli* O157:H7 lineages, twelve conserved regions were found. In addition to a number of identified regulatory genes and genes having a potential to act as virulence factors by regulating expression of effector genes directly involved in pathogenesis, hemolysins were also present within the conserved regions [147]. During cluster identification the conserved regions have been identified in *E. coli* lineages across different pathotypes, confirming also the conserved genetic structure of the three analyzed pore-forming cyto- / hemolysins.

5.2.3 Receptor targeted toxins (RTTs)

The group of RTTs represented a contrary group of both highly structural conserved toxins on the one hand and highly structural diverse toxins on the other hand. This group includes heat-labile and heat-stable enterotoxins (LT, EAST1, STa and STb), type III effector proteins (Cif, EspF, EspH, IpaB, IpgD, Map, Tir and VirA), cytotoxins (CdtVa-c, Cnf1 and Cnf2), toxins encoded within the colibactin locus (ClbA-Q), the dynamin like-protein LeoA, the lymphocyte inhibitory factor LifA (Efa1), *Shigella* enterotoxins (ShET1 and ShET2) and the metalloprotease YghJ.

Three toxins showed variants of their predicted three-dimensional protein structure and therefore indicate the possibility of functional changes. The toxins with identified variants of their three-dimensional protein models was the effector proteins EspF and EspH as well as the metalloprotease YghJ. In the end of the bioinformatics analyses, all the other toxins within the group of RTTs resulted in a single three-dimensional protein structure without any changes of their original function. It is possible that these enterotoxins, cytotoxins and effector proteins are embedded in conserved genomic regions of the chromosome as already discussed for the membrane damaging toxins. A number of potential virulence factors were identified within these regions, including iron transport systems and several regulatory genes [147].

The type III effector protein EspF resulted in this thesis into two protein variants (see Figure 21, p. 43) showing only differences in the number of their consecutive

homologues proline-rich repeats at the 3' end of their protein structure. Due to the proline-rich repeats, a short version of EspF for EHEC strains (Acc. No. AE005174) and a long version of EspF for EPEC strains (Acc. No. FM180568) was identified. The repeats were identical in number (47 AA) and their sequence similarity had no influence on the resulting protein function. The N-terminal region, containing the secretion signal is sufficient for the EspF secretion and translocation into the host cell, constituting a conserved region with phylogenetic differences between 98 to 100 % identity without functional changes in the expressed protein. Therefore, both EspF protein variants exhibit highly similar functions. Similar results were shown and discussed for EspF by Holmes *et al*. 2010. Holmes *et al.* show that the modular structure, the multiple function by infecting host cells and the acting site on different locations in the cells is not influenced by their number of proline-rich repeats. EspF is known to be one of the most multifunctional effector proteins to playing a role in several host cellular processes, including disruption of the epithelial barrier, modulation of the cytoskeleton and disruption of the nucleus up to apoptosis. Such a multifunctional effectiveness of this relatively small effector protein can explain its high genetic stability. For the EspF protein variant from *Citrobacter rodentium* functional changes are known showing mutations with up to 67% gene sequence difference in the N-terminal region between the 21st and the 74th AA. These AA changes alter the nucleus-disruption behavior of EspF in *Citrobacter rodentium*. A similar functional change of EspF could be observed for EPEC and EHEC strains if mutations in the N-terminal region occur. EspF from the rabbit-specific EPEC strain REDEC-1 with a gene sequence difference in the N-terminal region of 73% compared to the EPEC strain E2348/69 (Acc. No. FM180568) and only two proline-rich repeats could be such a candidate having functional differences [62]. Those short and genetically different *espF* gene sequences were excluded from this thesis on basis of the threshold of 85%. In this context, it can be discussed that the protein variant of the REDEC-1 strain, which was declared as EspF effector protein by Holmes *et al*. 2010 [62], could be a new effector protein.

In this thesis the analyses of the type III effector protein EspH has resulted in three distinct three-dimensional protein variants being summarized in Figure 22, p. 44. Common to all three protein variants was the combination of four α-helices that surround four parallel β-strands for EPEC strain E2348/69 (Acc. No. FM180568), three parallel β-strands for EHEC strain IHIT1190 (Acc. No. FM986651) and three β-strands arranged like a curve for EHEC strain 11128 (Acc. No. AP010960), whereby the spatial arrangements among the α-helices differ (see Figure 26, p. 53). This thesis presents for the first time how a protein model prediction of EspH can be realized. It is known that EspH disrupts the actin cytoskeletal structure in the host cell by modulating the assembly kinetics and morphology of actin pedestals by inhibiting a mammalian GTPase of the small Rho GTPase family, which are master regulators of actin cytoskeleton rearrangements. EPEC and EHEC strains are able to

inactivate the Ras homolog GTPase (Rho) of the host cell by expressing EspH. EspH binds directly to the conserved PH-DH domain of the cellular Rho Guanine Nucleotide Exchange Factors (RhoGEFs) and blocks therefore the Rho GTPase activation of the host cell inducing cytotoxic effects. At the same time bacterial RhoGEFs that are insensitive to EspH stimulate the GTPase activation again by mimicking mammalian RhoGEFs to allow cell adhesion [28, 146, 166]. Furthermore, J. Cherfils *et al.* 2013 reviewed that an effector protein from *Shigella* (IpgB2) harboring to EspH comparable spatial arrangement in its three-dimensional protein structure, representing three parallel ß-strads surrounded by six a-helices, also binds to the PH-DH domain of the cellular RhoGEFs to block the mammalian Rho GTPase activation [18]. With regard to the structural difference of IpgB2 in comparison to EspH, which binds the same RhoGEFs over the PH-DH domain, it may be possible that all three allelic variations of EspH are able to bind the mammalian RhoGEFs. In this thesis, such an EspH - RhoGEF complex was confirmed using structural modeling for all three EspH variations (see Figure 27, p. 55). X. Tu *et al.* 2003 showed clear differences between the morphology of the actin pedestals induced by EPEC and EHEC. EspH expressed by EHEC induces the formation of flat and faint actin structures, whereby EspH expressed by EPEC forms elongated pedestals and containes a high concentration of actin [159]. On basis of the differences in their three-dimensional arrangement, predicted in this thesis, it may be possible that the EspH protein variants of EHEC bind less efficient to the PH-DH domain of the RhoGEFs than the EspH protein variant of EPEC (see Figure 27, p. 55). These spatial differences give a possible answer for the formation of different actin pedestals of EHEC and EPEC strains. It remains to be proven if the differences in the functional mechanisms of the EspH variants from EHEC and EPEC contribute to the variances in pedestal formation or how the expression of the different EspH variants affect the mechanism which is used by EPEC and EHEC to induce the formation of pedestals.

The bioinformatics results of the metalloprotease YghJ, including 185 *yghJ* gene sequences, presented with 25 alleles a high number of allelic gene variants with gene sequence identities of 85% to 100% compared to the reference sequence. The allelic variants can be summarized at the end of the analysis in seven structurally different three-dimensional protein models (see Figure 28, p. 56). This thesis gives a first insight as to how a predicted three-dimensional protein model of YghJ can be realized. The prediction of these metalloprotease proteins is complex corresponding with the modeling of multi-domain proteins. Such a complex protein modeling is based on several already known protein domains and is therefore only a prediction of the whole protein model. One of these domains, included in all seven protein variants, is highly similar to a protein having an already known three-dimensional structure and function: the M60-like pfam13402 domain. Proteins containing this domain have the metalloprotease function to degrade host glycoproteins and

bind metal-ions. Furthermore, all seven identified YghJ protein variations exhibited the conserved sequence motif HEVGH in their M60-like pfam13402 domain, which belongs to the HExxH motif, and is known to be involved in the catalytic activity of metalloproteases of different bacterial species [102]. In addition to their common features, the seven predicted protein variants differed in their protein folding compactness and in their metal binding preference. The M60 like pfam13402 domain of YghJ 3D variant 1 (UPEC strain UTI89, Acc. No. CP000243), variant 2 (*E. albertii* strain NIAH_Bird_3, Acc. No. AP014855), variant 6 (NMEC strain CE10, Acc. No. CP003034) and variant 7 (EAEC strain 55989, Acc. No. CU928145) were consistent with the nickel binding M60 peptidase 4FCA (chain A) of *Bacillus anthracis* in a three-dimensional alignment. In this thesis, such a nickel binding M60 like pfam13402 domain for *E. coli,* including a nickel binding catalytic motif was predicted to illustrate the probability of YghJ-metalloproteases being active in the presence of nickel. But it may also be possible that this M60 like pfam13402 domain is active in the presence of zinc, because the catalytic motif of the M60 like pfam13402 domain of *E. coli* K12 strain (POCK95) prefers to bind zinc only over two histidine residues, which were also present in the nickel binding size of the M60 peptidase 4FCA (chain A) (see Figure 29, p. 58). These structural affinities of the zinc binding M60 like pfam13402 domain POCK95 of the *E. coli* K12 strain was exhibited by the YghJ 3D variant 3 (*E. coli* strain BL21(DE3), Acc. No. CP001509), variant 4 (EPEC strain CB9615, Acc. No. CP001846) and variant 5 (*E. coli* K12 sub-strain W3110, Acc. No. AP009048). With regard to the whole YghJ protein, the three-dimensional structure of the YghJ 3D variant 3 and 4 breaked up in separate bundles in contrast to the other compact YghJ models. Such a separated protein folding can lead to a nonfunctional metalloprotease. The YghJ 3D variant 5 showed the highest protein folding compactness of all the selected YghJ variants (see Figure 28, p.56).

The described differences in the three-dimensional protein structure of the YghJ variants indicate functional changes in their metal binding preference. It may be possible that the catalytic activity of the zinc and nickel binding variants is diverse and it may also be possible that the in separate bundles folded three-dimensional protein variants show no catalytic effects. It is known that zinc has been substituted by several divalent cations in numerous metalloproteases. Metalloproteases like YghJ are able to bind different cations with respect to zinc, causing a comparable, lower or completely inactive catalytic activity [41].

5.3 Influence of bioinformatics methods on the results

The sequence search to identify homologues nucleotide acid or amino acid sequences in bioinformatics databases often provides the first indications about the function and therefore about functional changes due to mutations. There are a number of software tools for screening databases in order to find homologues sequences, which share a common

evolutionary ancestor. Around 80% of genomic sequence samples share significant similarity with genes or proteins in sequence databases like NCBI-BLAST [118]. In the following chapter all bioinformatics methods, which were used in this thesis, will be discussed with regard to their accuracy and their interpretability on the results.

5.3.1 Whole genome sequencing

The basis for searching sequences in bioinformatics databases is the genetic information of single genes or rather the whole genomes including mobile genetic elements such as plasmids and prophages. The quality of the assembled transcripts during sequencing depends on biological and sequencing technics. After fragmentation of the complete DNA single fragments undergo several amplification cycles in the Polymerase-chain-reaction (PCR), which might cause errors due to the sequencing steps. Further errors can be accumulated during sequencing in the steps of library preparation and afterwards during the execution of assembly algorithms. To minimalized sequencing errors from true variants, each base or each coding region respectively is read more than one time. Comparing the alignment after sequencing to a reference genome, mismatches, deletions or insertions can be corrected but also inserted. Further, the comparison of *E. coli* whole genomes to a reference sequence is challenging due to the highly variable gene content among *E. coli* strains. Therefore the identification of single nucleotide variations can be complicated by various error sources [25], [61], [96].

5.3.2 Algorithm for alignment and phylogenetic tree construction

To identify similar sequences to the reference gene sequences, the basic local alignment search tool (BLASTn) was used. In contrast, to create a pairwise alignment, the global alignment was used in this thesis. The basic differences between a local and global alignment is that in a local alignment the query sequence match as a small continuous portion of the subject sequence. Whereas in a global alignment the match of the query sequence perform an end to end alignment with the subject. Therefore, if the size of the query sequence is dissimilar to the subject, a lot of gaps in the global alignment to stretch the query sequence on the size of the subject sequence will be inserted. The differences between these two alignment types is explained by their scoring matrices. In the global algorithm, the highest scoring correspond to the coordinates of the whole subject sequence, whereas in the local algorithm a single element with the highest score is done [3, 108, 144]. Global alignments are useful for comparing long sequences including more than one gene, whereas local alignments can be used to find homologous gens or domains in other sequences. For screening databases against a reference gene sequence, the local alignment method is meaningful. Comparing sequences to each other to have a look at polymorphisms, a global alignment is meaningful including gaps in missing parts in the query sequence.

In this thesis, to build a phylogenetic tree as genetic distance model for nucleotide acid sequences the Tamura Nei model was used and for amino acid sequences the Jukes Cantor model was used, both in combination with the tree build method UPGMA. The Tamura Nei and the Jukes Cantor model are both one of the most common genetic distanced models. All these models described the evolution of a single site within a set of sequences as substitution models. The differences between al substitution models is the equilibrium of the base frequency causing their different substitution matrices. The Tamura Nei model based on the assumption, that every base has a different equilibrium base frequency π ($\pi_A \neq \pi_G \neq \pi_C \neq \pi_T$) distinguishing different weighting between the two transition types ($A \leftrightarrow G \neq C \leftrightarrow T$) as well as between transitions and transversions ($A \leftrightarrow C$, $G \leftrightarrow T$ and $A \leftrightarrow T$). It is assumed that transversions occur at the same rate and assign an equal weighting, however different to both transition types [149]. A more complex substitution model is the Generalised time-reversible (GTR) model. This method includes besides different substitution rate parameters even described as well for the Tamura Nei model also equilibrium base frequency parameters, giving the frequency at each base on each site [152]. All the other substitution models like the Jukes Cantor model [66], the Kimura model [73], the Felsenstein model [34] or the HKY [55] model observing fewer weighting factors than the Tamura Nei model, assuming equal base frequencies and / or equal mutation rates. Therefore, the Tamura Nei model is the most complex genetic distanced model for nucleotide acid sequences which can be used in Geneious and combined directly with the tree build method UPGMA. The different weighting between mutations with the Tamura Nei model is sufficient for the bioinformatics analysis using the subsequent phylogenetic tree construction solely for structural cluster building to identify allelic gene sequences. The same argumentation explains the usage of the simplest genetic distance model the Jukes Cantor model for weighting mutations between amino acid sequences assuming equal base frequencies [66]. The more complex protein sequence structure in combination with this simple substitution matrix provide in the subsequent phylogenetic tree construction an optimal cluster building to identify different protein sequence variants.

To identify gene or protein sequence variants classified by their polymorphisms a phylogenetic tree based on genetic distances will be needed. In this thesis, a rooted phylogenetic tree was chosen to identify variants via cluster building referring to the reference sequence. Therefore, the phylogenetic tree construction was only used to arrange and classify gene or protein sequence variants. Such a rooted tree resulting from the data of a distance matrix can be built with the simple and fast agglomerative hierarchical clustering method UPGMA. The UPGMA algorithm join the closest sequences in a cluster, starting with the first two sequences, combining the nearest sequences at each clustering step [98]. Unrooted trees which can be built via the neighbor-joining method uses also distance data,

clustering sequences with the smallest amount together as neighbors. However, the tree construction can be compare with a star without a starting point like a reference sequence [133]. Therefore, the reference sequence clusters together with its nearest neighbors in the star formation and complicates a graphic analysis.

5.3.3 Prediction of secondary structures

In this thesis, the GOR V algorithm (see chapter 3.2.4, p. 30) in Geneious software was used to predict the secondary protein structure of all the amino acid sequence variants. Geneious accesses a number of public databases hosted by NCBI (http://www.ncbi.nlm.nih.gov) and UniProt (http://www.uniprot.org). To predict the secondary protein structure from the stand alone amino acid sequence many attempts will be needed [48]. Therefore, a comparison of amino acid sequences with already known secondary structures is needed to ensure the accuracy of the predicted model. Only structural information gives an insight into protein function and is also useful for motif detection as well as homology modeling. Therefore, a high-accuracy prediction of protein structures from its amino acid sequence is highly desirable.

The fifth version of the GOR method has an accuracy prediction of Q_3 by 73,5%. Based on different approaches including information theory, Bayesian statistics and evolutionary information, GOR V has been the most successful method. The GOR V database is based on 513 sequentially non-redundant domains, which contain 84.107 amino acid residues to ensure a highly representative set of available proteins calculating helices, strands, coils and turns that have the highest probability at each residue position of the input sequence by evolutionary information using PSI-BLAST. Homologues sequence information through multiple sequence alignments give a significant boost, adding triplet statistics to the accuracy and optimize various parameters.

GOR V is still a secondary structure prediction method providing a consensus method with a high accuracy prediction like other methods. Other methods such as PhD and PSIPRED reported an accuracy of about 76% and PREDATOR reached an accuracy of about 77%. The GOR V method is around 5% less accurate than the other methods, but it provides complementary information because it is based on different approaches like the information theory and Bayesian statistics in contrast to the other neural network based methods [75, 141]. Therefore, all the secondary structure prediction methods result in protein structure predictions that represent the actual protein structure only in a limited resemblance to its original biological structure. However, the GOR V method benefits because of its underlying fundamentally principles. Nevertheless, the resulting secondary protein structures allowed a first statement in predicting the effect of mutations and understanding the function as well as

functional differences of variations. In this thesis, the secondary protein structures of allelic variations of a toxin give a first hint of probably resulting differences.

5.3.4 Prediction of three-dimensional structures

To predict the three-dimensional protein structures of all amino acid variants of interest in this thesis, the open source web server I-Tasser was used. I-Tasser is based on simulating the folding process using only the amino acid sequences itself as input data, predicting the structure of proteins on basis of homologs of known predicted or experimentally determined three-dimensional structures, the template-based homology modeling or fold-recognition. The target sequences are first threaded through a representative PDB structure library with a pair-wise sequence identity cut-off of 70% to identify possible folds. Multiple domain protein and domain boundaries are automatically assigned [174]. Well above six million unique protein sequences have been deposited at time in the public databases hosted by NCBI (http://www.ncbi.nlm.nih.gov) and UniProt (http://www.uniprot.org).

The practical applications of protein structure prediction can be used to develop functional hypotheses about hypothetical proteins, single domains or possible functional differences during mutations. With the advent of large sequence databases and highly accurate profile-profile matching algorithms, it may be possible to model three-dimensional protein structures with less than 20% sequence identity to known protein structures. However, *ab initio* techniques that have been developed to predict protein folding are not error-free. All tools can be used in a fully automated way and the results must be analyzed in context of their biological knowledge. Long protein sequences often contain multiple domains, which have to be considered individually in the remaining protein sequence / model to ensure that these domains are modeled as a single functional region and give a biological sense without any other mixed structures [71 ,173]. On this occasion it is useful to predict domains, extracted from their whole protein sequence and compare the results to the whole protein model, as was shown in this thesis for the type three-effector protein EspH and the metalloprotease YghJ. Domains extracted from the query sequence can be individually identified using NCBI, PDB or PFAM.

In this thesis, many selected alleles of a single toxin result in the same three-dimensional protein model. In this case, a potentially needed reference model, modeled by software can lead to problems, because single changes of amino acids could not be sufficiently incorporated in the three-dimensional model. Point mutations in the query protein structure will not in general result in a different three-dimensional protein model, which causes inherent limitations in the template-based prediction algorithms merged by such subtle changes. Many three-dimensional structure predicting applications use loop modeling

techniques to model small insertions and repair deletions in the alignment to the template [71].

After comparison with the most known freely available structure prediction systems on the Web, I-Tasser was chosen, as it provides an optimal performance between a target size range of 10 to 1.500 residues in contrast to Phyre, which optimal performance is less than 1.000 residues. Combining predictions from many sources is the most reliable way to avoid erroneous results and to determine the most accurate protein model. *Ab initio* techniques can provide valuable hints to structural features in the absence of time intensive and expensive experimental derived structures. The community-wide Critical Assessment of Structure Prediction (CASP) experiments have been designed I-Tasser as one of the best method in structure prediction systems. I-Tasser has shown to be able to generate 75% of the sequences tested in a blind experiment by CASP. More than 80% of the templates were identified close to their native structures [71, 162, 174]. Nevertheless, functional analysis will be needed to confirm and expand the bioinformatics results in biological systems.

6. Conclusion

For infectious diseases with bacterial pathogens, early diagnosis is the key in risk assessment and thus disease management. Next generation sequencing (NGS) allows sequencing of a huge number of bacterial genomes. The whole genome sequence (WGS) provides all the genetic information allowing bacterial pathogens the expression of genes like toxin genes to infect the host and possess infectious diseases. Furthermore, genome sequencing is less labor-intensive, requires less time and identifies the bacteria including all its genomic features like drug-resistance, virulence and toxin genes to significantly influence the treatment outcome. Additionally, gene variants and genome rearrangements can be detected to gain conclusions to clinical and immunological phenotypes.

This thesis was focused on toxin gene profiling of a large set of *E. coli* strains from the ECore database, which was enhanced at the Institute of Microbiology and Epizootics as well as the open source database NCBI. One aim of this work was to create a reference genome called “toxome”, which includes all currently known toxin genes. The “toxome” of *E. coli* and *Shigella* spp. allows screening of all currently known toxin genes and its variants with the power of multiplexing in the whole genome sequences of the mentioned bacteria using a simple BLASTn algorithm. Therefore, the “toxome“ can be used as reference to assemble all toxin sequences of one *E. coli* strain after sequencing and achieve full-length hits of their encoded toxin genes.

The potentially observed occurrences of toxin genes in various *E. coli* and *Shigella* spp. strains as well as variations and combinations in the genome may lead to a better understanding of the toxin complex of these bacteria. Functional differences between protein variants within a toxin have not been thoroughly researched. This thesis gives a first overview about the occurrence and impact of genetic differences in toxin genes in the genome of *E. coli* and *Shigella* spp. regarding their resulted protein variants on basis of bioinformatics analysis.

The results showed that genetic variations of all the analyzed 39 toxin genes (listed in Table 3, pp. 23 ff.) were available after screening 492 whole genome sequences including the internal ECore database and 86 whole genome sequences, 72 plasmids, 22 PAIs and 276 genes including the NCBI database. The hypothesis of this thesis that allelic variants of toxins differ in their biological effect regarding their toxicity was confirmed *in silico* for the effector protein EspH, the metalloprotease YghJ, the Subtilase Cytotoxin SubAB, the serine protease autotransporter of *Enterobacteriaceae* (SPATEs) as well as for the Shiga-Toxins Stx1 and Stx2. The three-dimensional structure of protein variants and the possibility of resulting functional changes of the effector protein EspH and the metalloprotease YghJ were analyzed for the first time.

Based on the differences in the three-dimensional arrangement of the EspH variants, it may be possible that the protein variants bind different efficient to the conserved PH-DH domain of the mammalian Rho Guanine Nucleotide Exchange Factor (RhoGEFs). Therefore, an efficient binding blocks the Ras homolog GTPase (Rho) activation, which is a master regulator of the actin cytoskeleton rearrangement of the host cell, inducing cytotoxic effects. Bacterial RhoGEFs are insensitive to EspH stimulating the GTPase activation again to allow cell adhesion. The different efficient binding to the PH-DH domain putatively results in the formation of different actin pedestals as was shown in a previous cell-culture based study by X. Tu *et al.* 2003, analyzing EspH from EPEC and EHEC strains [159].

The three-dimensional differences in the protein structure of the YghJ variants indicated functional changes in their metal binding preference. A part of the YghJ protein variants confirmed in their catalytic M60 like pfam13402 domain with a nickel binding M60 peptidase 4FCA (chain A) of *Bacillus antrhacis* binding nickel by two histidine and one glutamate residues. The other part confirmed to a zinc binding M60 peptidase of *E. coli* K12 (POCK95) binding zinc only over two histidine residues. Metalloproteases are able to bind different cations causing a comparable, lower or completely inactive catalytic activity in respect to zinc [41]. Therefore, it may be possible that the catalytic activity of the YghJ protein variants was diverse. Furthermore, the YghJ protein variants differed in their protein folding compactness. Protein variants which breaks up in separate bundles in contrast to the compact models can lead to a nonfunctional metalloprotease.

Different variants of the Subtilase Cytotoxin SubAB are already known on basis of their gene level [44, 111]. In this thesis, the predicted three-dimensional protein variants of the Subtilase Cytotoxin SubAB were first analyzed with regard to a different pathogenic potential and toxicity. The varieties in the B-subunits which bind to the cell surface could be responsible for reduced or increased cytotoxic effects on target cells and host cell specificities. It is possible that the B-subunit variants display differences in receptor recognition by their glycolipid binding preference as was shown for the B-subunits of the Shiga-Toxin variants [70].

Furthermore, the twelve already known and different three-dimensional protein variants of the SPATEs indicated that the SPATEs members are already divided into allelic variations and split into two known functionally different classes demonstrating functional changes in substrate binding caused by phylogenetic differences. The cytotoxic class-1 SPATEs harbored a compact disulfide bond-containing domain (helix-turn-helix motif) which is able to bind only small intracellular substrates. In contrast, this domain was decreased by the lecithin-like immunomodulators class-2 SPATEs and enabls the binding of huge extracellular substrates like mucins [132].

The allelic variations of all the other analyzed toxins expressed by *E. coli* and *Shigella* spp. due to *in silico* analyses had no impact on the resulting three-dimensional protein structure, indicating no functional changes as in the case of EspF. The EspF variants showed only differences in the number of their consecutive homologues proline-rich repeats at the 3' end of their protein structure which does not affect the resulting protein function despite the identified phylogenetic differences.

The inclusion of toxins with already known functional changes of their allelic protein variants like the different Shiga-Toxin variants [40, 70, 84, 135] have revealed the same results, indicating that the chosen bioinformatics methods in this thesis obtained comparable results to other bioinformatics studies with clear hints toward functional changes of the protein. Such a systematic bioinformatics approach can be further use for other virulence factors being involved in the pathogenicity of *E. coli* and *Shigella* spp. strains. Furthermore, this bioinformatics approach can be use as manual, which tools are useful to predict *in silico* functional changes of toxin variants.

The next step with regard to the results of this thesis could be subsequent functional analyses of the identified different protein variants to evaluate their resulting toxicity on mammalian cells or with lactate dehydrogenase release (LDH) assays to indirectly evaluate cell death [130]. The characterization of the toxigenic potential of a toxin variant can be realized by using cell cytotoxicity assays based on a subsequent microscopic assessment of toxin-induced cell damage. Especially for the intestinal mucin degrading metalloprotease YghJ, the intestinal epithelial cell-lineage Caco-2 is useful. The cells can be infected with a bacteria-free supernatant of transformed bacterial cells harboring a plasmid/vector with an expression cassette encoding the toxin variant of interest. The transformed bacterial cells can be also used for protein expression of the target protein to obtain purified proteins via chromatographic methods to analyze the toxigenic potential of the toxin protein only [90, 148] and for further mass spectrometry or calorimetry based analyses. For example, metal substitution assays can be performed with respect to identify the metal binding preferences measuring the absorption spectra [50] or measuring the amount of heat during the binding reaction [39].

7. Danksagung

Mein besonderer Dank gilt Herrn Prof. Dr. Lothar H. Wieler, Präsident des Robert-Koch-Instituts (RKI), für die Möglichkeit meine Dissertation am Institut für Mikrobiologie und Tierseuchen der Freien Universität Berlin anfertigen zu dürfen. Ebenso bedanke ich mich für die Betreuung und die Begutachtung meiner Arbeit sowie die fachlichen Diskussionen und Ideen.

Bedanken möchte ich mich ebenfalls bei Frau Prof. Dr. Haike Antelmann (Institut für Biologie der FU-Berlin) für die so kurzfristige Begutachtung meiner Arbeit als Zweitgutachterin sowie Herrn Dr. Torsten Semmler (RKI) für die Unterstützung im Bereich der Bioinformatik.

Die Teilnahme als assoziierte Doktorandin im Graduiertenkolleg (GRK 1673) der DFG in der Indo German Research Training Group „Functional Molecular Infection Epidemiology“ ermöglichte mir im Februar 2015 den Besuch der Winter School an der Universität in Hyderabad, Indien. Ein großes Dankeschön für die freundliche Betreuung und Kooperation vor Ort geht an Frau Prof. Lalitha Guruprasad (University of Hyderabad) und Herrn Dr. Kunchur Guruprasad (Center for Cellular and Molecular Biology - CCMB).

Ein herzliches Dankeschön geht an Frau Dr. Inga Eichhorn für die zahlreichen Hilfestellungen, fachlichen Fragen und die exzellente wissenschaftliche Betreuung am Institut. Liebe Inga, deine fachliche sowie private Unterstützung haben einen großen Anteil zum Gelingen dieser Arbeit beigetragen.

Weiterhin danke ich dem gesamten Team des Instituts für Mikrobiologie und Tierseuchen für die immer freundliche Zusammenarbeit und das ich so viel Zusätzliches im Bereich der Veterinärmedizin lernen durfte. Liebe Katharina, Lisa und Inga, es war eine besonders schöne Zeit mit euch zusammen im Büro zu Arbeiten.

Ebenso möchte ich mich an dieser Stelle herzlich bei meinen Eltern, meinem Bruder und meinem Freund für die große Hilfe bedanken. Ohne eure Unterstützung wäre diese Arbeit nicht möglich gewesen. Danke, dass ihr mich in dieser Phase meines Lebens immer bestärkt und an mich geglaubt habt.

Gewidmet ist diese Dissertationsschrift in Gedenken an meinen besten Freund.

8. Curriculum Vitae

For reasons of data protection,

the curriculum vitae is not included in the online version.

9. Erklärung

Hiermit erkläre ich, dass:

1. die vorliegende Dissertation selbstständig verfasst wurde,
2. die vorliegende Dissertation nur unter Verwendung der angegebenen Literatur und Hilfsmittel erstellt wurde,
3. wörtlich und sinngemäß aus anderen Quellen übernommene Teile der Arbeit deutlich als Zitat gekennzeichnet und mit Quellenangaben versehen sind.

Berlin, den 08.01.2019

Susanne Fleischmann

10. References

1. **Ahmed, N., Dobrindt, U., Hacker, J., & Hasnain, S. E.** 2008. Genomic fluidity and pathogenic bacteria: applications in diagnostics, epidemiology and intervention. Nat Rev Microbiol, **6**(5), 387-394.
2. **Alouf, J. E.** 2006. in *The Comprehensive Sourcebook of Bacterial Protein toxins* (eds Alouf, J. E. & Popoff, M. R.) Chapter 1: A 166-year story of bacterial protein toxins (1888-2004): from "diphteritic poison" to molecular toxinology (Vol. **Thd. Ed.**). Londin: Academic Press. 3-21.
3. **Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J.** 1990. Basic local alignment search tool. J Mol Biol, **215**(3), 403-410.
4. **Andrade, F. B., Abreu, A. G., Nunes, K. O., Gomes, T. A. T., Piazza, R. M. F., & Elias, W. P.** 2017. Distribution of serine protease autotransporters of *Enterobacteriaceae* in typical and atypical enteroaggregative *Escherichia coli*. Infect Genet Evol, **50**, 83-86.
5. **Asakura, H., Makino, S., Kobori, H., Watarai, M., Shirahata, T., Ikeda, T., & Takeshi, K.** 2001. Phylogenetic diversity and similarity of active sites of Shiga toxin (stx) in Shiga toxin-producing *Escherichia coli* (STEC) isolates from humans and animals. Epidemiol Infect, **127**(1), 27-36.
6. **Bai, X., Fu, S., Zhang, J., Fan, R., Xu, Y., Sun, H., . . . Xiong, Y.** 2018. Identification and pathogenomic analysis of an *Escherichia coli* strain producing a novel Shiga toxin 2 subtype. Sci Rep, **8**(1), 6756.
7. **Baker, D., & Sali, A.** 2001. Protein structure prediction and structural genomics. Science, **294**(5540), 93-96.
8. **Bardhan, P., Faruque, A. S., Naheed, A., & Sack, D. A.** 2010. Decrease in shigellosis-related deaths without *Shigella* spp.-specific interventions, Asia. Emerg Infect Dis, **16**(11), 1718-1723.
9. **Behrens, M., Sheikh, J., & Nataro, J. P.** 2002. Regulation of the overlapping pic/set locus in *Shigella flexneri* and enteroaggregative *Escherichia coli*. Infect Immun, **70**(6), 2915-2925.
10. **Belousov, M. V., Bondarev, S. A., Kosolapova, A. O., Antonets, K. S., Sulatskaya, A. I., Sulatsky, M. I., . . . Nizhnikov, A. A.** 2018. M60-like metalloprotease domain of the Escherichia coli YghJ protein forms amyloid fibrils. PLoS One, **13**(1), e0191317.
11. **Betts, M. J., & Russel, R. B.** 2008. in *Bioinformatics for Geneticists: A bioinformatics primer for the analysis of genetic data* (eds. M. R. Barnes) Chapter13: Amino-Acid Properties and Consequences of Substitutions (Vol. **Sec. Ed.**). West Sussex: John Wiley & Sons. 311-342.
12. **Bezine, E., Vignard, J., & Mirey, G.** 2014. The cytolethal distending toxin effects on Mammalian cells: a DNA damage perspective. Cells, **3**(2), 592-615.
13. **Bielaszewska, M., Middendorf, B., Kock, R., Friedrich, A. W., Fruth, A., Karch, H., . . . Mellmann, A.** 2008. Shiga toxin-negative attaching and effacing *Escherichia coli*: distinct clinical associations with bacterial phylogeny and virulence traits and inferred in-host pathogen evolution. [Research Support, Non-U.S. Gov't]. Clin Infect Dis, **47**(2), 208-217.
14. **Biomatter Ltd.** (2015, May 22, 2015). Geneious 7.1 Manual, from http://assets.geneious.com/documentation/geneious/GeneiousManual7.1.pdf
15. **Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., . . . Lander, E. S.** 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet, **22**(3), 231-238.
16. **Casjens, S.** 2003. Prophages and bacterial genomics: what have we learned so far? Mol Microbiol, **49**(2), 277-300.
17. **Cassady-Cain, R. L., Blackburn, E. A., Alsarraf, H., Dedic, E., Bease, A. G., Bottcher, B., . . . Stevens, M. P.** 2016. Biophysical Characterization and Activity of Lymphostatin, a Multifunctional Virulence Factor of Attaching and Effacing *Escherichia coli*. J Biol Chem, **291**(11), 5803-5816.

18. **Cherfils, J., & Zeghouf, M.** 2013. Regulation of small GTPases by GEFs, GAPs, and GDIs. Physiol Rev, **93**(1), 269-309.
19. **Croxen, M. A., Law, R. J., Scholz, R., Keeney, K. M., Wlodarska, M., & Finlay, B. B.** 2013. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev, **26**(4), 822-880.
20. **Cuff, J. A., & Barton, G. J.** 1999. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins, **34**(4), 508-519.
21. **Cuff, J. A., & Barton, G. J.** 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, **40**(3), 502-511.
22. **David, A.** 2004. in Handbook of Proteolytic Enzyme (eds. Barret, A. J., Rawlings, N. D. & Woessner, J. E.), Catalytic mechanisms for metallopeptidases (Vol. **2nd ed**). San Diago, Calif, USA: Academic Press. 268-289.
23. **Davison, J.** 1999. Genetic exchange between bacteria in the environment. Plasmid, **42**(2), 73-91.
24. **DebRoy, C., & Maddox, C. W.** 2001. Identification of virulence attributes of gastrointestinal *Escherichia coli* isolates of veterinary significance. Anim Health Res Rev, **2**(2), 129-140.
25. **Deng, F., Jia, X., Lai, S., Liu, Y., & Chen, S.** 2015. [Transcript assembly and quality assessment]. Sheng Wu Gong Cheng Xue Bao, **31**(9), 1271-1278.
26. **Dobrindt, U., Blum-Oehler, G., Nagy, G., Schneider, G., Johann, A., Gottschalk, G., & Hacker, J.** 2002. Genetic structure and distribution of four pathogenicity islands (PAI I(536) to PAI IV(536)) of uropathogenic *Escherichia coli* strain 536. Infect Immun, **70**(11), 6365-6372.
27. **Dobrindt, U., & Hacker, J.** 2006. in *The Comprehensive Sourcebook of Bacterial Protein toxins* (eds Alouf, J. E. & Popoff, M. R.) Chapter 3: Mobile genetic elements and pathogenicity islands encoding bacteial toxins (Vol. **Thd. Ed.**). London: Academic Press. 44-63.
28. **Dong, N., Liu, L., & Shao, F.** 2010. A bacterial effector targets host DH-PH domain RhoGEFs and antagonizes macrophage phagocytosis. EMBO J, **29**(8), 1363-1376.
29. **Donohue-Rolfe, A., Keusch, G. T., Edson, C., Thorley-Lawson, D., & Jacewicz, M.** 1984. Pathogenesis of *Shigella diarrhea*. IX. Simplified high yield purification of Shigella toxin and characterization of subunit composition and function by the use of subunit-specific monoclonal and polyclonal antibodies. J Exp Med, **160**(6), 1767-1781.
30. **Dutta, P. R., Sui, B. Q., & Nataro, J. P.** 2003. Structure-function analysis of the enteroaggregative *Escherichia coli* plasmid-encoded toxin autotransporter using scanning linker mutagenesis. J Biol Chem, **278**(41), 39912-39920.
31. **Eichhorn, I.** (2016). Microevolution of epidemiological highly relevant non-O157 enterohemorrhagic *Escherichia coli* (EHEC). Dissertation, Freie Universität Berlin, Berlin. Retrieved from https://refubium.fu-berlin.de/handle/fub188/10712
32. **Eslava, C., Navarro-Garcia, F., Czeczulin, J. R., Henderson, I. R., Cravioto, A., & Nataro, J. P.** 1998. Pet, an autotransporter enterotoxin from enteroaggregative *Escherichia coli*. Infect Immun, **66**(7), 3155-3163.
33. **Fabbri, A., Travaglione, S., & Fiorentini, C.** 2010. *Escherichia coli* cytotoxic necrotizing factor 1 (CNF1): toxin biology, in vivo applications and therapeutic potential. Toxins (Basel), **2**(2), 283-296.
34. **Felsenstein, J., & Churchill, G. A.** 1996. A Hidden Markov Model approach to variation among sites in rate of evolution. Mol Biol Evol, **13**(1), 93-104.
35. **Feng, D. F., & Doolittle, R. F.** 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol, **25**(4), 351-360.
36. **Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., . . . Bateman, A.** 2008. The Pfam protein families database. Nucleic Acids Res, **36**(Database issue), D281-288.
37. **Fraser, M. E., Chernaia, M. M., Kozlov, Y. V., & James, M. N.** 1994. Crystal structure of the holotoxin from *Shigella dysenteriae* at 2.5 A resolution. Nat Struct Biol, **1**(1), 59-64.

38. **Fraser, M. E., Fujinaga, M., Cherney, M. M., Melton-Celsa, A. R., Twiddy, E. M., O'Brien, A. D., & James, M. N.** 2004. Structure of shiga toxin type 2 (Stx2) from *Escherichia coli* O157:H7. J Biol Chem, **279**(26), 27511-27517.
39. **Freyer, M. W., & Lewis, E. A.** 2008. Isothermal titration calorimetry: experimental design, data analysis, and probing macromolecule/ligand binding and kinetic interactions. Methods Cell Biol, **84**, 79-113.
40. **Friedrich, A. W., Bielaszewska, M., Zhang, W. L., Pulz, M., Kuczius, T., Ammon, A., & Karch, H.** 2002. *Escherichia coli* harboring Shiga toxin 2 gene variants: frequency and association with clinical symptoms. J Infect Dis, **185**(1), 74-84.
41. **Fukasawa, K. M., Hata, T., Ono, Y., & Hirose, J.** 2011. Metal preferences of zinc-binding motif on metalloproteases. J Amino Acids, **2011**, 574816.
42. **Fukasawa, K. M., Hirose, J., Hata, T., & Ono, Y.** 2010. In rat dipeptidyl peptidase III, His(5)(6)(8) is essential for catalysis, and Glu(5)(0)(7) or Glu(5)(1)(2) stabilizes the coordination bond between His(4)(5)(5) or His(4)(5)(0) and zinc ion. Biochim Biophys Acta, **1804**(10), 2063-2069.
43. **Fuller, C. A., Pellino, C. A., Flagler, M. J., Strasser, J. E., & Weiss, A. A.** 2011. Shiga toxin subtypes display dramatic differences in potency. Infect Immun, **79**(3), 1329-1337.
44. **Funk, J., Stoeber, H., Hauser, E., & Schmidt, H.** 2013. Molecular analysis of subtilase cytotoxin genes of food-borne Shiga toxin-producing *Escherichia coli* reveals a new allelic subAB variant. BMC Microbiol, **13**, 230.
45. **Furniss, R. C. D., Low, W. W., Mavridou, D. A. I., Dagley, L. F., Webb, A. I., Tate, E. W., & Clements, A.** 2018. Plasma membrane profiling during enterohemorrhagic *E. coli* infection reveals that the metalloprotease StcE cleaves CD55 from host epithelial surfaces. J Biol Chem, **293**(44), 17188-17199.
46. **Gaastra, W., & Svennerholm, A. M.** 1996. Colonization factors of human enterotoxigenic *Escherichia coli* (ETEC). Trends Microbiol, **4**(11), 444-452.
47. **Gallegos, K. M., Conrady, D. G., Karve, S. S., Gunasekera, T. S., Herr, A. B., & Weiss, A. A.** 2012. Shiga toxin binding to glycolipids and glycans. PLoS One, **7**(2), e30368.
48. **Garnier, J., Gibrat, J. F., & Robson, B.** 1996. GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol, **266**, 540-553.
49. **Garnier, J., Osguthorpe, D. J., & Robson, B.** 1978. Analysis of Accuracy and Implications of Simple Methods for Predicting Secondary Structure of Globular Proteins. Journal of Molecular Biology, **120**(1), 97-120.
50. **Gomis-Ruth, F. X., Grams, F., Yiallouros, I., Nar, H., Kusthardt, U., Zwilling, R., . . . Stocker, W.** 1994. Crystal structures, spectroscopic features, and catalytic properties of cobalt(II), copper(II), nickel(II), and mercury(II) derivatives of the zinc endopeptidase astacin. A correlation of structure and proteolytic activity. J Biol Chem, **269**(25), 17111-17117.
51. **Grasso, F., & Frisan, T.** 2015. Bacterial Genotoxins: Merging the DNA Damage Response into Infection Biology. Biomolecules, **5**(3), 1762-1782.
52. **Hackett, R., & Kam, P. C.** 2007. Botulinum toxin: pharmacology and clinical developments: a literature review. Med Chem, **3**(4), 333-345.
53. **Harrington, S. M., Dudley, E. G., & Nataro, J. P.** 2006. Pathogenesis of enteroaggregative *Escherichia coli* infection. FEMS Microbiol Lett, **254**(1), 12-18.
54. **Hartland, E. L., & Leong, J. M.** 2013. Enteropathogenic and enterohemorrhagic *E. coli*: ecology, pathogenesis, and evolution. Front Cell Infect Microbiol, **3**, 15.
55. **Hasegawa, M., Kishino, H., & Yano, T.** 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol, **22**(2), 160-174.
56. **Head, S. C., Karmali, M. A., & Lingwood, C. A.** 1991. Preparation of VT1 and VT2 hybrid toxins from their purified dissociated subunits. Evidence for B subunit modulation of a subunit function. J Biol Chem, **266**(6), 3617-3621.
57. **Henderson, I. R., Cappello, R., & Nataro, J. P.** 2000. Autotransporter proteins, evolution and redefining protein secretion: response. Trends Microbiol, **8**(12), 534-535.

58. **Henderson, I. R., Navarro-Garcia, F., & Nataro, J. P.** 1998. The great escape: structure and function of the autotransporter proteins. Trends Microbiol, **6**(9), 370-378.
59. **Henkel, J. S., Baldwin, M. R., & Barbieri, J. T.** 2010. Toxins from bacteria. EXS, **100**, 1-29.
60. **Hews, C. L., Tran, S. L., Wegmann, U., Brett, B., Walsham, A. D. S., Kavanaugh, D., . . . Schuller, S.** 2017. The StcE metalloprotease of enterohaemorrhagic *Escherichia coli* reduces the inner mucus layer and promotes adherence to human colonic epithelium ex vivo. Cell Microbiol, **19**(6).
61. **Hoffmann, S., Stadler, P. F., & Strimmer, K.** 2015. A simple data-adaptive probabilistic variant calling model. Algorithms Mol Biol, **10**, 10.
62. **Holmes, A., Muhlen, S., Roe, A. J., & Dean, P.** 2010. The EspF effector, a bacterial pathogen's Swiss army knife. Infect Immun, **78**(11), 4445-4453.
63. **Jacob-Dubuisson, F., Fernandez, R., & Coutte, L.** 2004. Protein secretion through autotransporter and two-partner pathways. Biochim Biophys Acta, **1694**(1-3), 235-257.
64. **Jandhyala, D. M., Thorpe, C. M., & Magun, B.** 2012. Ricin and Shiga toxins: effects on host cell signal transduction. Curr Top Microbiol Immunol, **357**, 41-65.
65. **Johnson, T. J., & Nolan, L. K.** 2009. Pathogenomics of the virulence plasmids of *Escherichia coli*. Microbiol Mol Biol Rev, **73**(4), 750-774.
66. **Jukes, T., & Cantor, C.** 1969. Evolution of protein molecules. New York: Academic Press. 21-32.
67. **Kaper, J. B., Nataro, J. P., & Mobley, H. L.** 2004. Pathogenic *Escherichia coli*. Nat Rev Microbiol, **2**(2), 123-140.
68. **Kaper, J. B., & O'Brien, A. D.** 2014. Overview and Historical Perspectives. Microbiol Spectr, **2**(6).
69. **Karch, H., Schubert, S., Zhang, D., Zhang, W., Schmidt, H., Olschlager, T., & Hacker, J.** 1999. A genomic island, termed high-pathogenicity island, is present in certain non-O157 Shiga toxin-producing *Escherichia coli* clonal lineages. Infect Immun, **67**(11), 5994-6001.
70. **Karve, S. S., & Weiss, A. A.** 2014. Glycolipid binding preferences of Shiga toxin variants. PLoS One, **9**(7), e101173.
71. **Kelley, L. A., & Sternberg, M. J.** 2009. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc, **4**(3), 363-371.
72. **Khan, S., Mian, H. S., Sandercock, L. E., Chirgadze, N. Y., & Pai, E. F.** 2011. Crystal structure of the passenger domain of the Escherichia coli autotransporter EspP. J Mol Biol, **413**(5), 985-1000.
73. **Kimura, M.** 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, **16**(2), 111-120.
74. **Klink, B. U., Barden, S., Heidler, T. V., Borchers, C., Ladwein, M., Stradal, T. E., . . . Heinz, D. W.** 2010. Structure of *Shigella* IpgB2 in complex with human RhoA: implications for the mechanism of bacterial guanine nucleotide exchange factor mimicry. J Biol Chem, **285**(22), 17197-17208.
75. **Kloczkowski, A., Ting, K. L., Jernigan, R. L., & Garnier, J.** 2002. Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. Proteins, **49**(2), 154-166.
76. **Kotloff, K. L., Nataro, J. P., Blackwelder, W. C., Nasrin, D., Farag, T. H., Panchalingam, S., . . . Levine, M. M.** 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. Lancet, **382**(9888), 209-222.
77. **Lai, X. H., Arencibia, I., Johansson, A., Wai, S. N., Oscarsson, J., Kalfas, S., . . . Uhlin, B. E.** 2000. Cytocidal and apoptotic effects of the ClyA protein from *Escherichia coli* on primary and cultured monocytes and macrophages. Infect Immun, **68**(7), 4363-4367.

78. **Laohachai, K. N., Bahadi, R., Hardo, M. B., Hardo, P. G., & Kourie, J. I.** 2003. The role of bacterial and non-bacterial toxins in the induction of changes in membrane transport: implications for diarrhea. Toxicon, **42**(7), 687-707.
79. **Lathem, W. W., Grys, T. E., Witowski, S. E., Torres, A. G., Kaper, J. B., Tarr, P. I., & Welch, R. A.** 2002. StcE, a metalloprotease secreted by *Escherichia coli* O157:H7, specifically cleaves C1 esterase inhibitor. Mol Microbiol, **45**(2), 277-288.
80. **Le Gall, T., Mavris, M., Martino, M. C., Bernardini, M. L., Denamur, E., & Parsot, C.** 2005. Analysis of virulence plasmid gene expression defines three classes of effectors in the type III secretion system of *Shigella flexneri*. Microbiology, **151**(Pt 3), 951-962.
81. **Le Nours, J., Paton, A. W., Byres, E., Troy, S., Herdman, B. P., Johnson, M. D., . . . Beddoe, T.** 2013. Structural basis of subtilase cytotoxin SubAB assembly. J Biol Chem, **288**(38), 27505-27516.
82. **Levine, M. M., Xu, J. G., Kaper, J. B., Lior, H., Prado, V., Tall, B., . . . Wachsmuth, K.** 1987. A DNA probe to identify enterohemorrhagic *Escherichia coli* of O157:H7 and other serotypes that cause hemorrhagic colitis and hemolytic uremic syndrome. J Infect Dis, **156**(1), 175-182.
83. **Leyton, D. L., Sevastsyanovich, Y. R., Browning, D. F., Rossiter, A. E., Wells, T. J., Fitzpatrick, R. E., . . . Henderson, I. R.** 2011. Size and conformation limits to secretion of disulfide-bonded loops in autotransporter proteins. J Biol Chem, **286**(49), 42283-42291.
84. **Lindgren, S. W., Samuel, J. E., Schmitt, C. K., & O'Brien, A. D.** 1994. The specific activities of Shiga-like toxin type II (SLT-II) and SLT-II-related toxins of enterohemorrhagic *Escherichia coli* differ when measured by Vero cell cytotoxicity but not by mouse lethality. Infect Immun, **62**(2), 623-631.
85. **Ling, H., Pannu, N. S., Boodhoo, A., Armstrong, G. D., Clark, C. G., Brunton, J. L., & Read, R. J.** 2000. A mutant Shiga-like toxin IIe bound to its receptor Gb(3): structure of a group II Shiga-like toxin with altered binding specificity. Structure, **8**(3), 253-264.
86. **Linhartova, I., Bumba, L., Masin, J., Basler, M., Osicka, R., Kamanova, J., . . . Sebo, P.** 2010. RTX proteins: a highly diverse family secreted by a common mechanism. FEMS Microbiol Rev, **34**(6), 1076-1112.
87. **Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. E.** 2001. Molecular Cell Biology, Chapter 3: Protein Structure and Funktion (Vol. **Fourth Ed**). New York: W. H. Freeman and Company. 51-62.
88. **Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., & Darnell, J. E.** 2001. Molecular Cell Biology, Chapter 8: Genetic Analysis in Cell Biology (Vol. **Fourth Ed**). New York: W. H. Freeman and Company. 254-293.
89. **Ludwig, A., von Rhein, C., Bauer, S., Huttinger, C., & Goebel, W.** 2004. Molecular analysis of cytolysin A (ClyA) in pathogenic Escherichia coli strains. J Bacteriol, **186**(16), 5311-5320.
90. **Luo, Q., Kumar, P., Vickers, T. J., Sheikh, A., Lewis, W. G., Rasko, D. A., . . . Fleckenstein, J. M.** 2014. Enterotoxigenic *Escherichia coli* secretes a highly conserved mucin-degrading metalloprotease to effectively engage intestinal epithelial cells. Infect Immun, **82**(2), 509-521.
91. **Majowicz, S. E., Scallan, E., Jones-Bitton, A., Sargeant, J. M., Stapleton, J., Angulo, F. J., . . . Kirk, M. D.** 2014. Global Incidence of Human Shiga Toxin-Producing *Escherichia coli* Infections and Deaths: A Systematic Review and Knowledge Synthesis. Foodborne Pathogens and Disease, **11**(6), 447-455.
92. **Marchler-Bauer, A., Anderson, J. B., Derbyshire, M. K., DeWeese-Scott, C., Gonzales, N. R., Gwadz, M., . . . Bryant, S. H.** 2007. CDD: a conserved domain database for interactive domain family analysis. Nucleic Acids Res, **35**(Database issue), D237-240.
93. **McDaniel, T. K., Jarvis, K. G., Donnenberg, M. S., & Kaper, J. B.** 1995. A genetic locus of enterocyte effacement conserved among diverse enterobacterial pathogens. Proc Natl Acad Sci U S A, **92**(5), 1664-1668.

94. **Meisen, I., Rosenbruck, R., Galla, H. J., Huwel, S., Kouzel, I. U., Mormann, M., . . . Muthing, J.** 2013. Expression of Shiga toxin 2e glycosphingolipid receptors of primary porcine brain endothelial cells and toxin-mediated breakdown of the blood-brain barrier. Glycobiology, **23**(6), 745-759.
95. **Melton-Celsa, A. R.** 2014. Shiga Toxin (Stx) Classification, Structure, and Function. Microbiol Spectr, **2**(4), EHEC-0024-2013.
96. **Meyts, I., Bosch, B., Bolze, A., Boisson, B., Itan, Y., Belkadi, A., . . . Casanova, J. L.** 2016. Exome and genome sequencing for inborn errors of immunity. J Allergy Clin Immunol, **138**(4), 957-969.
97. **Michelacci, V., Tozzoli, R., Caprioli, A., Martinez, R., Scheutz, F., Grande, L., . . . Morabito, S.** 2013. A new pathogenicity island carrying an allelic variant of the Subtilase cytotoxin is common among Shiga toxin producing *Escherichia coli* of human and ovine origin. Clin Microbiol Infect, **19**(3), E149-156.
98. **Michener, C. D., & Sokal, R. R.** 1957. A Quantitative Approach to a Problem in Classification. Evol, **11**(2), 130-162.
99. **Michie, K. A., Boysen, A., Low, H. H., Moller-Jensen, J., & Lowe, J.** 2014. LeoA, B and C from enterotoxigenic *Escherichia coli* (ETEC) are bacterial dynamins. PLoS One, **9**(9), e107211.
100. **Muthing, J., Meisen, I., Zhang, W., Bielaszewska, M., Mormann, M., Bauerfeind, R., . . . Karch, H.** 2012. Promiscuous Shiga toxin 2e and its intimate relationship to Forssman. Glycobiology, **22**(6), 849-862.
101. **Muthing, J., Schweppe, C. H., Karch, H., & Friedrich, A. W.** 2009. Shiga toxins, glycosphingolipid diversity, and endothelial cell injury. Thromb Haemost, **101**(2), 252-264.
102. **Nakjang, S., Ndeh, D. A., Wipat, A., Bolam, D. N., & Hirt, R. P.** 2012. A novel extracellular metallopeptidase domain shared by animal host-associated mutualistic and pathogenic microbes. PLoS One, **7**(1), e30287.
103. **Nataro, J. P., Deng, Y., Cookson, S., Cravioto, A., Savarino, S. J., Guers, L. D., . . . Tacket, C. O.** 1995. Heterogeneity of enteroaggregative *Escherichia coli* virulence demonstrated in volunteers. J Infect Dis, **171**(2), 465-468.
104. **Nataro, J. P., & Kaper, J. B.** 1998. Diarrheagenic *Escherichia coli*. Clin Microbiol Rev, **11**(1), 142-201.
105. **National Center for Biotechnology Information.** (2013). The New BLAST Results Page. *NCBI Handout Series - New BLAST*, from ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_NewBLAST.pdf
106. **Navarro-Garcia, F., Canizalez-Roman, A., Burlingame, K. E., Teter, K., & Vidal, J. E.** 2007. Pet, a non-AB toxin, is transported and translocated into epithelial cells by a retrograde trafficking pathway. Infect Immun, **75**(5), 2101-2109.
107. **Navarro-Garcia, F., Canizalez-Roman, A., Vidal, J. E., & Salazar, M. I.** 2007. Intoxication of epithelial cells by plasmid-encoded toxin requires clathrin-mediated endocytosis. Microbiology, **153**(Pt 9), 2828-2838.
108. **Needleman, S. B., & Wunsch, C. D.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol, **48**(3), 443-453.
109. **Nirenberg, M. W., Jones, O. W., Leder, P., Clark, B. F. C., Sly, W. C., & Pestka, S.** 1963. On the Coding of Genetic Information. Cold Spring Harb Symp Quant Biol, **28**, 549–557.
110. **Nougayrede, J. P., Homburg, S., Taieb, F., Boury, M., Brzuszkiewicz, E., Gottschalk, G., . . . Oswald, E.** 2006. Escherichia coli induces DNA double-strand breaks in eukaryotic cells. Science, **313**(5788), 848-851.
111. **Nuesch-Inderbinen, M. T., Funk, J., Cernela, N., Tasara, T., Klumpp, J., Schmidt, H., & Stephan, R.** 2015. Prevalence of subtilase cytotoxin-encoding subAB variants among Shiga toxin-producing *Escherichia coli* strains isolated from wild ruminants and sheep differs from that of cattle and pigs and is predominated by the new allelic variant subAB2-2. Int J Med Microbiol, **305**(1), 124-128.

112. **Ohlson, T., Wallner, B., & Elofsson, A.** 2004. Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods. Proteins, **57**(1), 188-197.
113. **Olsnes, S., Reisbig, R., & Eiklid, K.** 1981. Subunit structure of *Shigella* cytotoxin. J Biol Chem, **256**(16), 8732-8738.
114. **Orden, J. A., Horcajo, P., de la Fuente, R., Ruiz-Santa-Quiteria, J. A., Dominguez-Bernal, G., & Carrion, J.** 2011. Subtilase cytotoxin-coding genes in verotoxin-producing *Escherichia coli* strains from sheep and goats differ from those from cattle. Appl Environ Microbiol, **77**(23), 8259-8264.
115. **Ostroff, S. M., Tarr, P. I., Neill, M. A., Lewis, J. H., Hargrett-Bean, N., & Kobayashi, J. M.** 1989. Toxin genotypes and plasmid profiles as determinants of systemic sequelae in *Escherichia coli* O157:H7 infections. J Infect Dis, **160**(6), 994-998.
116. **Parsot, C.** 2005. *Shigella* spp. and enteroinvasive *Escherichia coli* pathogenicity factors. FEMS Microbiol Lett, **252**(1), 11-18.
117. **Paton, A. W., Srimanote, P., Talbot, U. M., Wang, H., & Paton, J. C.** 2004. A new family of potent AB(5) cytotoxins produced by Shiga toxigenic *Escherichia coli*. J Exp Med, **200**(1), 35-46.
118. **Pearson, W. R.** 2013. Selecting the Right Similarity-Scoring Matrix. Curr Protoc Bioinformatics, **43**, 3 5 1-9.
119. **Protein Data Bank** (PDB). (2015, 22.09.2015). Statistics, from http://www.rcsb.org/pdb/statistics
120. **Pupo, G. M., Lan, R., & Reeves, P. R.** 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A, **97**(19), 10567-10572.
121. **Ratnam, S., March, S. B., Ahmed, R., Bezanson, G. S., & Kasatiya, S.** 1988. Characterization of *Escherichia coli* serotype O157:H7. J Clin Microbiol, **26**(10), 2006-2012.
122. **Rawlings, N. D., & Barret, A. J.** 2004. in Handbook of Proteolytic Enzyme (eds. Barret, A. J., Rawlings, N. D. & Woessner, J. E.), Introduction: metallopeptidases and their clans (Vol. **2nd ed**). San Diego, Calif, USA: Academic Press. 231-268.
123. **Reeves, P. R., Liu, B., Zhou, Z., Li, D., Guo, D., Ren, Y., . . . Wang, L.** 2011. Rates of mutation and host transmission for an *Escherichia coli* clone over 3 years. PLoS One, **6**(10), e26907.
124. **Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K., & Whittam, T. S.** 2000. Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature, **406**(6791), 64-67.
125. **Reisbig, R., Olsnes, S., & Eiklid, K.** 1981. The cytotoxic activity of *Shigella* toxin. Evidence for catalytic inactivation of the 60 S ribosomal subunit. J Biol Chem, **256**(16), 8739-8744.
126. **Robert-Koch-Institut**. 2011. Bericht: Abschließende Darstellung und Bewertung der epidemiologischen Erkenntnisse im EHEC O104:H4 Ausbruch, Deutschland 2011. **Berlin 2011**, 1-44.
127. **Robert-Koch-Institut**. 2011. Informationen zum EHEC-/HUS-Ausbruchsgeschehen von Mai bis Juli 2011 in Deutschland - Ende des Ausbruchs. Epidemiol Bull, **2011**(31), 295-296.
128. **Robert-Koch-Institut**. 2015. Aktuelle Daten und Informationenen zu Infektionskrankheiten und Public Health. Epidemiol Bull, **2015**(20), 165-174.
129. **Robert-Koch-Institut**. 2018. Aktuelle Daten und Informationenen zu Infektionskrankheiten und Public Health. Epidemiol Bull, **2018 (21)**, 199-204.
130. **Roberts, P. H., Davis, K. C., Garstka, W. R., & Bhunia, A. K.** 2001. Lactate dehydrogenase release assay from Vero cells to distinguish verotoxin producing *Escherichia coli* from non-verotoxin producing strains. J Microbiol Methods, **43**(3), 171-181.

131. **Roderer, D., Benke, S., Muller, M., Fah-Rechsteiner, H., Ban, N., Schuler, B., & Glockshuber, R.** 2014. Characterization of variants of the pore-forming toxin ClyA from Escherichia coli controlled by a redox switch. Biochemistry, **53**(40), 6357-6369.
132. **Ruiz-Perez, F., & Nataro, J. P.** 2014. Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity, and role in virulence. Cell Mol Life Sci, **71**(5), 745-770.
133. **Saitou, N., & Nei, M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol, **4**(4), 406-425.
134. **Sandvig, K., Bergan, J., Dyve, A. B., Skotland, T., & Torgersen, M. L.** 2010. Endocytosis and retrograde transport of Shiga toxin. Toxicon, **56**(7), 1181-1185.
135. **Scheutz, F., Teel, L. D., Beutin, L., Pierard, D., Buvens, G., Karch, H., . . . O'Brien, A. D.** 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. J Clin Microbiol, **50**(9), 2951-2963.
136. **Schmidt, H., Bielaszewska, M., & Karch, H.** 1999. 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, **65**(9), 3855-3861.
137. **Schmidt, H., & Hensel, M.** 2004. Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev, **17**(1), 14-56.
138. **Schubert, S., Picard, B., Gouriou, S., Heesemann, J., & Denamur, E.** 2002. *Yersinia* high-pathogenicity island contributes to virulence in *Escherichia coli* causing extraintestinal infections. Infect Immun, **70**(9), 5335-5337.
139. **Sears, C. L., & Kaper, J. B.** 1996. Enteric bacterial toxins: mechanisms of action and linkage to intestinal secretion. Microbiol Rev, **60**(1), 167-215.
140. **Semmler, T.** 2018. Population Genetics of *Escherichia coli* - a Genomic Approach: Dissertation, Freie Universität Berlin.
141. **Sen, T. Z., Jernigan, R. L., Garnier, J., & Kloczkowski, A.** 2005. GOR V server for protein secondary structure prediction. Bioinformatics, **21**(11), 2787-2788.
142. **Seyahian, E. A., Oltra, G., Ochoa, F., Melendi, S., Hermes, R., Paton, J. C., . . . Zotta, E.** 2017. Systemic effects of Subtilase cytotoxin produced by Escherichia coli O113:H21. Toxicon, **127**, 49-55.
143. **Sims, G. E., & Kim, S. H.** 2011. Whole-genome phylogeny of *Escherichia coli*/*Shigella* group by feature frequency profiles (FFPs). Proc Natl Acad Sci U S A, **108**(20), 8329-8334.
144. **Smith, T. F., & Waterman, M. S.** 1981. Identification of common molecular subsequences. J Mol Biol, **147**(1), 195-197.
145. **Speyer, F. J., Lengyel, P., Basilio, C., Wahba, A. J., Gardner, R. S., & Ochoa, S.** 1963. Synthetic polynucleotides and the amino acid code. Cold Spring Harb Symp Quant Biol, **28**, 559–567.
146. **Spiering, D., & Hodgson, L.** 2011. Dynamics of the Rho-family small GTPases in actin regulation and motility. Cell Adh Migr, **5**(2), 170-180.
147. **Steele, M., Ziebell, K., Zhang, Y., Benson, A., Konczy, P., Johnson, R., & Gannon, V.** 2007. Identification of *Escherichia coli* O157:H7 genomic regions conserved in strains with a genotype associated with human infection. Appl Environ Microbiol, **73**(1), 22-31.
148. **Szabady, R. L., Yanta, J. H., Halladin, D. K., Schofield, M. J., & Welch, R. A.** 2011. TagA is a secreted protease of Vibrio cholerae that specifically cleaves mucin glycoproteins. Microbiology, **157**(Pt 2), 516-525.
149. **Tamura, K., & Nei, M.** 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol, **10**(3), 512-526.
150. **Tapader, R., Bose, D., & Pal, A.** 2017. YghJ, the secreted metalloprotease of pathogenic *E. coli* induces hemorrhagic fluid accumulation in mouse ileal loop. Microb Pathog, **105**, 96-99.

151. **Tauschek, M., Strugnell, R. A., & Robins-Browne, R. M.** 2002. Characterization and evidence of mobilization of the LEE pathogenicity island of rabbit-specific strains of enteropathogenic *Escherichia coli*. Mol Microbiol, **44**(6), 1533-1550.
152. **Tavaré, S.** 1986. Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences. Lectures on Mathematics in the Life Sciences, **17**(American Mathematical Society), 57-86.
153. **Tesh, V. L.** 2012. The induction of apoptosis by Shiga toxins and ricin. Curr Top Microbiol Immunol, **357**, 137-178.
154. **Thapar, N., & Sanderson, I. R.** 2004. Diarrhoea in children: an interface between developing and developed countries. Lancet, **363**(9409), 641-653.
155. **Thompson, C. N., Duy, P. T., & Baker, S.** 2015. The Rising Dominance of *Shigella sonnei*: An Intercontinental Shift in the Etiology of Bacillary Dysentery. PLoS Negl Trop Dis, **9**(6), e0003708.
156. **Tobe, T., Hayashi, T., Han, C. G., Schoolnik, G. K., Ohtsubo, E., & Sasakawa, C.** 1999. Complete DNA sequence and structural analysis of the enteropathogenic *Escherichia coli* adherence factor plasmid. Infect Immun, **67**(10), 5455-5462.
157. **Torres, A. G., & Kaper, J. B.** 2002. Pathogenicity islands of intestinal *E. coli*. Curr Top Microbiol Immunol, **264**(1), 31-48.
158. **Tsutsuki, H., Yahiro, K., Suzuki, K., Suto, A., Ogura, K., Nagasawa, S., . . . Noda, M.** 2012. Subtilase cytotoxin enhances *Escherichia coli* survival in macrophages by suppression of nitric oxide production through the inhibition of NF-kappaB activation. Infect Immun, **80**(11), 3939-3951.
159. **Tu, X., Nisan, I., Yona, C., Hanski, E., & Rosenshine, I.** 2003. EspH, a new cytoskeleton-modulating effector of enterohaemorrhagic and enteropathogenic Escherichia coli. Mol Microbiol, **47**(3), 595-606.
160. **Turner, S. M., Scott-Tucker, A., Cooper, L. M., & Henderson, I. R.** 2006. Weapons of mass destruction: virulence factors of the global killer enterotoxigenic *Escherichia coli*. FEMS Microbiol Lett, **263**(1), 10-20.
161. **Wallace, A. J., Stillman, T. J., Atkins, A., Jamieson, S. J., Bullough, P. A., Green, J., & Artymiuk, P. J.** 2000. *E. coli* hemolysin E (HlyE, ClyA, SheA): X-ray crystal structure of the toxin and observation of membrane pores by electron microscopy. Cell, **100**(2), 265-276.
162. **Wang, Y., Virtanen, J., Xue, Z., & Zhang, Y.** 2017. I-TASSER-MR: automated molecular replacement for distant-homology proteins using iterative fragment assembly and progressive sequence truncation. Nucleic Acids Res, **45**(W1), W429-W434.
163. **Webb, S. A., & Kahler, C. M.** 2008. Bench-to-bedside review: Bacterial virulence and subversion of host defences. Crit Care, **12**(6), 234.
164. **Weinstein, D. L., Jackson, M. P., Perera, L. P., Holmes, R. K., & O'Brien, A. D.** 1989. In vivo formation of hybrid toxins comprising Shiga toxin and the Shiga-like toxins and role of the B subunit in localization and cytotoxic activity. Infect Immun, **57**(12), 3743-3750.
165. **Welch, R. A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., . . . Blattner, F. R.** 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci U S A, **99**(26), 17020-17024.
166. **Wong, A. R., Raymond, B., Collins, J. W., Crepin, V. F., & Frankel, G.** 2012. The enteropathogenic *E. coli* effector EspH promotes actin pedestal formation and elongation via WASP-interacting protein (WIP). Cell Microbiol, **14**(7), 1051-1070.
167. **Xia, X., Meng, J., McDermott, P. F., Ayers, S., Blickenstaff, K., Tran, T. T., . . . Zhao, S.** 2010. Presence and characterization of shiga toxin-producing *Escherichia coli* and other potentially diarrheagenic *E. coli* strains in retail meats. Appl Environ Microbiol, **76**(6), 1709-1717.
168. **Yahiro, K., Satoh, M., Morinaga, N., Tsutsuki, H., Ogura, K., Nagasawa, S., . . . Noda, M.** 2011. Identification of subtilase cytotoxin (SubAB) receptors whose

signaling, in association with SubAB-induced BiP cleavage, is responsible for apoptosis in HeLa cells. Infect Immun, **79**(2), 617-627.

169. **Yamamoto, T., Wakisaka, N., Sato, F., & Kato, A.** 1997. Comparison of the nucleotide sequence of enteroaggregative *Escherichia coli* heat-stable enterotoxin 1 genes among diarrhea-associated *Escherichia coli*. FEMS Microbiol Lett, **147**(1), 89-95.
170. **Yen, Y. T., Kostakioti, M., Henderson, I. R., & Stathopoulos, C.** 2008. Common themes and variations in serine protease autotransporters. Trends Microbiol, **16**(8), 370-379.
171. **Yu, J. Y., Jeon, H. G., Kang, Y. H., Kim, E. C., Sohn, C. K., & Lee, B. K.** 2001. Characterization of Shiga toxin genes in Shiga toxin-producing *Escherichia coli* isolated in Korea. Direct Submission to NCBI.
172. **Yuan, X., Johnson, M. D., Zhang, J., Lo, A. W., Schembri, M. A., Wijeyewickrema, L. C., . . . Leyton, D. L.** 2018. Molecular basis for the folding of beta-helical autotransporter passenger domains. Nat Commun, **9**(1), 1395.
173. **Zhang, C., Mortuza, S. M., He, B., Wang, Y., & Zhang, Y.** 2018. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins, **86 Suppl 1**, 136-151.
174. **Zhang, Y.** 2007. Template-based modeling and free modeling by I-TASSER in CASP7. Proteins, **69 Suppl 8**, 108-117.
175. **Zhang, Y.** 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics, **9**, 40.
176. **Zhou, Z., Li, X., Liu, B., Beutin, L., Xu, J., Ren, Y., . . . Wang, L.** 2010. Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. PLoS One, **5**(1), e8700.
177. **Zuo, G., Xu, Z., & Hao, B.** 2013. *Shigella* strains are not clones of *Escherichia coli* but sister species in the genus *Escherichia*. Genomics Proteomics Bioinformatics, **11**(1), 61-65.

11. Appendix

Table A 1: Internal database ECore (*E. coli* Reference Collection)

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
6551_2#14		*Shigella*		human					
6551_3#12		*Shigella*		human					
6551_3#16		*Shigella*		human					
6613_1#18		*Shigella*		human					
6613_2#13		*Shigella*		human					
6613_2#16		*Shigella*		human					
6613_2#8		*Shigella*		human					
6613_3#12		*Shigella*		human					
6613_3#21		*Shigella*		human					
6613_3#3		*Shigella*		human					
6613_3#5		*Shigella*		human					
6613_4#15		*Shigella*		human					
6711_2#6		*E. coli*	ETEC	human					
6732_2#1		*E. coli*	ETEC	human	healthy	University of Gothenburg, Sweden	2003	Guatemala	
6732_2#20		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1990	Indonesia	
6732_2#6		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Egypt	
6753_7#10		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1986/87	Bangladesh	
6753_7#15		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989		
6753_7#16		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Zaire	
6753_7#18		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989		
6753_7#2		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1985		
6753_7#21		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1983	Japan	
6753_7#23		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1987	Japan	
6753_8#14		*E. coli*	ETEC	human					
6753_8#21		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Egypt	
6753_8#22		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Egypt	
7067_5#15		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7067_5#28		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2007	Bangladesh	
7067_5#5		*E. coli*	ETEC	human					
7067_5#9		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7114_1#2		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2011	Bangladesh	
7114_1#22		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2007	Bangladesh	
7114_1#25		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2011	Bangladesh	
7114_1#26		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2011	Bangladesh	

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7114_1#4		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2011	Bangladesh	
7114_1#6		*E. coli*	ETEC	human		University of Gothenburg, Sweden		Bolivia	
7521_1#11		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2007	Thailand	
7521_1#17		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1998/00	Mexico	
7521_1#19		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1998/00	Guatemala	
7521_1#25		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1998/00	Guatemala	
7521_1#26		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Mexico	
7521_1#28		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Mexico	
7521_1#3		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Mexico	
7521_1#31		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1987	Japan	
7521_1#35		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Mexico	
7521_1#37		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Guatemala	
7521_1#38		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Guatemala	
7521_1#4		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Guatemala	
7521_1#43		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1987	Japan	
7521_1#7		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000/01	Guatemala	
7521_1#9		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1998/00	Mexico	
7521_2#52		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1998/00	Mexico	
7521_2#55		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2002/03	Guatemala	
7521_2#67		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2003	Guatemala	
7521_2#71		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2003	Guatemala	
7521_2#72		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2003	Guatemala	
7521_2#79		*E. coli*	ETEC	human					
7521_2#84		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Egypt	
7521_2#93		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Egypt	
7521_2#94		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000	Egypt	
7521_3#1		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2000	Egypt	
7521_3#10		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2001	Egypt	
7521_3#16		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1996	Egypt	
7521_3#17		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1996	Indonesia	
7521_3#2		*E. coli*	ETEC	human					
7521_3#21		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2001	Egypt	
7521_3#24		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_3#26		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7521_3#3		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_3#31		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2001	Egypt	
7521_3#32		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_3#38		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1983	Japan	
7521_3#39		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_3#47		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_3#8		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_4#53		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2001	Egypt	
7521_4#55		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_4#58		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1989	Argentina	
7521_4#61		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#62		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#64		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#65		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#66		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#67		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#68		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#69		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#70		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#71		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#75		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#78		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#91		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1997	Indonesia	
7521_4#93		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2008	Bolivia	
7521_4#96		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2008	Bolivia	
7553_7#67		*E. coli*	ETEC	human		University of Gothenburg, Sweden	2007	Bolivia	
7553_7#68		*E. coli*	aEPEC	human					
7553_7#71		*E. coli*	aEPEC	human					
7554_4#18		*E. coli*	aEPEC	human					
7558_1#1		*E. coli*	ETEC	human		University of Gothenburg, Sweden	1998	Egypt	
7558_1#2		*E. coli*	ETEC	human					
7738_5#10	12220	*E. coli*		chicken	healthy	Free University Berlin, IMT	2006	Germany	
7738_5#12	12543	*E. coli*		horse	sick	Synlab Augsburg	2007	Germany	
7738_5#13	13301	*E. coli*		turkey	sick	AniCon Laboratory, Höltinghausen	2007	Germany	

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7738_5#14	13350	*E. coli*		pig	sick	BVL - GermVet	2007	Germany	423
7738_5#17	13483	*E. coli*		pig	sick	BVL - GermVet	2007	Germany	4510
7738_5#18	13780	*E. coli*	aEPEC	human	healthy	University Köln, Institute of Genetics	2001	Germany	
7738_5#19	13800	*E. coli*		human	sick	University Köln, Institute of Genetics	2001	Germany	
7738_5#2	10497	*E. coli*		horse	sick	Free University Berlin, IMT	2005	Germany	
7738_5#23	13834	*E. coli*		human	UTI	University Köln, Institute of Genetics	2001	Germany	
7738_5#25	14116	*E. coli*		turkey	sick	AniCon Laboratory, Höltinghausen	2007	Germany	
7738_5#26	14124	*E. coli*		turkey	sick	AniCon Laboratory, Höltinghausen	2007	Germany	
7738_5#27	14719	*E. coli*		chicken	healthy	Free University Berlin, IMT	2008	Germany	
7738_5#28	14939	*E. coli*		turkey	sick	Free University Berlin, IMT	2008	Germany	
7738_5#29	14973	*E. coli*		cat	UTI	BfT - GermVet-Studie	2004	Germany	220
7738_5#3	10676	*E. coli*		chicken	environ	Free University Berlin, ITU	2005	Germany	V/7
7738_5#30	15014			cat	UTI	BfT - GermVet-Studie	2004	Germany	1103
7738_5#31	15026	*E. coli*		dog	sick	BfT - GermVet-Studie	2005	Germany	1370
7738_5#32	16001	*E. coli*	STEC	cattle		Free University Berlin, IMT	2008	Germany	
7738_5#33	16007	*E. coli*		cattle		Free University Berlin, IMT	2008	Germany	
7738_5#34	16009	*E. coli*		cattle		Free University Berlin, IMT	2008	Germany	
7738_5#35	17832	*E. coli*	STEC	human		Robert-Koch Institute, Wernigerode	2008	Germany	08-00912
7738_5#37	19611	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_5#38	19859	*E. coli*		cattle		Free University Berlin, IMT	2009	Germany	
7738_5#39	15191	*E. coli*		turkey	sick	Free University Berlin, IMT	2008	Germany	
7738_5#4	10740	*E. coli*		environ	environ	Free University Berlin, ITU	2003	Germany	V/9
7738_5#40	19946	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_5#41	20163	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_5#42	20164	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_5#43	20231	*E. coli*		cattle		Free University Berlin, IMT	2009	Germany	
7738_5#44	20319	*E. coli*		cattle		Free University Berlin, IMT	2009	Germany	
7738_5#45	15388	*E. coli*		cat	Enteritis	BfT - GermVet-Studie	2004	Germany	607
7738_5#46	15390	*E. coli*		dog	Enteritis	BfT - GermVet-Studie	2004	Germany	636
7738_5#47	15406	*E. coli*		cat	Enteritis	BfT - GermVet-Studie	2004	Germany	919
7738_5#48	15455	*E. coli*		pig	UTI	BfT - GermVet-Studie	2004	Germany	159
7738_5#5	11327	*E. coli*		chicken	healthy	Free University Berlin, ITU	2005	Germany	V/8
7738_5#8	12022	*E. coli*	NMEC	human	sick	Johns Hopkins University, School of Medicine	2006	USA	
7738_6#50	15470	*E. coli*		pig	sick	BfT - GermVet-Studie	2004	Germany	850

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7738_6#52	15523	*E. coli*		pig	UTI	BfT - GermVet-Studie	2005	Germany	2483
7738_6#53	15530	*E. coli*		pig	UTI	BfT - GermVet-Studie	2005	Germany	2724
7738_6#54	20321	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_6#55	27034	*E. coli*		pig		Free University Berlin, IMT	2010	Germany	
7738_6#57	16223	*E. coli*		turkey	sick	Free University Berlin, IMT	2008	Germany	
7738_6#60	18286	*E. coli*		dog	sick	VetMed Laboratory Ludwigsburg	2009	Germany	VB 941950
7738_6#61	18291	*E. coli*		horse	sick	VetMed Laboratory Ludwigsburg	2009	Germany	VB 980758
7738_6#62	18321	*E. coli*		horse	sick	VetMed Laboratory Ludwigsburg	2009	Germany	VB 969619
7738_6#65	18411	*E. coli*		human	healthy	University Würzburg, IHM	2006	Germany	Delta 99
7738_6#67	18449	*E. coli*		human	UTI	University Würzburg, IHM	2006	Germany	Delta 77
7738_6#68	18570	*E. coli*		dog	sick	VetMed Laboratory Ludwigsburg	2009	Netherlands	VB 984674
7738_6#69	18895	*E. coli*		chicken	sick	Free University Berlin, IMT	2009	Germany	
7738_6#70	18988	*E. coli*		horse	sick	VetMed Laboratory Ludwigsburg	2009	Germany	VB 961872
7738_6#71	19327	*E. coli*		dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 974460
7738_6#73	19340	*E. coli*		dog	Enteritis	VetMed Laboratory Ludwigsburg	2009	Germany	VB 933419
7738_6#74	19343	*E. coli*		dog	Enteritis	VetMed Laboratory Ludwigsburg	2009	Germany	VB 933496
7738_6#76	19526	*E. coli*		chicken	Enteritis	VetMed Laboratory Ludwigsburg	2009	Germany	VB 935842
7738_6#77	19743	*E. coli*		ape	Enteritis	VetMed Laboratory Ludwigsburg	2009	Germany	VB 950960
7738_6#78	19765	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 989928
7738_6#80	19779	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 990942
7738_6#81	19792	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	France	BF 131099
7738_6#82	19803	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 991265
7738_6#84	19814	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 991463
7738_6#86	19834	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 992436
7738_6#87	19836	*E. coli*	ExPEC	cat	UTI	VetMed Laboratory Ludwigsburg	2009	France	BF 131319
7738_6#89	20122	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Netherlands	VM 781901
7738_6#90	20148	*E. coli*	ExPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	France	BF 131795
7738_6#91	20434	*E. coli*		cat	Enteritis	VetMed Laboratory Ludwigsburg	2010	Germany	VB 913406
7738_6#92	20441	*E. coli*	ExPEC	cat	UTI	VetMed Laboratory Ludwigsburg	2010	Italy	VB 956886
7738_6#93	20469	*E. coli*	ExPEC	horse	WI	VetMed Laboratory Ludwigsburg	2010	Hungary	VB 956305
7738_6#94	20601	*E. coli*		dog	Enteritis	VetMed Laboratory Ludwigsburg	2010	Italy	VB 915561
7738_6#96	21440	*E. coli*	ExPEC	horse	UTI	VetMed Laboratory Ludwigsburg	2010	Germany	VB 984973
7738_7#11	22246	*E. coli*		horse	UTI	VetMed Laboratory Ludwigsburg	2010	Netherlands	VB 952527
7738_7#14	22533	*E. coli*	ExPEC	dog	Enteritis	VetMed Laboratory Ludwigsburg	2010	Norway	VB 903857

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7738_7#15	22537	*E. coli*	ExPEC	dog	Enteritis	VetMed Laboratory Ludwigsburg	2010	Netherlands	VB 903902
7738_7#17	22577	*E. coli*	ExPEC	dog	Enteritis	VetMed Laboratory Ludwigsburg	2010	Austria	VB 904456
7738_7#18	22589	*E. coli*	ExPEC	dog	Enteritis	VetMed Laboratory Ludwigsburg	2010	Austria	VB 904860
7738_7#19	2265	*E. coli*		chicken	Hepatitis	State-VetMed Investigation Office Aulendorf	1999	Germany	G748/1/99
7738_7#2	22136	*E. coli*	ExPEC	dog	Enteritis	VetMed Laboratory Ludwigsburg	2010	Germany	VB 900618
7738_7#20	22668	*E. coli*	ExPEC	cat	Enteritis	VetMed Laboratory Ludwigsburg	2010	Netherlands	VB 905443
7738_7#22	2358	*E. coli*		chicken	sick	LUFA Nord-West, Oldenburg (ITT)	1998	Germany	415/98
7738_7#25	27031	*E. coli*		pig	healthy	Free University Berlin, IMT	2010	Germany	
7738_7#26	27035	*E. coli*		pig	healthy	Free University Berlin, IMT	2010	Germany	
7738_7#27	27043	*E. coli*		pig	healthy	Free University Berlin, IMT	2010	Germany	
7738_7#28	27056	*E. coli*		pig	healthy	Free University Berlin, IMT	2010	Germany	
7738_7#29	27064	*E. coli*		pig	healthy	Free University Berlin, IMT	2010	Germany	
7738_7#30	27073	*E. coli*		pig	healthy	Free University Berlin, IMT	2010	Germany	
7738_7#32	4514	*E. coli*		chicken	sick	Free University Berlin, IMT	2000	Germany	
7738_7#34	4542	*E. coli*		chicken	sick	Free University Berlin, IMT	2000	Germany	
7738_7#35	5020	*E. coli*		chicken	sick	Free University Berlin, IMT	2000	Germany	
7738_7#36	5215	*E. coli*		chicken	Septicemia	Free University Berlin, IMT	2001	Germany	
7738_7#37	5472	*E. coli*		horse	Vaginitis	Free University Berlin, IMT	2002	Germany	
7738_7#41	9272	*E. coli*	NMEC	human	NM	University Würzburg, IHM	2004	USA	E 817
7738_7#42	3734	*E. coli*		cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Germany	RW2070
7738_7#43	3735	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Germany	RW1703
7738_7#44		*E. coli*	STEC						
7738_7#45		*E. coli*	STEC						
7738_7#46	3855	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Germany	RW1300
7738_7#47	3856	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Belgium	562/380S
7738_7#48		*E. coli*	aEPEC						
7738_7#5	22139	*E. coli*	ExPEC	cat	Enteritis	VetMed Laboratory Ludwigsburg	2010	Germany	VB 900961
7738_8#49		*E. coli*	STEC						
7738_8#50		*E. coli*	STEC						
7738_8#51		*E. coli*	STEC						
7738_8#52	4040	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	1989	Germany	413/W003
7738_8#53		*E. coli*							
7738_8#54		*E. coli*	STEC						
7738_8#55		*E. coli*	STEC						

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7738_8#56	14505	*E. coli*	aEPEC	mouse	healthy	Friedrich Löffler Institute, Riems	2007	Germany	07/0324/209
7738_8#57	19064	*E. coli*	aEPEC	human		University Münster, Medical Centre	2009	Germany	910/00
7738_8#58	19005	*E. coli*		dog	sick	VetMed Laboratory Ludwigsburg	2009	Germany	VB 900961
7738_8#59		*E. coli*	STEC						
7738_8#60	3173	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Germany	RW1706
7738_8#61	3748	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Belgium	377/W066
7738_8#62		*E. coli*	STEC						
7738_8#63	15302	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2008	Germany	
7738_8#64	15967	*E. coli*	STEC	cattle	healthy	Free University Berlin, IMT	2008	Germany	
7738_8#65		*E. coli*	aEPEC						
7738_8#66		*E. coli*	aEPEC						
7738_8#67		*E. coli*	aEPEC						
7738_8#68	19996	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_8#69	20165	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_8#70	20226	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_8#71	20239	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
7738_8#72	27058	*E. coli*		pig		Free University Berlin, IMT	2010	Germany	
7748_7#13		*E. coli*	aEPEC	human					
7748_7#16		*E. coli*	aEPEC	human					
7748_7#18		*E. coli*	aEPEC	human					
7748_7#19		*E. coli*	aEPEC	human					
7748_7#21		*E. coli*	aEPEC	human					
7748_7#27		*E. coli*	aEPEC	human					
7748_7#32		*E. coli*	aEPEC	human					
7748_7#36		*E. coli*	aEPEC	human					
7748_7#6		*E. coli*	aEPEC	human					
7790_1#50		*E. coli*	aEPEC	human					
7790_1#53		*E. coli*	aEPEC	human					
7790_1#60		*E. coli*	aEPEC	human					
7790_1#79		*E. coli*	aEPEC	human					
7790_1#91		*E. coli*	aEPEC	human					
7853_6#1		*E. coli*	EPEC	human					
7853_6#18		*E. coli*	EPEC	human					
7853_6#21		*E. coli*	EPEC	human					

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
7853_6#27		*E. coli*	EPEC	human					
7853_6#29		*E. coli*	EPEC	human					
7853_6#30		*E. coli*	EPEC	human					
7853_6#45		*E. coli*	EPEC	human					
7853_6#7		*E. coli*	EPEC	human					
7853_7#50		*E. coli*	EPEC	human					
7853_7#52		*E. coli*	EPEC	human					
7853_7#54		*E. coli*	EPEC	human					
7853_7#56		*E. albertii*							
7853_7#62		*E. coli*		human					
7853_7#65		*E. coli*	EPEC	human					
7853_7#66		*E. coli*	EPEC	human					
7853_7#69		*E. coli*	EPEC	human					
7853_7#70		*E. coli*	EPEC	human					
7853_7#73		*E. coli*	EPEC	human					
7853_7#75		*E. coli*	EPEC	human					
7853_7#76		*E. coli*	EPEC	human					
7853_7#78		*E. coli*	EPEC	human					
7853_7#81		*E. coli*	EPEC	human					
7853_7#83		*E. coli*	EPEC	human					
7907_8#27		*E. coli*	EPEC	human					
7907_8#29		*E. coli*	EPEC	human					
7907_8#7		*E. coli*	EPEC	human					
8001_7#11		*E. coli*	STEC						
8001_7#12		*E. coli*	STEC						
8001_7#13	3749	*E. coli*	STEC						
8001_7#14	3753	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	1990	USA (MN)	900105
8001_7#15	3778	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Germany	RW2087
8001_7#19		*E. coli*	STEC						
8001_7#21	3846	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Belgium	561/379S
8001_7#22		*E. coli*	STEC						
8001_7#23		*E. coli*	STEC						
8001_7#24	3870	*E. coli*	STEC	cattle	healthy	Justus-Liebig-University Giessen, IHIT	1995	Germany	GS0845-1
8001_7#25	3874	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Belgium	551/344S

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
8001_7#27		*E. coli*	STEC						
8001_7#28	3898	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Belgium	555/370S
8001_7#29		*E. coli*	STEC						
8001_7#3	1947	*E. coli*	STEC			Free University Berlin, IMT	1999	Germany	
8001_7#30			STEC						
8001_7#31	4271	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	1889	Germany	410/89
8001_7#35	4642	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#37	4644	*E. albertii*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#4	2141	*E. coli*	EPEC	bird	Enteritis	Justus-Liebig-University Giessen, IHIT	1994	Germany	311/94
8001_7#42	4649	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#46	4653	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#48	4655	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#5	2247	*E. coli*		bird	Septicemia	State-VetMed Investigation Office Aulendorf	1999	Germany	Z205
8001_7#51	4658	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#52	4659	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#54	4661	*E. coli*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#56	4663	*E. albertii*		human		Free University Berlin, IMT	2001	Brazil	
8001_7#59	5067	*E. coli*		dog	Vaginitis	Justus-Liebig-University Giessen, IHIT	1996	Germany	319/96
8001_7#60	5068	*E. coli*		dog	Enteritis	Justus-Liebig-University Giessen, IHIT	1996	Germany	320/96
8001_7#64	5072	*E. coli*		dog	Obstipation	Justus-Liebig-University Giessen, IHIT	1996	Germany	324/96
8001_7#65	5073	*E. coli*		dog	Enteritis	Justus-Liebig-University Giessen, IHIT	1996	Germany	325/96
8001_7#66	5074	*E. coli*		dog	Enteritis	Justus-Liebig-University Giessen, IHIT	1996	Germany	326/96
8001_7#67	5075	*E. coli*		dog	Enteritis	Justus-Liebig-University Giessen, IHIT	1996	Germany	327/96
8001_7#68	5076	*E. coli*		dog	Enteritis	Justus-Liebig-University Giessen, IHIT	1996	Germany	328/96
8001_7#7	3153	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	1989	Germany	0067
8001_7#73	5976	*E. alvei*		human	Enteritis	University of Helsinki, Faculty of VetMed	1990/91	Bangladesh	19982
8001_7#74	5977	*E. alvei*		human	Enteritis	University of Helsinki, Faculty of VetMed	1990/91	Bangladesh	9194
8001_7#75	5978	*E. alvei*		human	Enteritis	University of Helsinki, Faculty of VetMed	1990/91	Bangladesh	10457
8001_7#8	3168	*E. coli*	STEC			Justus-Liebig-University Giessen, IHIT	2000	Germany	2087
8001_7#9	3174	*E. coli*	aEPEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2000	Germany	RW2174
8016_2#1	15301	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2008	Germany	
8016_2#10	19615	*E. coli*	STEC	cattle		Free University Berlin, IMT	2010	Germany	
8016_2#11	19636	*E. coli*	STEC	cattle		Free University Berlin, IMT	2010	Germany	
8016_2#12	19648	*E. coli*	STEC	cattle		Free University Berlin, IMT	2010	Germany	

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
8016_2#13	19947	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2009	Germany	
8016_2#14	20217	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2009	Germany	
8016_2#15	20245	*E. coli*	aEPEC	dog	UTI	VetMed Laboratory Ludwigsburg	2009	Germany	VB 996164
8016_2#16	20337	*E. coli*	STEC	cattle		Free University Berlin, IMT	2009	Germany	
8016_2#19	21664	*E. coli*		human		Imperial College London, Dept. Life Sc.	2010	Spain	JB-4
8016_2#2	15911	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2008	Germany	P8054/02
8016_2#3	15915	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2008	Germany	0739/03
8016_2#4	15922	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2008	Germany	P7404/03-1
8016_2#46	21183	*E. coli*		human	sick	Charité Berlin, Institute of Hygiene	2010	Germany	EF1335
8016_2#49	17887	*E. coli*		horse	WI	VetMed Laboratory Ludwigsburg	2008	Germany	VB 964041.2
8016_2#5		*E. coli*	STEC						
8016_2#51	22073	*E. coli*		cat	UTI	VetMed Laboratory Ludwigsburg	2010	Germany	VB 998118.1
8016_2#53	21200	*E. coli*		human	sick	Charité Berlin, Institute of Hygiene	2010	Germany	EF1353
8016_2#56	18354	*E. coli*		horse	sick	VetMed Laboratory Ludwigsburg	2009	Germany	VB 977411.2
8016_2#6	15981	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	2008	Germany	P7278/07
8016_2#7	15987	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	2008	Germany	P3146/08-1
8016_2#8	15988	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	2008	Germany	P3146/08-2
8016_2#9	15989	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	2008	Germany	P3146/08-3
9352_8#10	7858	*E. coli*		human	healthy	Michigan State University, NFSTC	1979	USA (NY)	ECOR-02
9352_8#11	7859	*E. coli*		dog	healthy	Michigan State University, NFSTC		USA (MA)	ECOR-03
9352_8#12	7860	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-04
9352_8#13	7861	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-05
9352_8#14	7862	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-06
9352_8#15	7863	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-07
9352_8#16	7864	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-08
9352_8#17	7865	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-09
9352_8#18	7866	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-10
9352_8#19	7867	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-11
9352_8#20	7868	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-12
9352_8#21	7869	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-13
9352_8#22	7870	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-14
9352_8#23	7871	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-15
9352_8#24	7872	*E. coli*		leopard	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-16
9352_8#25	7873	*E. coli*		pig	healthy	Michigan State University, NFSTC		Indonesia	ECOR-17

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
9352_8#26	7874	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-18
9352_8#27	7875	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-19
9352_8#28	7876	*E. coli*		cattle	healthy	Michigan State University, NFSTC		Bali	ECOR-20
9352_8#29	7877	*E. coli*		cattle	healthy	Michigan State University, NFSTC		Bali	ECOR-21
9352_8#30	7878	*E. coli*		bull	healthy	Michigan State University, NFSTC		Bali	ECOR-22
9352_8#31	7879	*E. coli*		elephant	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-23
9352_8#32	7880	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-24
9352_8#33	7881	*E. coli*		dog	healthy	Michigan State University, NFSTC		USA (NY)	ECOR-25
9352_8#34	7882	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (MA)	ECOR-26
9352_8#35	7883	*E. coli*		giraffe	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-27
9352_8#36	7884	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-28
9352_8#37	7885	*E. coli*		rat	healthy	Michigan State University, NFSTC		USA (NV)	ECOR-29
9352_8#38	7886	*E. coli*		bison	healthy	Michigan State University, NFSTC		Canada	ECOR-30
9352_8#39	7887	*E. coli*		leopard	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-31
9352_8#40	7859	*E. coli*		giraffe	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-32
9352_8#41	7860	*E. coli*		sheep	healthy	Michigan State University, NFSTC		USA (CA)	ECOR-33
9352_8#42	7861	*E. coli*		dog	healthy	Michigan State University, NFSTC		USA (MA)	ECOR-34
9352_8#43	7862	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-35
9352_8#44	7863	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-36
9352_8#45	7864	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-37
9352_8#46	7865	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-38
9352_8#47	7866	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-39
9352_8#48	7867	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-40
9352_8#49	7746	*E. coli*		human	healthy	Michigan State University, NFSTC	1982	Tonga	ECOR-41
9352_8#50	7869	*E. coli*		human	healthy	Michigan State University, NFSTC	1979	USA (MA)	ECOR-42
9352_8#51	7870	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-43
9352_8#52	7871	*E. coli*		cougar	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-44
9352_8#53	7872	*E. coli*		pig	healthy	Michigan State University, NFSTC		Indonesia	ECOR-45
9352_8#54	7873	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-46
9352_8#55	7874	*E. coli*		sheep	healthy	Michigan State University, NFSTC		New Guinea	ECOR-47
9352_8#56	7875	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-49
9352_8#57	7876	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-50
9352_8#58	7877	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (MA)	ECOR-51
9352_8#59	7747	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-52

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
9352_8#6	7878	*E. coli*		human	UTI	Max-Planck-Institute for Infection Biology		Sweden	ECOR-48
9352_8#60	7879	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-53
9352_8#61	7880	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-54
9352_8#62	7881	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-55
9352_8#63	7882	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-56
9352_8#64	7883	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-57
9352_8#65	7884	*E. coli*		lion	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-58
9352_8#66	7885	*E. coli*		human	healthy	Michigan State University, NFSTC	1979	USA (MA)	ECOR-59
9352_8#67	7886	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-60
9352_8#68	7887	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-61
9352_8#69	7858	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-62
9352_8#70	7860	*E. coli*		human	healthy	Michigan State University, NFSTC		Sweden	ECOR-63
9352_8#71	7861	*E. coli*		human	UTI	Michigan State University, NFSTC		sweden	ECOR-64
9352_8#9	7863	*E. coli*		human	healthy	Michigan State University, NFSTC		USA (IA)	ECOR-01
9425_1#1	7864	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-65
9425_1#2	7865	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-66
9425_1#25									
9425_1#26									
9425_1#27									
9425_1#28									
9425_1#29									
9425_1#3	7923	*E. coli*		goat	healthy	Michigan State University, NFSTC		Indonesia	ECOR-67
9425_1#30									
9425_1#31									
9425_1#32									
9425_1#33									
9425_1#34									
9425_1#35									
9425_1#36									
9425_1#37									
9425_1#38									
9425_1#39									
9425_1#4	7924	*E. coli*		giraffe	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-68
9425_1#40									

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
9425_1#41			STEC						
9425_1#42									
9425_1#43									
9425_1#44									
9425_1#45									
9425_1#46									
9425_1#47									
9425_1#5	7925	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-69
9425_1#6	7926	*E. coli*		monkey	healthy	Michigan State University, NFSTC		USA (WA)	ECOR-70
9425_1#7	7927	*E. coli*		human	ABU	Michigan State University, NFSTC		Sweden	ECOR-71
9425_1#8	7928	*E. coli*		human	UTI	Michigan State University, NFSTC		Sweden	ECOR-72
ae005174		*E. coli*	EHEC	human	HC	University of Wisconsin, Genome Center, ref. Genome	1982	USA (MI)	EDL933
ae014075		*E. coli*	UPEC	human	UTI	University of Wisconsin, Genome Center, ref. Genome	2002	USA (MD)	CFT073
am946981		*E. coli*	COM	human	healthy	Austrian Center of Bioph., ref. Genome	2008	Austria	BL21(DE3)
ap009048		*E. coli*	COM	human	healthy	NAIST, ref. Genome	1950	Japan	K12-W3110
ap009240		*E. coli*	COM	human	healthy	Masahira Hattori University, ref. Genome	2008	Japan	SE11
ap009378		*E. coli*	COM	human	healthy	Masahira Hattori University, ref. Genome	2008	Japan	SE15
ap010953		*E. coli*	EHEC	human	Diarrhea	Masahira Hattori University, ref. Genome	2001	Japan	11368
ap010958		*E. coli*	EHEC	human	Diarrhea	Masahira Hattori University, ref. Genome	2001	Japan	12009
ap010960		*E. coli*	EHEC	human	Diarrhea	Masahira Hattori University, ref. Genome	2001	Japan	11128
ap012030		*E. coli*	K-12				1980		ATCC 33849
ba000007		*E. coli*	EHEC	human	HC	Miyazaki Medical College, ref. Genome	1996	Japan	Sakai
cp000243		*E. coli*	UPEC	human	UTI	Washington University, Genome Center, ref. Genome	2001	USA	UTI89
cp000247		*E. coli*	UPEC	human	UTI	University Göttingen, IMG, ref. Genome	1982	Germany	536
cp000468		*E. coli*	APEC	chicken	Septicemia	Iowa State University, VetMed, ref. Genome	2003	USA (IA)	APECO1
cp000800		*E. coli*	ETEC	human	Enteritis	J. Craig Venter Institute, ref.Genome	< 1972	Bangladesh	E24377A
cp000802		*E. coli*	COM	human	healthy	J. Craig Venter Institute, ref.Genome	1977	USA	HS
cp000819		*E. coli*	COM	human	healthy	Korea Research Institute of Biosc., ref. Genome	1950	Korea	REL606
cp000946		*E. coli*	COM	human	healthy	J. Craig Venter Institute, ref.Genome			ATCC 8739
cp000948		*E. coli*	COM	human	healthy	University of Wisconsin, Genome Center, ref. Genome	1950	USA	K12-DH10B
cp000970		*E. coli*	COM	environ		J. Craig Venter Institute, ref.Genome	2005	USA (SC)	SMS-3-5
cp001164		*E. coli*	EHEC	human	Enteritis	J. Craig Venter Institute, ref.Genome	2006	Spain	EC4115
cp001396		*E. coli*	COM	human	healthy	J. Craig Venter Institute, ref.Genome	1950	USA	K12-BW2952
cp001665		*E. coli*	COM	human	healthy	U.S. Dept. of Energy Joint Genome Inst., ref. Genome	2009	USA	BL21

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
cp001671		*E. coli*		human	ABU	University Goettingen, Reference Genome (NCBI)	1990	Germany	ABU 83972
cp001846		*E. coli*	aEPEC	human	Enteritis	Nankai University, TEDA, ref. Genome	2003	Germany	CB9615
cp001855		*E. coli*	AIEC	human	IBD - CD	Public Health Agency of Canada, ref. Genome	2010		NRG 857C
cp001925		*E. coli*	EHEC	human	HUS	CDC, China, ref. Genome	1999	China	Xuzhou21
cp001969		*E. coli*	NMEC	human	NM	Novartis Vaccines & Diagnostics, ref. Genome	1976	Finland	IHE3034
cp002167		*E. coli*	AIEC	human	IBD - CD	Research & Testing Laboratory USA, ref. Genome	2007	Canada	UM146
cp002185		*E. coli*	environ	environ		State University of New Jersey, ATCC, ref. Genome	1943	USA	ATCC 9637
cp002211		*E. coli*	UPEC	human	UTI	Nankai University, TEDA, ref. Genome	2005	USA	D i2
cp002291		*E. coli*				Nankai University, TEDA, ref. Genome	1985	USA	P12b
cp002516		*E. coli*	environ			University of Florida, Dept. Microbiology, ref. Genome	1991	USA	ATCC 55124
cp002729		*E. coli*	ETEC	pig	Diarrhea	University of Minnesota, VetBiomed Sc., ref. Genome	2007	USA (MN)	UMNK88
cp002797		*E. coli*	UPEC	human	Prostatitis	University of Hyderabad, Dept. Biotech., ref. Genome	2010	India (Pune)	NA114
cp003034		*E. coli*	NMEC	human	NM	Johns Hopkins University, Pediatrics, ref. Genome	2006	USA	CE10
cp003109		*E. coli*	EPEC	human	Enteritis	CA Dept. of Public Health, ref. Genome	1974	USA (CA)	RM12579
cu928145		*E. coli*	EAEC	human	Diarrhea	University Pierre et Marie Curie, IPG, ref. Genome	2001	Central Africa	55989
cu928160		*E. coli*	COM	human	healthy	University Pierre et Marie Curie, IPG, ref. Genome	1980	France	IAI1
cu928161		*E. coli*	NMEC	human	NM	University Pierre et Marie Curie, IPG, ref. Genome	1999	France	S88
cu928162		*E. coli*	COM	human	healthy	University Pierre et Marie Curie, IPG, ref. Genome	2000	France	ED1a
cu928163		*E. coli*	UPEC	human	UTI	University Pierre et Marie Curie, IPG, ref. Genome	1999	USA (MN)	UMN026
cu928164		*E. coli*	UPEC	human	UTI	University Pierre et Marie Curie, IPG, ref. Genome	1980	France	IAI39
ea_b090		*E. albertii*		bird	healthy	CDC, USA, ref. Genome	2001	Australia	B090
ea_b156		*E.albertii*		bird	healthy	CDC, USA, ref. Genome	2001	Australia	B156
ec_b088		*E. coli*		bird		Gordon Lab Australia, ref. Genome		Australia	B088
ec_e1118		*E. coli*		environ		Gordon Lab Australia, ref. Genome		Australia	E1118
ec_e1492		*E. coli*		environ		Broad Institute			E1492
ec_h442		*E. coli*		human					H442
ec_h605		*E. coli*		human		Gordon Lab Australia, ref. Genome		Australia	H605
ec_m863		*E. coli*		human				Australia	M863
ec_ta004		*E. coli*		human				Australia	TA004
ec_tw10509		*E. coli*		human				Guinea	TW10509
ef_b253		*E. fergusonii*							B253
fm180568		*E. coli*	EPEC	human	Diarrhea	University of Maryland, School of Medicine	1978	USA (MD)	E2348/69
fn649414		*E. coli*	ETEC	human	Diarrhea	Wellcome Trust Sanger Institute, ref. Genome	< 1972	Bangladesh	H10407
u00096		*E. coli*	COM	human	healthy	University of Wisconsin, Genome Center, ref. Genome	1997	USA	MG1655

Sanger No. / NCBI No.	IMT No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.
	15993	*E. coli*	aEPEC	cattle		Free University Berlin, IMT	2008	Germany	
	24405	*E. coli*	aEPEC	ruminant		University of Veterinary Medicine Hannover	2010	Germany	80538-9x/5
	15938	*E. coli*	STEC	cattle		Justus-Liebig-University Giessen, IHIT	2008	Germany	P4501/05-1
	15937	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	2008	Germany	P3675/05-1
	18537	*E. coli*	STEC	food		University Hohenheim, Dept. Food Microbiology	2009	Germany	3000
	21678	*E. coli*	aEPEC	human		Imperial College London, Dept. Life Sc.	2010	Australia	LH-5
	3800	*E. coli*	STEC	human		Justus-Liebig-University Giessen, IHIT	1986	USA (ID)	3030A-86
	477	*E. coli*	STEC	cattle	Enteritis	Justus-Liebig-University Giessen, IHIT	1998	Germany	413/W003

IMT: Institut für Mikrobiologie und Tierseuchen Pathotype: ExPEC: Extraintestinal pathogenic *E. coli*, APEC: Avian pathogenic *E. coli,* NMEC: Neonatal meningitis *E. coli*, UPEC: Uropathogenic *E. coli*; InPEC: Intestinal pathogenic *E. coli,* EAEC: Enteroaggregative *E. coli*, EHEC: Enterohaemorrhagic *E. coli*, EIEC: Enteroinvasive *E. coli*, EPEC: Enteropathogenic *E. coli*, aEPEC: atypical Enteropathogenic *E. coli,* ETEC: Enterotoxigenic *E. coli*, STEC: Shiga-Toxin-producing *E. coli*; COM: Commensal; Environ: Environment; Disease: ABU: Asymtomatic bacteriuria, HC: Hemorrhagic colitis, HUS: Hemolytic uremic syndrome, NM: Neonatal meningitis, UTI: Urinary tract infection, WI: Wound infection, IBD-CD: Inflammatory Bowel Disease – Chron's Disease

Table A 2: Collected NCBI database

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
AB011549	*E. coli*	EHEC	human	HUS	Osaka University, Research Institute for Microbial Diseases	1998	Japan	Sakai	Plasmid (pO157)
AB011677	*E. coli*				National Institute of Health, Department of Pathology, Tokyo	1994	Japan		Gene (LT)
AB048223	*E. coli*	STEC	deer		Souichi Makino Obihiro University, Dept. of veterinary microbiology	2000	Japan	#S-3	Gene (*stx* 2AB)
AB048224	*E. coli*	STEC	deer		Souichi Makino Obihiro University, Dept. of veterinary microbiology	2000	Japan	#S-5	Gene (*stx* 2AB)
AB048227	*E. coli*	STEC	deer		Souichi Makino Obihiro University, Dept. of veterinary microbiology	2000	Japan	#S-8	Gene (*stx* 2AB)
AB048236	*E. coli*	STEC	sheep		Souichi Makino Obihiro University, Dept. of veterinary microbiology	2000	Japan	#HI-11	Gene (*stx* 2AB)
AB083044	*E. coli*	STEC			Aichi Prefectural Institute of Public Health, Nagoya Kitaku	2003	Japan	AI2001/52	Gene (*stx* 1AB)
AB255435	*E. coli*	EAEC			Niigata University, Department of Infectious Disease Control		Japan	DIJ1	Plasmid (pO86A1)
AB355659	*E. coli*	EPEC	human	sick	Tadasuke Ooka University of Miyazaki, Dept. Infectious Disease	2007	Brazil	ICC223	Gene (*tir*)
AB426049	*E. coli*	EPEC	human	Diarrhea	Stanford University, Dept. of Microbiology and Immunology	1983	USA (MD)	B171-8	PAI (GEI0.81)
AB426060	*E. coli*	EAEC	human	Diarrhea	University Pierre et Marie Curie, Atelier de BioInformatique, Paris	2001	Central Africa	GEI3.10	PAI (GEI3.10)
AB472687	*E. coli*	STEC	human	sick	Yoshiki Etoh Fukuaka Institute of Health an Environmental Sciences	2008	Japan	F08/101-31	Gene (*stx* 2AB)
AB472834	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-5	Gene (*cdt Va-c*)
AB472835	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-6	Gene (*cdt Va-c*)
AB472839	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-10	Gene (*cdt Va-c*)
AB472840	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-11	Gene (*cdt Va-c*)
AB472857	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-13	Gene (*cdt Va-c*)
AB472860	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-16	Gene (*cdt Va-c*)
AB472861	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-17	Gene (*cdt Va-c*)
AB472870	*E. coli*		human	Diarrhea	Atsushi Hinenoya Osaka University, School of Life	2008	Japan	AH-26	Gene (*cdt Va-c*)
AB646137	*E. coli*	EHEC			University of Miyazaki, Frontier Science Reserch Center	2002	Japan	990281	Gene (*hly* E)
AB646138	*E. coli*	EPEC			University of Miyazaki, Frontier Science Reserch Center	2001	Japan	F57076	Gene (*hly* E)
AB646152	*E. coli*	EHEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 05091870	Gene (*hly* E)
AB646154	*E. coli*	EHEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 05091856	Gene (*hly* E)
AB646155	*E. coli*	EHEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 05091863	Gene (*hly* E)
AB646156	*E. coli*	EHEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 05091862	Gene (*hly* E)
AB646158	*E. coli*	EIEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 05091045	Gene (*hly* E)
AB646162	*E. coli*	ETEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 0509796	Gene (*hly* E)
AB646163	*E. coli*	ETEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 0509343	Gene (*hly* E)
AB646164	*E. coli*	ETEC			Osaka University, Research Institute for Microbial Diseases	2011	Japan	RIMD 0509792	Gene (*hly* E)
AB702969	*E. coli*	ETEC	human	sick	National Center for Global Health and Medicine, Tokyo	2012	Japan	4266 delta cssB::Km	Plasmid (pCss165Kan)
AB839651	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B1	Gene (*cdt Va-c*)
AB839652	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B3	Gene (*cdt Va-c*)
AB839653	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B5	Gene (*cdt Va-c*)
AB839654	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B7	Gene (*cdt Va-c*)
AB839657	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B21	Gene (*cdt Va-c*)
AB839658	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B43	Gene (*cdt Va-c*)
AB839659	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B49	Gene (*cdt Va-c*)
AB839660	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B55	Gene (*cdt Va-c*)
AB839661	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B56	Gene (*cdt Va-c*)
AB839662	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B61	Gene (*cdt Va-c*)
AB839663	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B65	Gene (*cdt Va-c*)
AB839664	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B68	Gene (*cdt Va-c*)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
AB839665	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B78	Gene (*cdt Va-c*)
AB839666	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B87	Gene (*cdt Va-c*)
AB839667	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B88	Gene (*cdt Va-c*)
AB839668	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B91	Gene (*cdt Va-c*)
AB839669	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B98	Gene (*cdt Va-c*)
AB839670	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	B100	Gene (*cdt Va-c*)
AB839671	*E. coli*		cattle	healthy	Atsushi Hinenoya Osaka University, School of Life	2013	Japan	S26	Gene (*cdt Va-c*)
AE005174	*E. coli*	EHEC	human	HC	University of Wisconsin, Genetics Laboratory, Madison	1998	USA (MI)	EDL933	Genome
AE005674	*S. flexneri*		human	Shigellosis	State Key Laboratory for Molecular Virology & Genetics, Beijing	1984	China	2a str. 301	Genome
AE014073	*S. flexneri*		human		University of Wisconsin, Genetics Laboratory, Madison	1964	USA	2a str. 2457T	Genome
AE014075	*E. coli*	UPEC	human	UTI	University of Wisconsin, Genetics Laboratory, Madison	2002	USA (WI)	CFT073	Genome
AEKA000000	*E. coli*	ETEC	human	sick	Georgia Institute of Technology, School of Biology and Bioinformatics	1987	India	TW10509	Genome
AF022236	*E. coli*	EPEC	human	Diarrhea	University of Maryland, Center for Vaccine Development, Baltimore	1978	USA (MD)	E2348/69	PAI
AF025311	*E. coli*	STEC	human	HUS	Women's and Children's Hospital, Microbiology, South Australia	1997	Australia	O111:H-	PAI (intimin receptor)
AF043471	*E. coli*	EHEC			University of Guelph, Ontario Veterinary College, Pathobiology	1998	Canada	EC920006	Gene (*ehx* A)
AF047364	*S. flexneri*		human	Shigellosis	Institut Pasteur, Molecular Pathogenic Microbiology, Paris	1998	France	5a str. M90T	Gene (*vir* A)
AF056581	*E. coli*	EAEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	E. coli 042	Gene (*pet*)
AF070067	*E. coli*	STEC	human	HUS	Women & Children Hospital, Molecular Microbiology, NorthAdelaide	1998	Australia	95SF2	Gene (*tir*)
AF074613	*E. coli*	EHEC	human	HC	University of Wisconsin, Genetics Laboratory, Madison	1998	USA (MI)	EDL933	Plasmid (pO157)
AF097644	*E. coli*	EAEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	E. coli 042	Toxin set
AF125993	*E. coli*	EHEC	human	HUS	University of British Columbia, Biotechnology Lab, Vancouver	1999	Canada	86/24	Gene (*tir*)
AF153317	*S. dysenteriae*				Massachusetts General Hospital, Infectious Disease Division, Boston	1999	USA (MA)	str. 1	Gene (*stx* 1AB)
AF159462	*E. coli*	EHEC			Royal Children's Hospital, Microbiological Research, Parkville	1999	Australia		Gene (*efa* 1/*lif* A)
AF160993	*E. coli*	STEC	human	healthy	Umea University, Microbiology	1999	Sweden	ECOR 5	Gene (EAST1)
AF160996	*E. coli*	STEC	human	healthy	Umea University, Microbiology	1999	Sweden	ECOR 10	Gene (EAST1)
AF160999	*E. coli*	STEC	giraffe	healthy	Umea University, Microbiology	1999	Sweden	ECOR 32	Gene (EAST1)
AF161001	*E. coli*	STEC	sheep	healthy	Umea University, Microbiology	1999	Sweden	ECOR 33	Gene (EAST1)
AF161002	*E. coli*	STEC	human	ABU	Umea University, Microbiology	1999	Sweden	ECOR 71	Gene (EAST1)
AF200692	*S. flexneri*		human		Monash University, Department of Microbiology	2001	Australia	2a she -PAI	PAI (she)
AF289092	*E. coli*	UPEC	human	UTI	University of Maryland, Microbiology and Immunology	2000	USA (MD)	CFT073	Gene (*sat*)
AF297061	*E. coli*	EPEC			University of Maryland, Center for Vaccine Development, Baltimore		USA		Gene (*esp* C)
AF348706	*S. flexneri*		human		University of Wisconsin, Genetics Laboratory, Madison	2001	USA	5a str. WR501	Plasmid (pWR501)
AF386526	*S. flexneri*		human	Shigellosis	State Key Laboratory for Molecular Virology & Genetics, Beijing	1984	China	2a str. 301	Plasmid (pCP301)
AF399919	*E. coli*	STEC	human	HUS	Adelaide University, Department Of Molecular Biosciences	2001	Australia	98NK2	Plasmid (pO113)
AF401292	*E. coli*	EHEC	human		University Wuerzburg, Institute of Hygiene and Microbiology	1996	Germany	3072/96	Plasmid (pSFO157)
AF411067	*E. coli*	EAEC	human	Diarrhea	Institut Pasteur, Bacteriology and Mycology, Paris	2001	France	55989	Gene (EAST1)
AF461173	*E. coli*	STEC	food		Dept. of Micobiology, National Institute of Health, Korea	2001	Korea	FD930	Gene (*stx* 2AB)
AF483828	*E. coli*		cattle	Diarrhea	North Dakota State University, Veterinary & Microbiological Sc	2002	USA	96-1913	Gene (*cnf* 1)
AF483829	*E. coli*		cattle	Diarrhea	North Dakota State University, Veterinary & Microbiological Sc	2002	USA	5383-2	Gene (*cnf* 1)
AF497476	*E. coli*				National Institute of Research Agriculture Toulouse	2002	France		Gene (*efa* 1/*lif* A)
AJ133705	*E. coli*				University of Maryland, Dept. of Infectious Diseases, Baltimore	1999	USA		Gene (*efa* 1/*lif* A)
AJ223631	*E. coli*		human	WI	Institute of Molecular Biological Sciences, Amsterdam	1998	Netherlands	EB1	Plasmid (pColV-K30)
AJ238954	*E. coli*		human	HC	Hospital Ramony Cajal, Madrid	1999	Spain	CECT 4267	Gene (*hly* E)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
AJ278144	*E. coli*	STEC	human	Diarrhea	University Wuerzburg, Bioinformatics Center, Germany	1996/00	Germany	4797/97	PAI
AJ303141	*E. coli*	STEC	cattle		Free University of Berlin, Inst. of Microbiology and Epizootics	2001	Germany	RW1374	PAI (I)
AJ459584	*E. coli*	EHEC	human	sick	University Muenster, Institute of Hygiene	1989	Germany	493/89	Gene (*efa* 1/*lif* A)
AJ488511	*E. coli*	UPEC	human	UTI	University Goettingen, Institute of Microbiology	1982	Germany	536	PAI (I)
AJ494981	*E. coli*	UPEC	human	UTI	University Goettingen, Institute of Microbiology	1982	Germany	536	PAI (II)
AJ508930	*E. coli*	EHEC	human	sick	University Muenster, Institute of Hygiene	1989	Germany	493/89	Gene (*cdt Va-c*)
AJ633129	*E. coli*	EPEC	human	Diarrhea	University Muenster, Institute of Hygiene	2002	Germany	0181-6/86	PAI (LEE)
AJ868113	*E. coli*	ETEC	human	Diarrhea	London School of Medicine, Inst. of Cell & Molecular Science	2004	UK	WS-1858B	Gene (STb)
AL391753	*S. flexneri*		human		Institut Pasteur, Genotyping of Pathogens, Paris	2000	France	5a str. WR100	Plasmid (pWR100)
AM229678	*E. coli*	NMEC	human	NM	University Wuerzburg, Inst. f. Moleculare Infectrion Biology	1976	Finland	IHE3034	Gene (*clb* A-Q)
AM230662	*E. coli*	EHEC	human	HUS	University Muenster, Institute of Hygiene	2007	Germany	3385/00	Gene (*stx* 1AB)
AM230663	*E. coli*	EHEC	human	HUS	University Muenster, Institute of Hygiene	2007	Germany	04-06263	Gene (*stx* 1AB)
AM261284	*E. coli*	COM			Institute of Microbiology, Videnska	2006	Czech Rep.	A0 34/86	PAI (pO83)
AM690759	*E. coli*		human	ABU	Lund University, Institute for Molecular Infection biology	1974	Sweden	ABU 83972	Gene (*hly* A)
AM690760	*E. coli*		human	ABU	University Wuerzburg, Institute for Molecular Infection biology	1974	Sweden	ABU 27	Gene (*hly* A)
AM690761	*E. coli*		human	ABU	University Wuerzburg, Institute for Molecular Infection biology	1974	Sweden	ABU 37	Gene (*hly* A)
AM904726	*E. coli*	STEC			Norwegian Institute of Public Health, Infectious Diseases Control	2007	Norway	FHI 1106-1092	Gene (*stx* 2AB)
AP009048	*E. coli*	COM			Nara Institute of Science and Technology, Ikoma	2006	Japan	K-12 W3110	Genome
AP010910	*E. coli*	ETEC	human	Diarrhea	Fujita Health University, Department of Microbiology	2008	Japan	H10407	Plasmid (pEntH10407)
AP010953	*E. coli*	EPEC			Masahira Hattori University of Tokyo, Frontier Sciences	2008	Japan	11368	Genome
AP010954	*E. coli*	EPEC			Masahira Hattori University of Tokyo, Frontier Sciences	2008	Japan	11368	Plasmid (pO26_1)
AP010958	*E. coli*	EHEC	human	Diarrhea	Masahira Hattori University of Tokyo, Frontier Sciences	2001	Japan	12009	Genome
AP010959	*E. coli*	EPEC			Masahira Hattori University of Tokyo, Frontier Sciences	2008	Japan	12009	Plasmid (pO103)
AP010960	*E. coli*	EPEC			Masahira Hattori University of Tokyo, Frontier Sciences	2008	Japan	11128	Genome
AP010963	*E. coli*	EPEC			Masahira Hattori University of Tokyo, Frontier Sciences	2008	Japan	11128	Plasmid (pO111_3)
AP014654	*E. coli*	ETEC	human	Diarrhea	Osaka City University, Food and Human Health Sciences	2014	Japan	O169:H41	Plasmid (pEntYN10)
AP014855	*E. albertii*		bird		Tetsuya Hayashi Kyushu University, Department of Bacteriology	2015	Japan	NIAH_Bird_3	Genome
AP014856	*E. albertii*		human	Diarrhea	Tetsuya Hayashi Kyushu University, Department of Bacteriology	2015	Japan	CB9786	Genome
AP014857	*E. albertii*		human	Diarrhea	Tetsuya Hayashi Kyushu University, Department of Bacteriology	2015	Japan	EC06-170	Genome
AY029075	*S. boydii*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2431	Gene (*ipg* D)
AY029076	*S. dysenteriae*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ3684	Gene (*ipg* D)
AY029077	*S. dysenteriae*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ3474	Gene (*ipg* D)
AY029078	*S. dysenteriae*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ3472	Gene (*ipg* D)
AY029079	*S. flexneri*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2439	Gene (*ipg* D)
AY029080	*S. flexneri*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2441	Gene (*ipg* D)
AY029081	*S. flexneri*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2443	Gene (*ipg* D)
AY029082	*S. flexneri*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2445	Gene (*ipg* D)
AY029083	*S. flexneri*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2448	Gene (*ipg* D)
AY029085	*S. flexneri*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2449	Gene (*ipg* D)
AY029087	*S. dysenteriae*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ2436	Gene (*ipg* D)
AY029089	*S. dysenteriae*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ3475	Gene (*ipg* D)
AY029090	*S. boydii*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2426	Gene (*ipg* D)
AY029092	*S. boydii*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ3689	Gene (*ipg* D)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
AY029093	*S. boydii*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2720	Gene (*ipg* D)
AY029094	*S. boydii*		human		Laboratory of Sante Publique, Quebec	2000	Canada (QC)	LSPQ2432	Gene (*ipg* D)
AY029095	*S. boydii*		human		Canadian National Laboratory for Enteric Pathogens	2000	Canada (QC)	LSPQ3690	Gene (*ipg* D)
AY029096	*E. coli*	EIEC	human		State Health Laboratory, Perth	1997	Argentinia	C906-91	Gene (*ipg* D)
AY029097	*E. coli*	EIEC	human		International Escherichia and Klebsiella Centre, Denmark	1997	USA	C499-89	Gene (*ipg* D)
AY029098	*S. sonnei*		human		American Type Culture Collection (ATCC), Manassas		USA (VA)	ATCC 22930	Gene (*ipg* D)
AY098990	*E. coli*	EIEC	human		University of Clinical Farmaceutics, Clinical anlalysis & Toxicology	2002	Brazil	FBC124-13	Gene (*ipg* D)
AY128535	*E. coli*	EPEC	cattle		National Institute of Research Agriculture Toulouse	2003	France	CF11201	Gene (*cif*)
AY128536	*E. coli*	EHEC	human		National Institute of Research Agriculture Toulouse	2003	France	ED-31	Gene (*cif*)
AY128537	*E. coli*	EPEC	human		National Institute of Research Agriculture Toulouse	2003	France	EF-26	Gene (*cif*)
AY128538	*E. coli*	EHEC	human		National Institute of Research Agriculture Toulouse	2003	France	PMK5	Gene (*cif*)
AY128539	*E. coli*	EHEC	human		National Institute of Research Agriculture Toulouse	2003	France	H19	Gene (*cif*)
AY128540	*E. coli*	EPEC	pig		National Institute of Research Agriculture Toulouse	2003	France	1390	Gene (*cif*)
AY128541	*E. coli*	EPEC	rabbit		National Institute of Research Agriculture Toulouse	2003	France	C102	Gene (*cif*)
AY128542	*E. coli*	EPEC	cattle		National Institute of Research Agriculture Toulouse	2003	France	C/15333	Gene (*cif*)
AY128544	*E. coli*	EPEC	human		National Institute of Research Agriculture Toulouse	2003	France	EF-33	Gene (*cif*)
AY151282	*E. coli*	APEC	chicken	Septicemia	University of Guelph, Ontario Veterinary College	2002	Canada	Ec222	PAI
AY163491	*E. coli*	ETEC	human	Diarrhea	University of Tennessee, Dept. of Infectious Disease, Memphis	2003	USA (TN)	H10407	Gene (*eat* A)
AY170851	*E. coli*	STEC	cattle		University of Munich, Institute for Hygiene and Technology of Food	2002	Germany	MHI813	Gene (*stx* 1AB)
AY206437	*S. flexneri*		human		University of Sydney, School of Molecular and Microbial Biosciences	2002	Australia	623	Plasmid (set1)
AY206439	*S. flexneri*		human		University of Sydney, School of Molecular and Microbial Biosciences	2003	Australia	M1382	Gene (*ipg* D, *ipa* B)
AY206439	*S. flexneri*		human		University of Sydney, School of Molecular and Microbial Biosciences	2003	Australia	623	Gene (*ipa* B)
AY206441	*S. flexneri*		human		University of Sydney, School of Molecular and Microbial Biosciences	2003	Australia	M1382	Gene (*vir* A)
AY258503	*E. coli*	EHEC	human	sick	Monash University, Department of Microbiology	2001	Australia	EH41	Plasmid (pO113)
AY365042	*E. coli*	STEC	human	Diarrhea	University Muenster, Institute of Hygiene	2001	Germany	9282/01	Gene (*cdt Va-c*)
AY365043	*E. coli*	STEC	human	Diarrhea	University Muenster, Institute of Hygiene	2001	Germany	5249/01	Gene (*cdt Va-c*)
AY365044	*E. coli*	STEC	human	Diarrhea	University Muenster, Institute of Hygiene	2002	Germany	9063/02	Gene (*cdt Va-c*)
AY365045	*E. coli*	STEC	human	HUS	University Muenster, Institute of Hygiene	1996	Germany	2996/96	Gene (*cdt Va-c*)
AY545598	*E. coli*	APEC			Iowa State University, Veterinary Microbiology	2004	USA	A2363	Plasmid (pAPEC-O2-ColV)
AY576656	*E. coli*	STEC	human	HUS	University Hospital Frankfurt Main, Inst. of Medical Microbiology	1998	Germany	3232/96	Gene (*hly* E)
AY576657	*E. coli*	EIEC	human	Diarrhea	University Hospital Frankfurt Main, Inst. of Medical Microbiology	2004	Germany	4608-58	Gene (*hly* E)
AY576659	*E. coli*	EAEC	human	Diarrhea	University Hospital Frankfurt Main, Inst. of Medical Microbiology	1994	Germany	5477/94	Gene (*hly* E)
AY576660	*E. coli*	ETEC	human	Diarrhea	University Hospital Frankfurt Main, Inst. of Medical Microbiology	1999	Germany	G1253	Gene (*hly* E)
AY604009	*E. coli*	ETEC	pig		University of Munich, Department of Anaesthesiology	2004	Canada (ON)		Gene (*sep* A)
AY627215	*E. coli*	EIEC	human		University of Sao Paulo, Institute of Biomedical Sciences	1977	Brazil	FBC12430	Gene (*ipg* D)
AY627216	*E. coli*	EIEC	human		Biomedical Reference Laboratory, Melbourne		Australia	R411/55	Gene (*ipg* D)
AY627217	*E. coli*	EIEC	human		Biomedical Reference Laboratory, Melbourne		Australia	111/55	Gene (*ipg* D)
AY627218	*E. coli*	EIEC	human		Biomedical Reference Laboratory, Melbourne		Australia	223	Gene (*ipg* D)
AY627222	*E. coli*	EIEC			University of New South Wales, Biotechnology and Biomolecular Sc.	2004	Australia	M2332	Gene (*ipg* D)
AY627226	*E. coli*	EIEC	human		University of Sao Paulo, Institute of Biomedical Sciences		Brazil	FBC12406	Gene (*ipg* D)
AY627234	*E. coli*	EIEC	human		University of Sao Paulo, Institute of Biomedical Sciences	1985	Brazil	FBC16406	Gene (*ipg* D)
AY696755	*E. albertii*				Michigan State University, Microbial Evolution Laboratory	2004	USA	19982	Gene (*cdt Va-c*)
AY840983	*E. coli*		rabbit	Diarrhea	University of Maryland, Center for Vaccine Development, Baltimore	2004		RDEC-1	Gene (efa1/lifA)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
AY879342	*S. flexneri*		human		State Key Laboratory for Molecular Virology & Genetics, Beijing	2005	China	pSF5	Plasmid (pSF5)
AY944737	*E. coli*	Environ	water		Environmental Microbial Safety Laboratory, Baltimore, Beltsville	2005	USA (MD)	Baisman's Run stream	Gene (*tir*)
CP000035	*S. dysenteriae*		human	Dysentery	State Key Laboratory for Molecular Virology & Genetics, Beijing	2004	China	Sd197	Plasmid (pSD1_197)
CP000036	*S. boydii*		human		State Key Laboratory for Molecular Virology & Genetics, Beijing	2004	China	Sb227	Genome
CP000037	*S. boydii*		human		State Key Laboratory for Molecular Virology & Genetics, Beijing	2004	China	Sb227	Plasmid (pSB4_227)
CP000038	*S. sonnei*		human		State Key Laboratory for Molecular Virology & Genetics, Beijing	2004	China	Ss046	Genome
CP000039	*S. sonnei*		human		State Key Laboratory for Molecular Virology & Genetics, Beijing	2004	China	Ss046	Plasmid (pSS_046)
CP000243	*E. coli*	UPEC		UTI	Washington University School of Medicine, Molecular Microbiology	2006	USA(MA)	UTI89	Genome
CP000247	*E. coli*	UPEC	human	UTI	Georg August University Goettingen, Inst. of Microbiology and Genetic	2006	Germany	536	Genome
CP000468	*E. coli*	APEC	poultry	sick	Iowa State University, Veterinary Microbiology	2006	USA	APEC O1	Genome
CP000795	*E. coli*	ETEC	human	Enteritis	J. Craig Venter Institute, Medical Center, Rockville	< 1972	Bangladesh	E24377A	Plasmid (pETEC_80)
CP000799	*E. coli*	ETEC	human	Enteritis	J. Craig Venter Institute, Medical Center, Rockville	< 1972	Bangladesh	E24377A	Plasmid (pETEC_74)
CP000836	*E. coli*	APEC			Arizona State University, Center for Infectious Diseases	2007	USA	chi7122	Plasmid (pAPEC-1)
CP000913	*E. coli*	ETEC	pig		University Goettingen, Institute of Microbiology	2007	Germany	EC2173	Plasmid (pTC1)
CP000970	*E. coli*	Environ			J. Craig Venter Institute, Medical Center, Rockville	2008	USA (MD)	SMS-3-5	Genome
CP001062	*S. boydii*				University of Maryland, Dept. Microbiology & Immunology, Baltimore	2008	USA	3083-94	Plasmid (pBS512_211)
CP001063	*S. boydii*		human	Shigellosis	J. Craig Venter Institute, Medical Center, Rockville	2008	USA	CDC 3083-94	Genome
CP001064	*E. coli*	EIEC	human		J. Craig Venter Institute, Medical Center, Rockville	2008	USA	53638	Plasmid (p53638_226)
CP001162	*E. coli*				University of Minnesota, Veterinary and Biomedical Sciences	2008	USA		Plasmid (Vir68)
CP001163	*E. coli*	EHEC	human	Enteritis	J. Craig Venter Institute, Medical Center, Rockville	2006	USA	EC4115	Plasmid (p0157)
CP001164	*E. coli*	EHEC	human	sick	J. Craig Venter Institute, Medical Center, Rockville	2006	USA (MD)	EC4115	Genome
CP001368	*E. coli*	EHEC	human	HUS	University of Washington, Samuel Miller Laboratory, Seattle	2006	USA	TW14359	Genome
CP001369	*E. coli*	EHEC	human	HUS	University of Washington, Samuel Miller Laboratory, Seattle	2006	USA	TW14359	Plasmid (pO157)
CP001383	*S. flexneri*		human	Shigellosis	Chinese Center for Disease Control and Prevention	2002	China	2002017	Genome
CP001384	*S. flexneri*		human	Shigellosis	Chinese Center for Disease Control and Prevention	2002	China	2002017	Plasmid (pSFxv_1)
CP001509	*E. coli*				Industrial Biotechnology and Bioenergy Research Center, Daejeon	2009	Korea	BL21(DE3)	Genome
CP001671	*E. coli*		human	ABU	Lund University, Institute for Molecular Infection biology	1974	Sweden	ABU 83972	Genome
CP001846	*E. coli*	EPEC	human	Diarrhea	Nankai University, TEDA School of Biological and Biotech. Sciences	2003	Germany	CB9615	Genome
CP001855	*E. coli*	AIEC	human	IBD-CD	Public Health Agency of Canada, Laboratory for Foodborne Zoonoses	2010	Canada	NRG 857C	Genome
CP001926	*E. coli*	EHEC	human	HUS	Chinese Center for Disease Control and Prevention	2010	China	Xuzhou21	Plasmid (pO157)
CP002167	*E. coli*	AIEC	human	IBD-CD	Research and Testing Laboratory, Lubbock	2010	USA (TX)	UM146	Genome
CP002212	*E. coli*	UPEC	dog	UTI	University of Minnesota, Veterans Affairs Medical Center	2005/08	USA (MN)	D i14	Genome
CP002732	*E. coli*	ETEC	pig	Diarrhea	University of Minnesota, Veterinary and Biomedical Sciences	2007	USA (MN)	UMNK88	Plasmid (pUMNK88)
CP002733	*E. coli*	ETEC	pig	Diarrhea	University of Minnesota, Veterinary and Biomedical Sciences	2007	USA (MN)	UMNK88	Plasmid (pUMNK88_Hly)
CP002967	*E. coli*				University of Florida, Microbiology and Cell Biology, Gainesville	2011	USA (FL)	W (ATCC 9637)	Genome
CP002970	*E. coli*				University of Florida, Microbiology and Cell Biology, Gainesville	2011	USA (FL)	KO11FL	Genome
CP003034	*E. coli*	NMEC	human	NM	Microbial Genome Center of Chinese Ministry of Public Health, Beijing	1970	China	CE10	Genome
CP003109	*E. coli*	EPEC	human	UTI	Western Regional Research Center/Produce Safety and Microbiology	1974	USA (CA)	RM12579	Genome
CP003289	*E. coli*	EHEC	human	HUS	CDC, Los Alamos National Laboratory, New Mexico	2011	Germany	2011C-3493	Genome
CP003291	*E. coli*	EHEC	human	HUS	CDC, Los Alamos National Laboratory, New Mexico	2011	Germany	2011C-3493	Plasmid (pESBL-EA11)
CP003297	*E. coli*	EHEC	human	HUS	National Center for Disease Control and Public Health, Tbilisi	2009	Georgia	2009EL-2050	Genome
CP003299	*E. coli*	EHEC	human	HUS	National Center for Disease Control and Public Health, Tbilisi	2009	Georgia	2009EL-2050	Plasmid (pAA-09EL50)
CP003301	*E. coli*	EHEC	human	HUS	National Center for Disease Control and Public Health, Tbilisi	2009	Georgia	2009EL-2071	Genome

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
CP003302	*E. coli*	EHEC	human	HUS	National Center for Disease Control and Public Health, Tbilisi	2009	Georgia	2009EL-2071	Plasmid (pAA-09EL71)
CP004056	*S. flexneri*		human		Chinese Center for Disease Control and Prevention	2013	China	2003036	Genome
CP004057	*S. flexneri*		human		Chinese Center for Disease Control and Prevention	2013	China	Shi06HN006	Genome
CP005930	*E. coli*	APEC	chicken		Free University Berlin, Institute of Microbiology and Epizootics	2001	Germany	IMT5155	Genome
CP005931	*E. coli*	APEC	chicken		Free University Berlin, Institute of Microbiology and Epizootics	2001	Germany	IMT5155	Plasmid (p1ColV5155)
CP006001	*E. coli*	ETEC	human	Diarrhea	J. Craig Venter Institute, Medical Center, Rockville	2005	Viet Nam	B7A	Plasmid (pEB3)
CP006002	*E. coli*	ETEC	human	Diarrhea	J. Craig Venter Institute, Medical Center, Rockville	2005	Viet Nam	B7A	Plasmid (pEB4)
CP006027	*E. coli*	STEC	Food		United States Department of Agriculture, Albany	2013	USA	RM13514	Genome
CP006028	*E. coli*	STEC	Food		United States Department of Agriculture, Albany	2013	USA	RM13515	Plasmid (pO145-13514)
CP006262	*E. coli*	EHEC	Food		United States Department of Agriculture, Albany	2013	USA	RM13516	Genome
CP006263	*E. coli*	EHEC	Food		United States Department of Agriculture, Albany	2013	USA	RM13517	Plasmid (pO145-13516)
CP006737	*S. dysenteriae*		human	Shigellosis	Armed Forces Research Institute of Medical Sciences	1968	Thailand	1617	Plasmid (pSLG231)
CP006784	*E. coli*				State Serum Institute, Microbiology & InfectControl, Copenhagen	2013	USA	JJ1886	Genome
CP006830	*E. coli*	APEC	chicken		Iowa State University, Veterinary Microbiology and Preventive Medicin	2013	USA (IA)	APEC O18	Genome
CP007038	*S. flexneri*		human		TEDA Institute of Biological and Biotechnology, Tianjin	2013	China	G1663	Plasmid (pG1663)
CP007133	*E. coli*	STEC	Food		United States Department of Agriculture, Albany	2007	Belgium	RM12761	Genome
CP007135	*E. coli*	STEC	Food		United States Department of Agriculture, Albany	2007	Belgium	RM12762	Plasmid (pO145-12761)
CP007136	*E. coli*	STEC	Food		United States Department of Agriculture, Albany	2010	USA	RM12581	Genome
CP007394	*E. coli*		water		University of Antwerp, Medical Microbiology	2012	Belgium	ST2747	Genome
CP007592	*E. coli*	EHEC	human	healthy	Bangladesh Agricultural University, Veterinary Science	2012	Bangladesh	Santai	Genome
CP007594	*E. coli*		pig	Diarrhea	Animal Nutrition Center, Inst. of Subtropical Agriculture, Hunan	2010	China	SEC470	Genome
CP007799	*E. coli*		human	healthy	University Wuerzburg, Bioinformatics Center, Germany	1917	Germany	Nissle 1917	Genome
CP008805	*E. coli*	STEC	cattle	healthy	University Park, 205 Wartik Laboratory	2014	USA	SS17	Genome
CP008806	*E. coli*	STEC	cattle	healthy	University Park, 205 Wartik Laboratory	2014	USA	SS18	Plasmid (pO157)
CP008958	*E. coli*	EHEC	human	HC	University of Wisconsin, Genome Center of Wisconsin, Madison	1982	USA (MI)	EDL933	Plasmid (p0157)
CP009072	*E. coli*		human	sick	CDC, Los Alamos National Laboratory, New Mexico	1946	USA (WA)	ATCC 25922	Genome
CP009105	*E. coli*	STEC	cattle		United States Department of Agriculture, Wyndmoor	2009	USA (CA)	RM9387	Plasmid (pO104_H7)
CP009107	*E. coli*	STEC	cattle	HC	United States Department of Agriculture, Wyndmoor	1994	USA (MT)	94-3024	Plasmid (pO104_H21)
CP009166	*E. coli*		cattle	Mastitis	University Muenster, Institute of Hygiene	2014	Germany	1303	Genome
CP009273	*E. coli*				University of Sherbrooke, Dept. of Microbiology	2014	Canada	K-12 BW25113	Genome
CP009644	*E. coli*				New England Biolabs, Ipswich	2014	USA	K-12 8ER2796	Genome
CP009859	*E. coli*		human	sick	National Human Genome Research Institute, Bethesda	2014	USA	ECONIH1	Genome
CP010151	*E. coli*		dog		Nanjing Agricultural University, Bioinformatics Center, Jiangsu	2014	China	D8	Genome
CP010304	*E. coli*	STEC	cattle	Diarrhea	United States Department of Agriculture, Nebraska	2013	USA (NE)	SS52	Genome
CP010305	*E. coli*	STEC	cattle	Diarrhea	United States Department of Agriculture, Nebraska	2013	USA (NE)	SS52	Plasmid (p0157)
CP010315	*E. coli*	APEC	poultry	Septicemia	University Goettingen, Institute of Microbiology	2014	Germany	789	Genome
CP010344	*E. coli*		cattle	Mastitis	University Muenster, Institute of Hygiene	2014	Germany	ECC-1470	Genome
CP010371	*E. coli*		human		Medical Center Drive, Rockville	2014	USA	6409	Genome
CP010830	*S. sonnei*		human		Kyung Hee University, Food Science and Biotechnology, Gyeonggi	2015	South Korea	FORC_011	Plasmid (pFORC11.1)
CP010876	*E. coli*	ExPEC	human	ABU	University of Minnesota, Veterinary and Biomedical Sciences	2012	USA (MN)	MNCRE44	Genome
CP011018	*E. coli*	UPEC	human	UTI	National University of Singapore, Medicine/Infectious Diseases	2013	Singapore	CI5	Genome
CP011134	*E. coli*	UPEC	human	ABU	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	VR50	Genome
CP011417	*E. coli*	EIEC	human	Diarrhea	Center for Food Safety and Applied Nutrition, United States FDA	2015	USA (MD)	CFSAN029787	Plasmid (pCFSAN029787_01)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
CP011511	*S. boydii*		human	Shigellosis	Korea Research Institute Of Bioscience and Biotechnology, Daejeon	2016	Korea	ATTC 9210	Genome
CP012142	*S. flexneri*		human		Hangzhou Center for Desease Control and Prevention, Zhejiang	2015	China	4c srt. 1205	Plasmid (1205p2)
CP012633	*E. coli*		human	Septicemia	Santa Clara University, Biology	2008	USA (CA)	SF-166	Genome
CP013025	*E. coli*	STEC	human	HUS	CDC, Atlanta, Georgia	2009	USA (GA)	2009C-3133	Genome
CP013029	*E. coli*	STEC	human	HUS	CDC, Atlanta, Georgia	2012	USA (GA)	2012C-4227	Genome
CP013658	*E. coli*		human	sick	University of Oxford	2005	UK	uk_P46212	Genome
CP013962	*E. coli*		pig	Diarrhea	Animal Nutrition Center, Inst. of Subtropical Agriculture, Hunan	2010	China	SEC470	Genome
CP014092	*E. coli*	Environ	well		University of Maryland, Institute for Genome Science	2015	USA (MD)	268-78-1	Genome
CP014197	*E. coli*	Environ	water		Culture Collection of Public Health England, London	1950	UK	MRE600	Genome
CP014488	*E. coli*				University of Minnesota, Veterinary and Biomedical Sciences	2016	USA (MI)	G749	Genome
CP016497	*E. coli*	UPEC	human	UTI	Chungnam National University, Dept. Of Microbiology and Med. Sc.	2013	Korea	UPEC 26-1	Genome
CR942285	*E. coli*	ETEC			Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2005	UK		Plasmid (pCoo)
CU651637	*E. coli*	AIEC	human	IBD-CD	Genoscope, Centre National de Sequence	2010	France	LF82	Genome
CU928145	*E. coli*	EAEC	human	Diarrhea	Genoscope, Centre National de Sequence	2009	Central Africa	55989	Genome
CU928159	*E. coli*	EAEC	human	Diarrhea	University Pierre et Marie Curie, Atelier de BioInformatique, Paris	2009	Central Africa	55989p	Plasmid (55989p)
CU928160	*E. coli*	COM	human	healthy	Genoscope, Centre National de Sequence	1980	France	IAI1	Genome
CU928161	*E. coli*	ExPEC	human	NM	Genoscope, Centre National de Sequence	1999	France	S88	Genome
CU928162	*E. coli*	COM	human	healthy	Genoscope, Centre National de Sequence	2000	France	ED1a	Genome
CU928163	*E. coli*	UPEC	human	UTI	Genoscope, Centre National de Sequence	1999	USA	UMN026	Genome
CU928164	*E. coli*	UPEC	human	UTI	Genoscope, Centre National de Sequence	1980	France	IAI39	Genome
D26468	*S. flexneri*		human	Shigellosis	University of Tokyo, Institute of Medical Science	1994	Japan	2a pMYSH6000	Gene (*vir* A)
DQ007019	*E. coli*		human		Royal Free Hospita, Paediatric Gastroenterology, London	2005	UK	GPG122-G57	Gene (*tir*)
DQ007021	*E. coli*		human		Royal Free Hospita, Paediatric Gastroenterology, London	2005	UK	CPG7	Gene (*tir*)
DQ381420	*E. coli*	APEC	poultry	sick	Iowa State University, Veterinary Microbiology	2006	USA	APEC O1	Plasmid (pAPEC-O1-ColBM)
DQ778054	*E. coli*		pig	Diarrhea	Hunan Agriculture University, College of Veterinary Medicine	2004	China	Eco412	Gene (LT)
ECOASTAA	*E. coli*	EAEC			Naval Medical Research Institut, Bethesda	1993	USA (MD)	L11241	Gene (EAST1)
ECOCNF2	*E. coli*				Uniformd Services University of Health Sciences, Dept. of Microbio.	1993	USA (MD)	711(pVir)	Gene (*cnf* 2)
ECOSTI	*E. coli*		cattle		Naval Medical Research Institut, Bethesda	1985	USA (MD)	M25607	Gene (STa)
ECU42629	*E. coli*	Environ	water		J. Craig Venter Institute, Medical Center, Rockville	2007	USA (MD)	109aa	Gene (*cnf* 1)
EF057802	*E. coli*	ETEC			Xiamen University, The Key Laboratory for Cell Biology & Engineering	2006	China	195	Gene (LT)
EF158843	*E. coli*		human	Diarrhea	Pasteur Institute of Iran, Molecular Biology Unit, Tehran	2006	Iran	163-3	Gene (*cdt Va-c*)
EF204919	*E. coli*	STEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR047	Gene (*ehx* A)
EF204920	*E. coli*	STEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR053	Gene (*ehx* A)
EF204921	*E. coli*	STEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR119	Gene (*ehx* A)
EF204922	*E. coli*	STEC	sheep		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR151	Gene (*ehx* A)
EF204923	*E. coli*	EPEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR158	Gene (*ehx* A)
EF204924	*E. coli*	STEC	human		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR270	Gene (*ehx* A)
EF204925	*E. coli*	STEC	sheep		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR340	Gene (*ehx* A)
EF204926	*E. coli*	STEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR374	Gene (*ehx* A)
EF204927	*E. coli*	STEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR670	Gene (*ehx* A)
EF204928	*E. coli*	STEC	cattle		Enteric Reference Laboratory, Porirua	2007	New Zealand	AGR674	Gene (*ehx* A)
EF204929	*E. coli*	STEC	human	Diarrhea	Food and Health, AgResearch Limited, Manawatu	2007	New Zealand	ER03/4238	Gene (*ehx* A)
EF441594	*E. coli*	STEC			Idaho State University, Biological Sciences	2004	USA (ID)	I8419	Gene (*stx* 1AB)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
EF441599	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	93-111	Gene (*stx* 2AB)
EF441602	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	493/89	Gene (*stx* 2AB)
EF441605	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	5905	Gene (*stx* 2AB)
EF441613	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	I6581	Gene (*stx* 2AB)
EF441616	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	I7606	Gene (*stx* 2AB)
EF441618	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	CL-3	Gene (*stx* 2AB)
EF441619	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	G5506	Gene (*stx* 2AB)
EF441621	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	88-1509	Gene (*stx* 2AB)
EF441622	*E. coli*	STEC			Idaho State University, Biological Sciences	2007	USA (ID)	90-1787	Gene (*stx* 2AB)
EM164354	*E. coli*	Environ	water		J. Craig Venter Institute, Medical Center, Rockville	2007	USA (MD)	105aa	Gene (*esp* C)
EU113242	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	136-I	Gene (LT)
EU113243	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	1372-1	Gene (LT)
EU113244	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	121-I	Gene (LT)
EU113245	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	225-IV	Gene (LT)
EU113246	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	PE0534	Gene (LT)
EU113247	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	PE0415	Gene (LT)
EU113248	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	4702-1	Gene (LT)
EU113249	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	214-III	Gene (LT)
EU113250	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	4692-1	Gene (LT)
EU113251	*E. coli*	ETEC	human	Diarrhea	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	4321-1	Gene (LT)
EU113252	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	PE0615	Gene (LT)
EU113253	*E. coli*	ETEC	human	healthy	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	63-V	Gene (LT)
EU113254	*E. coli*	ETEC	human	Diarrhea	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	61A-4	Gene (LT)
EU113255	*E. coli*	ETEC	human	Diarrhea	University of Sao Paulo, Institute of Biomedical Sciences	2008	Brazil	2781-5	Gene (LT)
EU273279	*E. coli*	EHEC	spinach		Pennsylvania State University, Dept. of Vet. and Biomed. Sciences	2006	USA(PA)	06E01767	Gene (*stx* 1AB)
EU330199	*E. coli*	APEC	chicken		University of Melbourne, Veterinary Science	2007	Australia	E3	Plasmid (pVM01)
EU700490	*E. coli*	EPEC	cattle		Friedrich-Loeffler-Institut, Institute for Animal Health, Wusterhausen	2008	Germany	WUS-01/27/017-1	Gene (*stx* 1AB)
EU871626	*E. coli*	STEC	human		Laboratory for Foodborne Zoonoses, Lethbridge	2008	Canada	ECI-1717	Gene (*tir*)
EU871627	*E. coli*				Public Health Agency of Canada, Laboratory for Foodborne Zoonoses	2008	Canada	EDS-58	Gene (*esp* F)
EU902125	*E. coli*	STEC	cattle		University Muenster, Institute of Hygiene	1989	Germany	493/89	Gene (*hly* E)
EU902128	*E. coli*	EPEC			University of Washington, School of Medicine, Dept. of Pediatrics	1997	USA	TB182A	Gene (*hly* E)
FJ386569	*E. coli*	STEC	human	HUS	United States Department of Agriculture, Wyndmoor	2008	USA	H30	Plasmid (pO26-Vir)
FJ664545	*E. coli*				Inst. Superiore di Sanita, Veterinary Public Health and Food Safety	2010	Italy	ED 591	Gene (*sub* AB)
FJ875095	*E. coli*	STEC	human	HUS	United States Department of Agriculture, Wyndmoor	2009	USA	83-75	Gene (*esp* P)
FM180012	*E. coli*	EPEC	human	sick	Federal Institute for Risk Assesment, Ref. Lab for *E. coli*	1983	Germany	C4115	Gene (*hly* A)
FM180568	*E. coli*	EPEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	1969	UK	E2348/69	Genome
FM201463	*E. coli*	aEPEC			University Muenster, Institute of Infectiology	2008	Germany	B6	PAI (LEE)
FM201464	*E. coli*				University Muenster, Institute of Infectiology	2006	Germany	9812	PAI (LEE)
FM986650	*E. coli*	EHEC	human	Diarrhea	Centers for Disease Control and Prevention, Maryland	1977	USA (MD)	DEC8A	PAI (LEE)
FM986651	*E. coli*	EHEC			University of Maryland, Center for Vaccine Development, Baltimore	2009	USA (MD)	IHIT 1190	PAI (LEE)
FM986652	*E. coli*	STEC	cattle	Diarrhea	University of Maryland, Center for Vaccine Development, Baltimore	2009	USA (MD)	537/89	PAI (LEE)
FM998838	*E. coli*	STEC	meat		University of Hohenheim, Food Microbiology	2008	Germany	TS01/08	Gene (*stx* 2AB)
FM998844	*E. coli*	STEC	sausage		University of Hohenheim, Food Microbiology	2007	Germany	TS09/07	Gene (*stx* 2AB)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
FM998851	*E. coli*	STEC	pork		University of Hohenheim, Food Microbiology	2008	Germany	TS17/08	Gene (*stx* 2AB)
FM998855	*E. coli*	STEC	meat		University of Hohenheim, Food Microbiology	2008	Germany	TS21/08	Gene (*stx* 2AB)
FM998860	*E. coli*	STEC	beef		University of Hohenheim, Food Microbiology	2008	Germany	TS27/08	Gene (*stx* 2AB)
FN554766	*E. coli*	EAEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	E. coli 042	Genome
FN554767	*E. coli*	EAEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	E. coli 042	Plasmid (pAA)
FN649414	*E. coli*	ETEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	H10407	Genome
FN649417	*E. coli*	ETEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	H10407	Plasmid (p666)
FN649418	*E. coli*	ETEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2009	UK	H10407	Plasmid (p948)
FN822745	*E. coli*	ETEC	human	Diarrhea	Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2010	China	1392/75	Plasmid (p1018)
GQ259888	*E. coli*	EHEC	human	HUS	University of Melbourne, National E. coli Reference Laboratory	2001	Australia	O6877	Plasmid (pO26-CRL)
GQ338312	*E. coli*	STEC	human		Laboratory for Foodborne Zoonoses, Lethbridge	2009	Canada	71074	PAI (LEE)
GQ429155	*E. coli*	STEC	beef		University of Maryland, Nutrition and Food Science	2009	USA (MD)	22813	Gene (*stx* 1AB)
GQ429160	*E. coli*	STEC	beef		University of Maryland, Nutrition and Food Science	2009	USA (MD)	N15018	Gene (*stx* 2AB)
GQ497943	*E. coli*	ExPEC	human		Institut Pasteur, Genotyping of Pathogens, Paris	2009	France	AL862	PAI (I)
GU363949	*E. coli*	EHEC	human	HUS	University of Idaho, Dept. of Microbiology, Moscow	2009	USA	ATCC 43894 (EDL 932)	Plasmid (pO157)
HE603111	*E. coli*	EHEC	human	HUS	University Giessen, Institut of Medical Microbiology	2001	Germany	HUSEC41	Plasmid (HUSEC41-2)
HE610901	*E. coli*	EHEC	human	HUS	University Giessen, Institut of Medical Microbiology	2011	Germany	HUSEC2011	Plasmid (pHUSEC2011-2)
HE616528	*S. sonnei*		human		Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2011	UK	53G	Genome
HE616529	*S. sonnei*				Wellcome Trust Sanger Institute, Pathogen Sequencing, Cambridge	2011	UK	53G	Plasmid (pA)
HF922624	*E. coli*	NMEC	human	NM	Institut Pasteur, Genotyping of Pathogens, Paris	2013	France	S286	Plasmid (pS286colV)
HG428755	*E. coli*	ExPEC			Free University Brussel, Bacterial Genetics and Physiology	2013	Belgium	PMV-1	Genome
HG941718	*E. coli*		human	UTI	University of Queensland, Centre for Infectious Disease Research	2011	UK	EC958	Genome
HM099896	*E. coli*		human		Niigata University, Department of Public Health, Japan	2010	Japan	7	Gene (EAST1)
HM099897	*E. coli*		human		Niigata University, Department of Public Health, Japan	2010	Japan	17	Gene (EAST1)
HM138194	*E. coli*	STEC	human	HUS	United States Department of Agriculture, Wyndmoor	2009	USA	83-75	Plasmid (pO145-NM)
HM581522	*E. coli*				Shanghai JiaoTong University, China	2010	China	Min27	Gene (*tir*)
HQ591459	*S. sonnei*				National Inst. of Genetic Engineering & Biotechnology, Tehran	2010	Iran	DY89	Gene (*ipa* B)
J04117	*S. flexneri*		human	Shigellosis	University of Tokyo, Institute of Medical Science	1989	Japan	M90T-W	Gene (*ipa* B)
JN130365	*E. coli*				University of Barcelona, Department of Microbiology	2011	Spain	374	Gene (*hly* A)
JPQG000000(	*E. coli*		deer		Vetsuisse Faculty of Zürich, Inst. For Food Safety and Hygiene	2013	Switzerland	str 48	Genome
JQ327853	*E. coli*	STEC			Wuhan Polytechnic University, Biological and Pharmaceutical Engin.	2011	China	EC150	Gene (*stx* 1AB)
JQ327854	*E. coli*	STEC			Wuhan Polytechnic University, Biological and Pharmaceutical Engin.	2011	China	EC169	Gene (*stx* 1AB)
JQ411011	*E. coli*	STEC	cattle		Faculty of Veterinary Medicine Udayana University, Bali	2012	Indonesia	SM-25(1)	Gene (*stx* 2AB)
JQ994271	*E. coli*		human		Inst. Superiore di Sanita, Veterinary Public Health and Food Safety	2012	Italy	ED 23	PAI (I)
JX050263	*E. coli*	EAEC			University Xochimilco, Dept. of human medicine, Mexico	2012	Mexico	49766	Gene (*sat*)
JX050263	*E. coli*	EAEC			University Autonoma Metropolitan Xochimilco, Sistemas Biologicos	2010	Mexico	49766	Gene (*sat*)
JX206444	*E. coli*	STEC	fish		Central Institute of Fisheries Education, Mumbai	2010	India	A7	Gene (*stx* 1AB)
JX504011	*E. coli*	ETEC	human	Diarrhea	State Research Center, Microbiology & Biotechnology, Obolensc	2011	Russia	118-5	PAI
KC853435	*E. coli*	APEC	chicken	Septicemia	State Key Laboratory of Agricultural Microbiology, Hebei	2013	China	ACN001	Plasmid (pACN001-B)
KF322032	*E. coli*	Environ	water		University of Barcelona, Microbiology	2013	Spain	O157:H7	Phage (fiAA91-ss)
KJ149553	*E. coli*	ExPEC			University Estadual de Campinas, Evolution & Biogenetics, Brazil	2014	Brazil	EC113	Gene (EAST1)
KJ149554	*E. coli*	ExPEC			University Estadual de Campinas, Evolution & Biogenetics, Brazil	2014	Brazil	322/EC	Gene (EAST1)
KJ149555	*E. coli*	EAEC			University Estadual de Campinas, Evolution & Biogenetics, Brazil	2014	Brazil	UEL01	Gene (EAST1)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
KJ466025	*E. coli*	EHEC	human	HUS	University Hospital Muenster, Inst. for Hygiene and nat. Lab. on HUS	2014	Germany	02-05924	Gene (EAST1)
KJ466027	*E. coli*	EHEC	human	HUS	University Hospital Muenster, Inst. for Hygiene and nat. Lab. on HUS	2014	Germany	99-04062	Gene (EAST1)
KJ466030	*E. coli*	EHEC	human	HUS	University Hospital Muenster, Inst. for Hygiene and nat. Lab. on HUS	2014	Germany	4728/00	Gene (EAST1)
KJ466034	*E. coli*	EHEC	human	HUS	University Hospital Muenster, Inst. for Hygiene and nat. Lab. on HUS	2001	Germany	HUSEC41	Gene (EAST1)
KJ747188	*E. coli*	EPEC	human	Diarrhea	University of Federal de São Paulo, Departament of Microbiology	2014	Brazil	A16	Gene (EAST1)
KP120709	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4023	Gene (*stx* 1AB)
KP120710	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4025	Gene (*stx* 1AB)
KP120714	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4077	Gene (*stx* 1AB)
KP120716	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4124	Gene (*stx* 1AB)
KP120717	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4001	Gene (*stx* 2AB)
KP120718	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4015	Gene (*stx* 2AB)
KP120719	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4035	Gene (*stx* 2AB)
KP120722	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4082	Gene (*stx* 2AB)
KP120723	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4086	Gene (*stx* 2AB)
KP120724	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4105	Gene (*stx* 2AB)
KP120725	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4110	Gene (*stx* 2AB)
KP120726	*E. coli*	STEC	yak		College of Life Science and Technology, Sichuan	2014	China	SWUN4160	Gene (*stx* 2AB)
KR094926	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB05	Gene (*vat*)
KR094927	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB07	Gene (*vat*)
KR094928	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB08	Gene (*vat*)
KR094929	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB10	Gene (*vat*)
KR094930	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB11	Gene (*vat*)
KR094931	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB15	Gene (*vat*)
KR094932	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB19	Gene (*vat*)
KR094933	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB22	Gene (*vat*)
KR094934	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB25	Gene (*vat*)
KR094935	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB26	Gene (*vat*)
KR094936	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB28	Gene (*vat*)
KR094937	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB29	Gene (*vat*)
KR094938	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB32	Gene (*vat*)
KR094939	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB37	Gene (*vat*)
KR094940	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB38	Gene (*vat*)
KR094941	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB41	Gene (*vat*)
KR094942	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB42	Gene (*vat*)
KR094943	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB44	Gene (*vat*)
KR094944	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB45	Gene (*vat*)
KR094945	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB47	Gene (*vat*)
KR094946	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB48	Gene (*vat*)
KR094947	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB50	Gene (*vat*)
KR094948	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB52	Gene (*vat*)
KR094949	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB55	Gene (*vat*)
KR094950	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB56	Gene (*vat*)
KR094951	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB57	Gene (*vat*)

NCBI No.	Species	Pathotype	Host	Disease	Source Laboratory	Year	Region	Common name / No.	Genome, Plasmid, PAI, Gene
KR094952	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB60	Gene (*vat*)
KR094953	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB63	Gene (*vat*)
KR094954	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB66	Gene (*vat*)
KR094955	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB70	Gene (*vat*)
KR094956	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB71	Gene (*vat*)
KR094957	*E. coli*	UPEC	human	UTI	University of Queensland, School of Chemestrie and Molecular Science	2015	Australia	PAB72	Gene (*vat*)
L11241	*E. coli*	EAEC	human	Diarrhea	Naval Medical Research Institute, Enteric Diseases, Bethesda	1993	USA	17-2	Gene (EAST1)
LC053401	*E. coli*		quail		Tetsuya Hayashi Kyushu University, Department of Bacteriology	2015	Japan	NIAH_Bird_32	PAI (Locus of *efa*)
M10133	*E. coli*	UPEC			University of Wisconsin, Dept. of Medical Microbiology, Madison	1985	USA	J96	Gene (*hly* A)
M18345	*E. coli*		human		University of Texas Medical Center, Dept. of Microbiology, Dallas	1988	Bangladesh	CRL 25090	Gene (STb)
M25607	*E. coli*					1985		clone pT5031	Gene (STa)
M29255	*E. coli*	ETEC			Astra Research Centre India, Bangalore	1989	India	clone pARC063	Gene (STb)
M34916	*E. coli*		human	Diarrhea	Stanford University, Dept. of Microbiology and Immunology	1983	Thailand	153837-2	Gene (STb)
M58746	*E. coli*	ETEC	human	Diarrhea	Wellcome Research Laboratories, Dept. Microbiology, NorthCarolina	1990	USA (KY)	18D	Gene (STa)
S81691	*E. coli*	ETEC	human	Diarrhea	National Center for Biotechnology Information, Bethesda	2014	Bangladesh	H10707-P	Plasmid (pCS1)
U01097	*E. coli*				University of Health Sciences, Dept. of Microbiology, Bethesda	1993	USA		Gene (*cnf* 2)
U35656	*S. flexneri*		human		University of Maryland, Center for Vaccine Development, Baltimore	1995	USA (MD)	2a	Toxin set
U42629	*E. coli*				University of Maryland, Center for Vaccine Development, Baltimore	1994	USA		Gene (*cnf* 1)
U97487	*S. flexneri*		human		Monash University, Department of Microbiology	1997	Australia	2a str. SBA1336	Gene (*sig* A)
V00275	*E. coli*	ETEC	human	Diarrhea	Juntendo University, Department of Bacteriology, Tokyo	1982	Japan	H10407	Gene (LT)
V00612	*E. coli*	ETEC	human			1980			Gene (STa)
X15319	*S. flexneri*		human	Shigellosis	University of Tokyo, Institute of Medical Science	1989	Japan	str 2a pMYSH6000	Gene (*ipa* B)
X70670	*E. coli*		human		University Nodo Nazionale Italiano, Bari	1993	Italy		Gene (*cnf* 1)
Y13614	*E. coli*	STEC	cattle		University Giessen, Institut of Medical Microbiology	1997	Germany	413/89-1	Gene (*pss* A)
Z36901	*E. coli*	STEC	human		Women's and Children's Hospital, Microbiology, South Australia	2005	Australia	DG131/3	Gene (*stx* 1AB)
Z48219	*S. flexneri*		human	Shigellosis	Institut Pasteur, Bacteriology and Mycology, Paris	1995	France	5a str. M90T Sm	Gene (*sep* A)

Pathotype: ExPEC: Extraintestinal pathogenic *E. coli*, APEC: Avian pathogenic *E. coli,* NMEC: Neonatal meningitis *E. coli*, UPEC: Uropathogenic *E. coli*; InPEC: Intestinal pathogenic *E. coli,* EAEC: Enteroaggregative *E. coli*, EHEC: Enterohaemorrhagic *E. coli*, EIEC: Enteroinvasive *E. coli*, EPEC: Enteropathogenic *E. coli*, aEPEC: atypical Enteropathogenic *E. coli,* ETEC: Enterotoxigenic *E. coli*, STEC: Shiga-Toxin-producing *E. coli*; COM: Commensal; Environ: Environment; Disease: ABU: Asymtomatic bacteriuria, HC: Hemorrhagic colitis, HUS: Hemolytic uremic syndrome, NM: Neonatal meningitis, UTI: Urinary tract infection, WI: Wound infection, IBD-CD: Inflammatory Bowel Disease – Chron's Disease

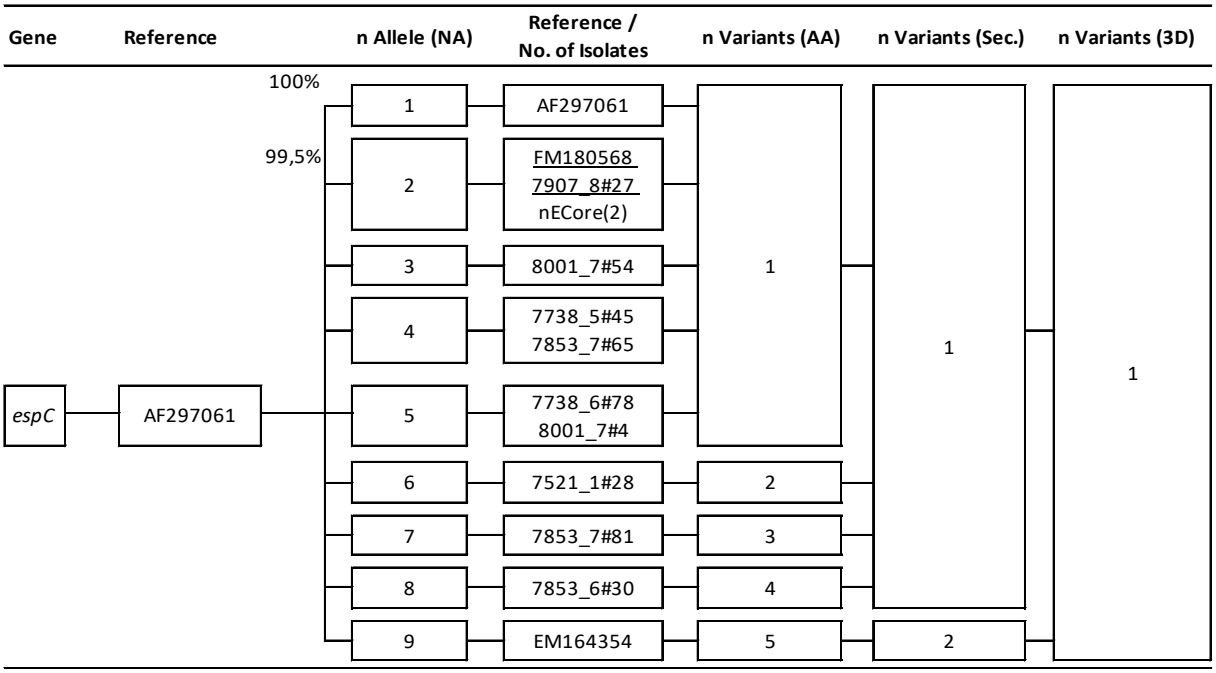

Figure A 1: *espC* (SPATE – EPEC secreted protein C) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF297061 (genetic identity in percent).

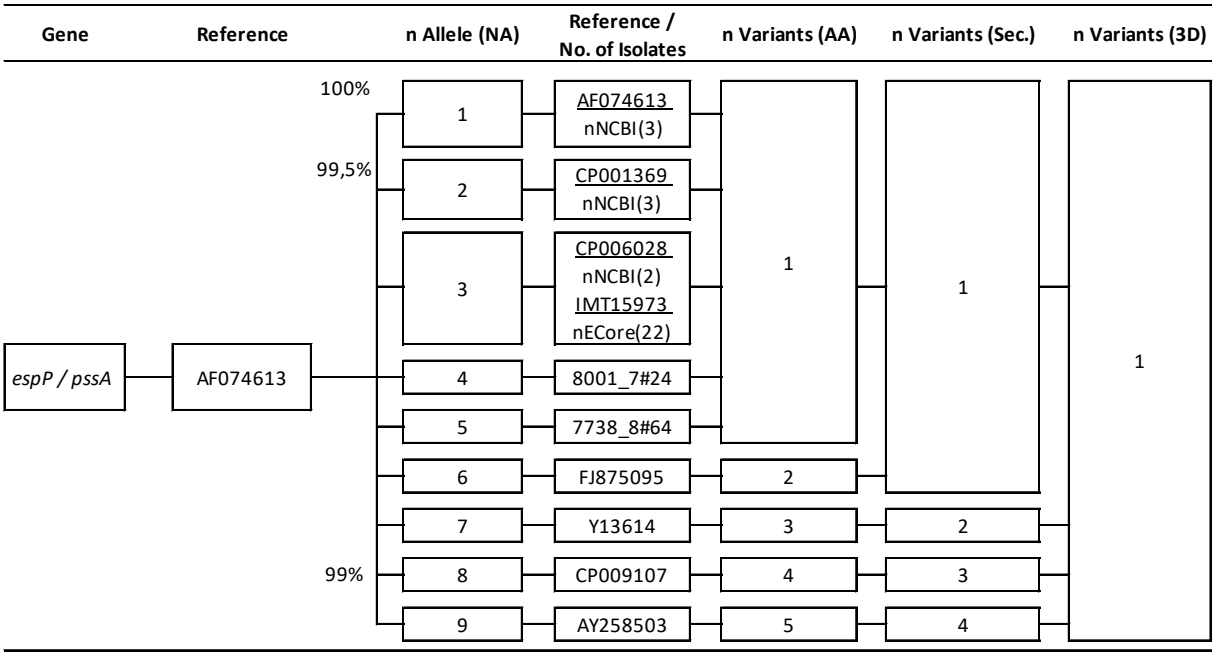

Figure A 2: *espP / pssA* (SPATE – Extracellular serine protease (EHEC) / protease secreted by STEC) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF074613 (genetic identity in percent).

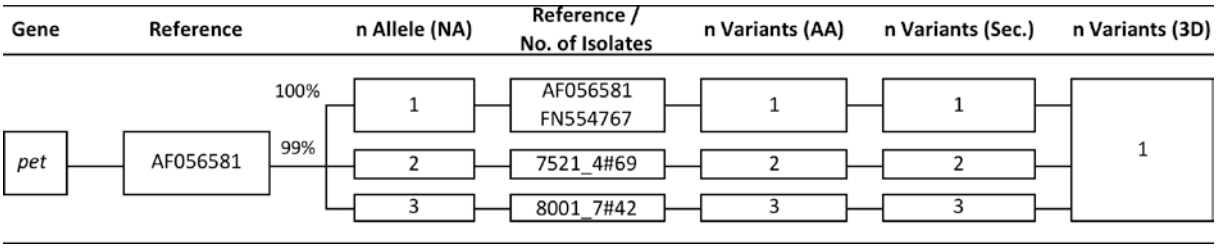

Figure A 3: *pet* (SPATE – Plasmid encoded toxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants with respect to the reference sequence AF056581 (genetic identity in percent).

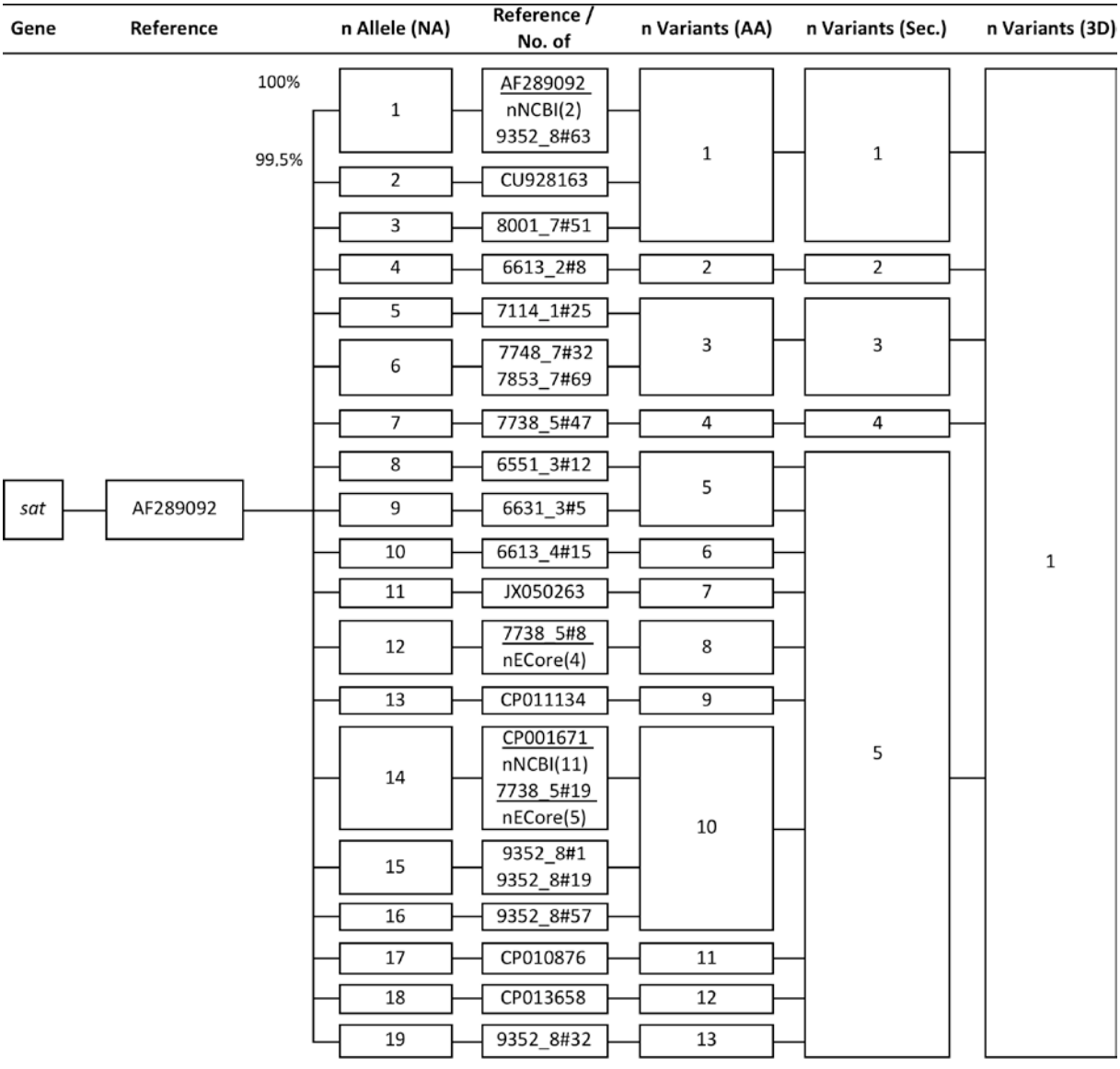

Figure A 4: *sat* (SPATE – Secreted autotransporter toxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF289092 (genetic identity in percent).

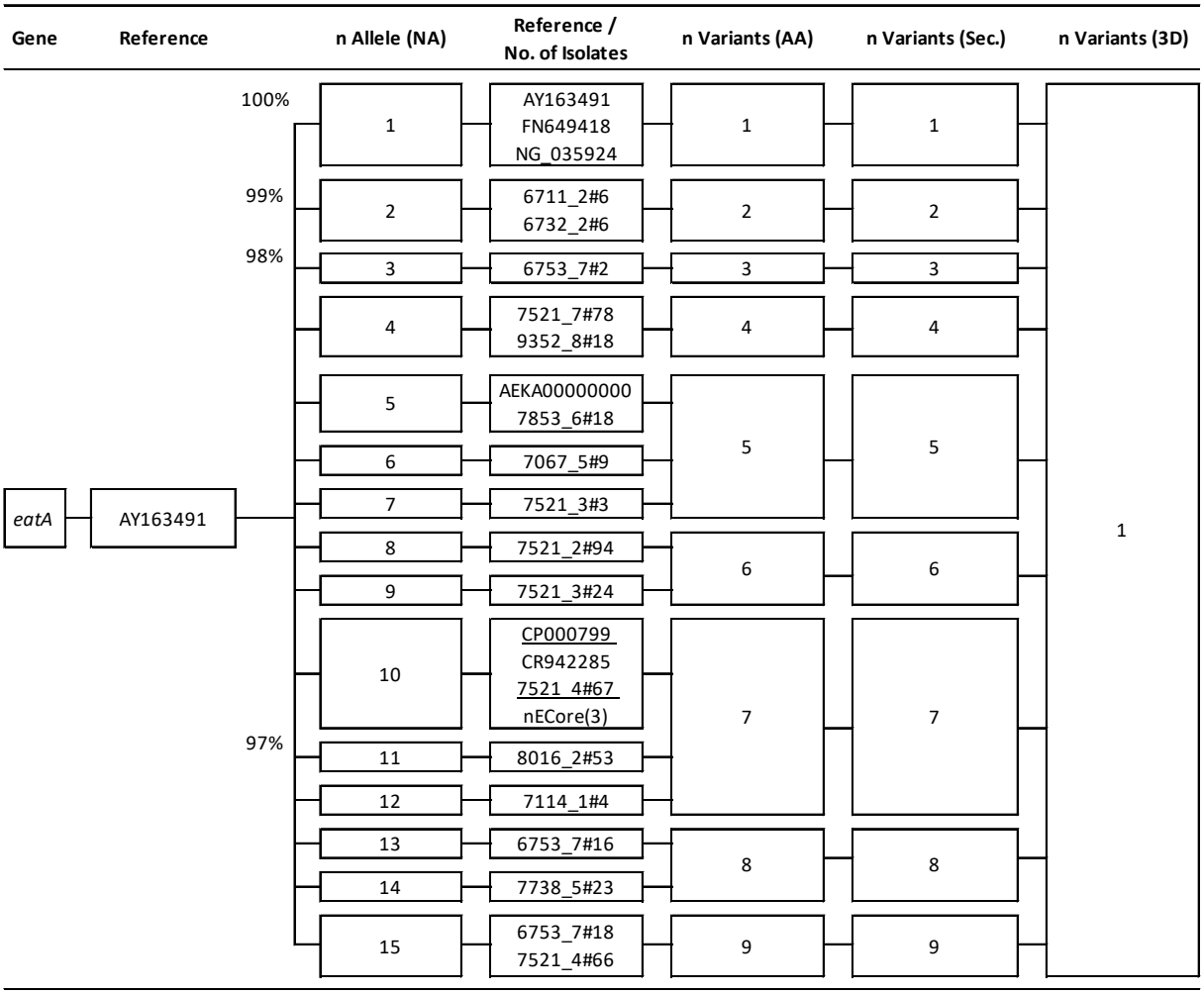

Figure A 5: *eatA* (SPATE – ETEC autotransporter A) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AY163491 (genetic identity in percent).

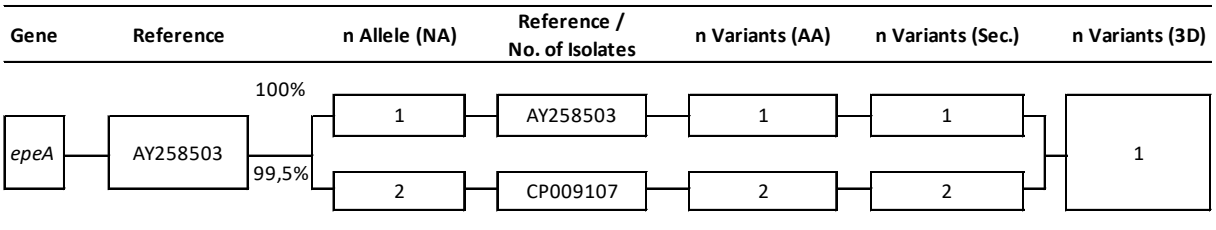

Figure A 6: *epeA* (SPATE –EHEC plasmid encoded autotransporter) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AJ258503 (genetic identity in percent).

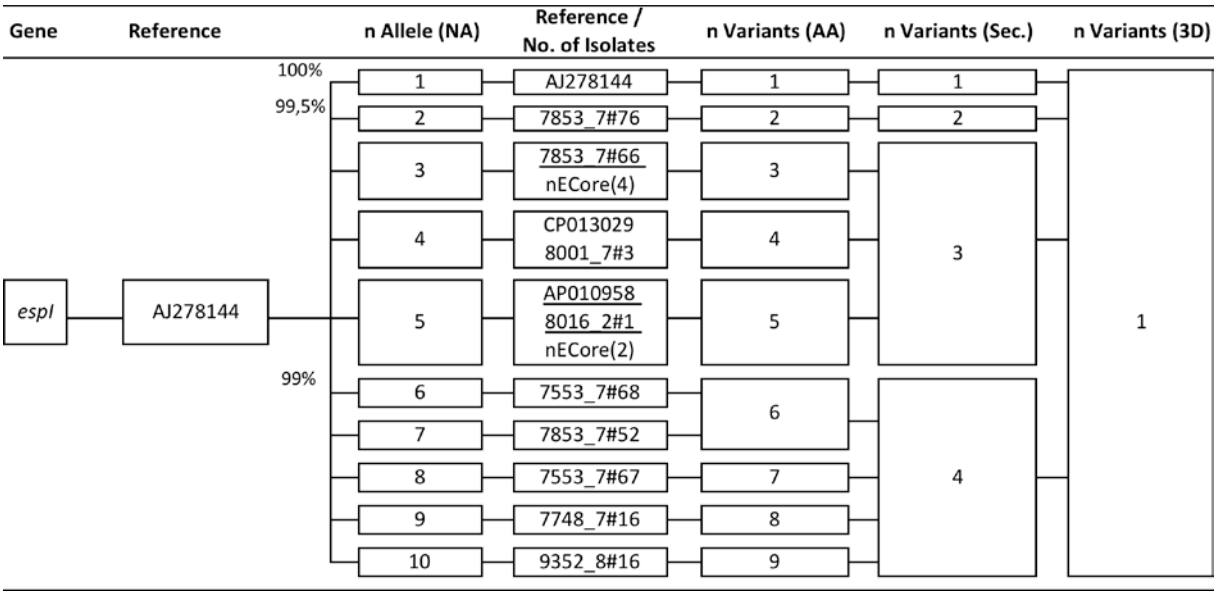

Figure A 7: *espI* (SPATE – *E. coli* secreted protease I) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AJ278144 (genetic identity in percent).

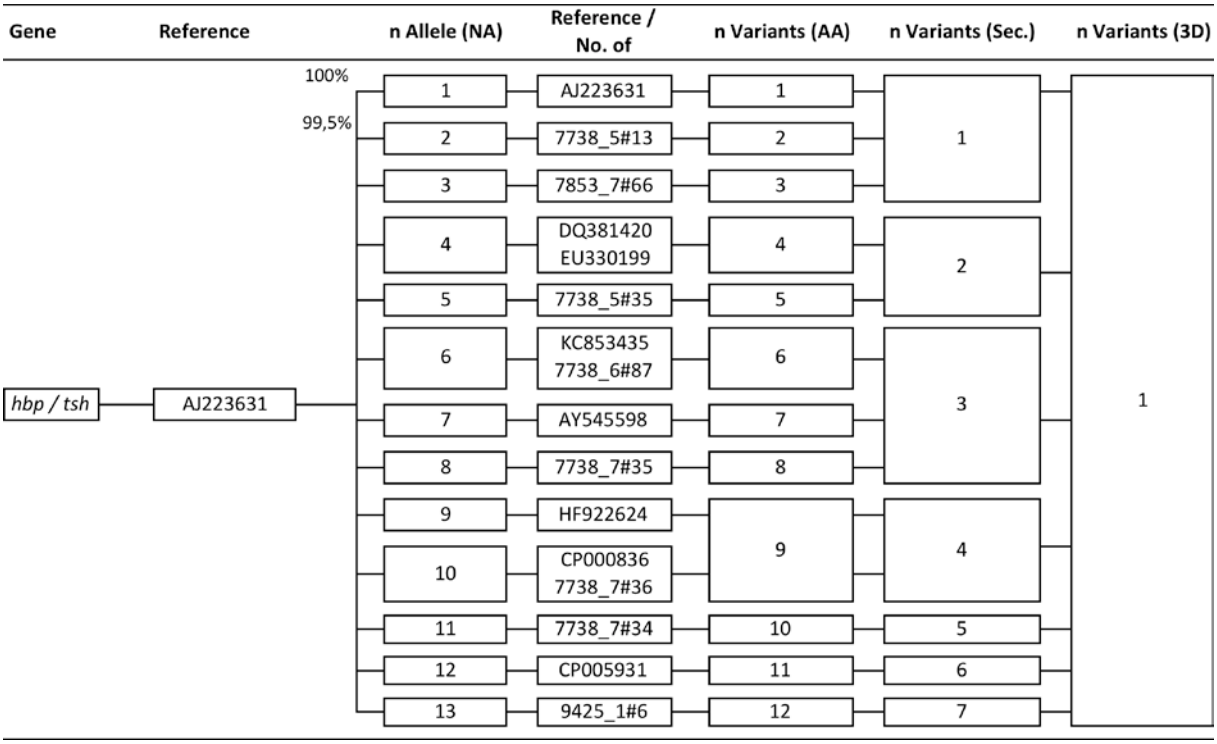

Figure A 8: *hbp / tsh* (SPATE – Hemoglobin binding protein / temperature sensitive hemagglutinin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AJ223631 (genetic identity in percent).

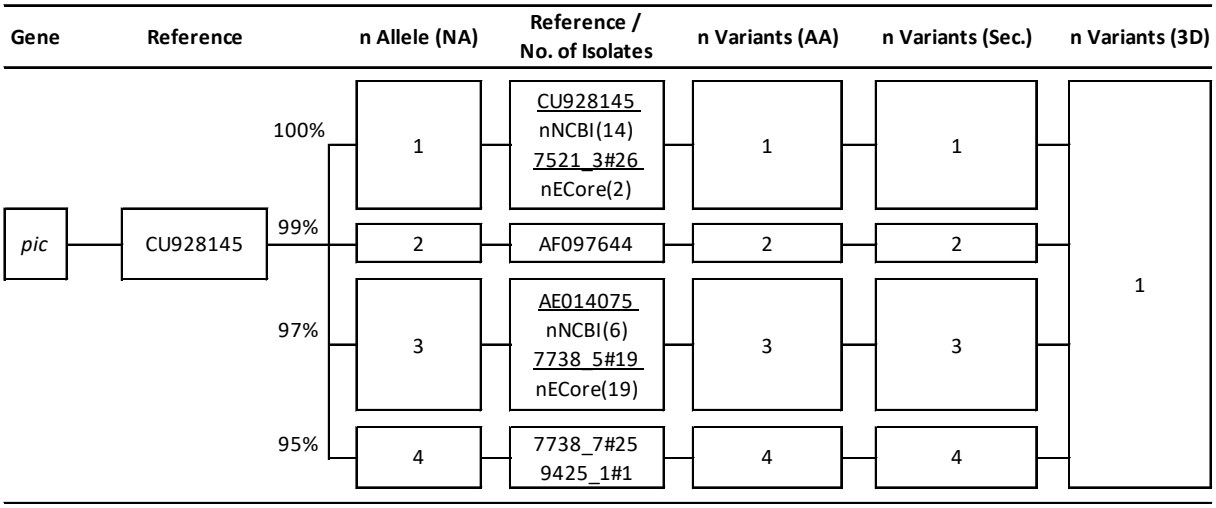

Figure A 9: *pic* (SPATE – Protease involved in intestinal colonization) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence CU928145 (genetic identity in percent).

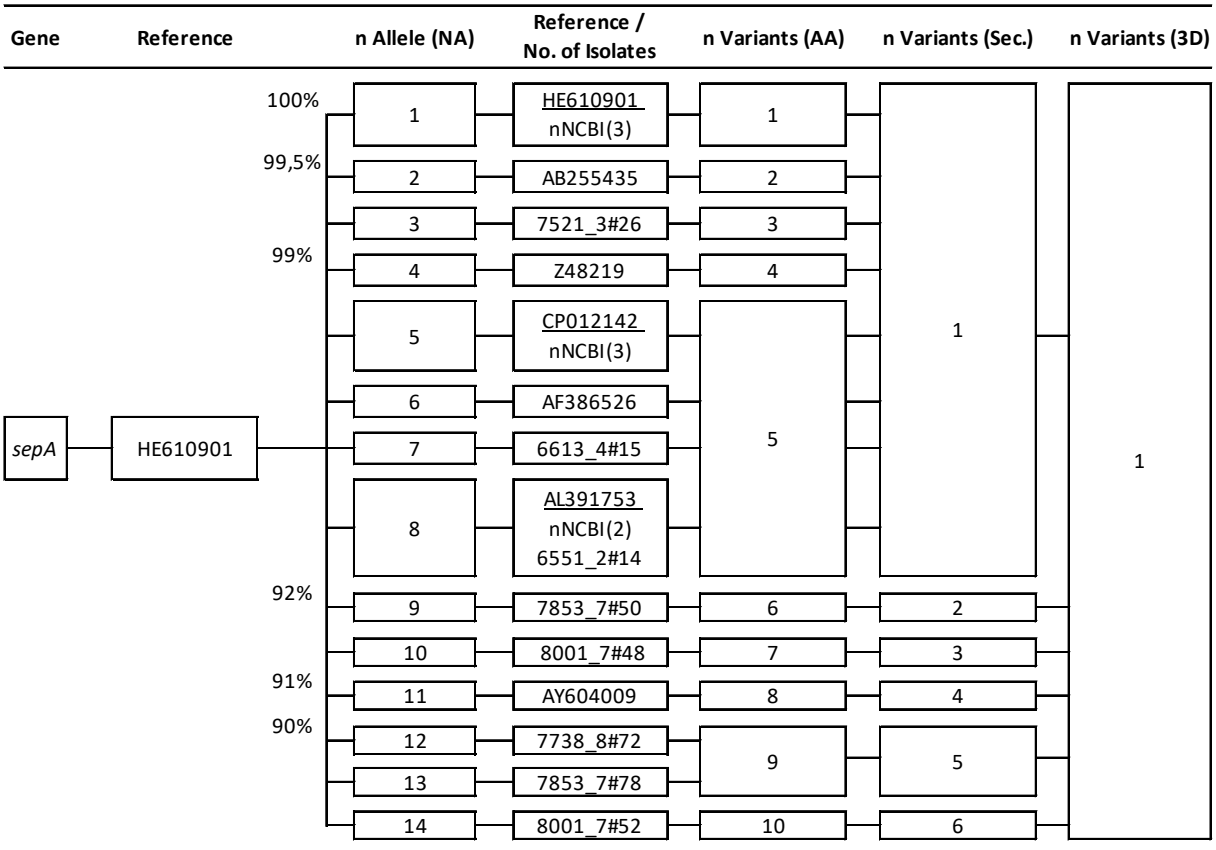

Figure A 10: *sepA* (SPATE – *Shigella* extracellular protein) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence HE610901 (genetic identity in percent).

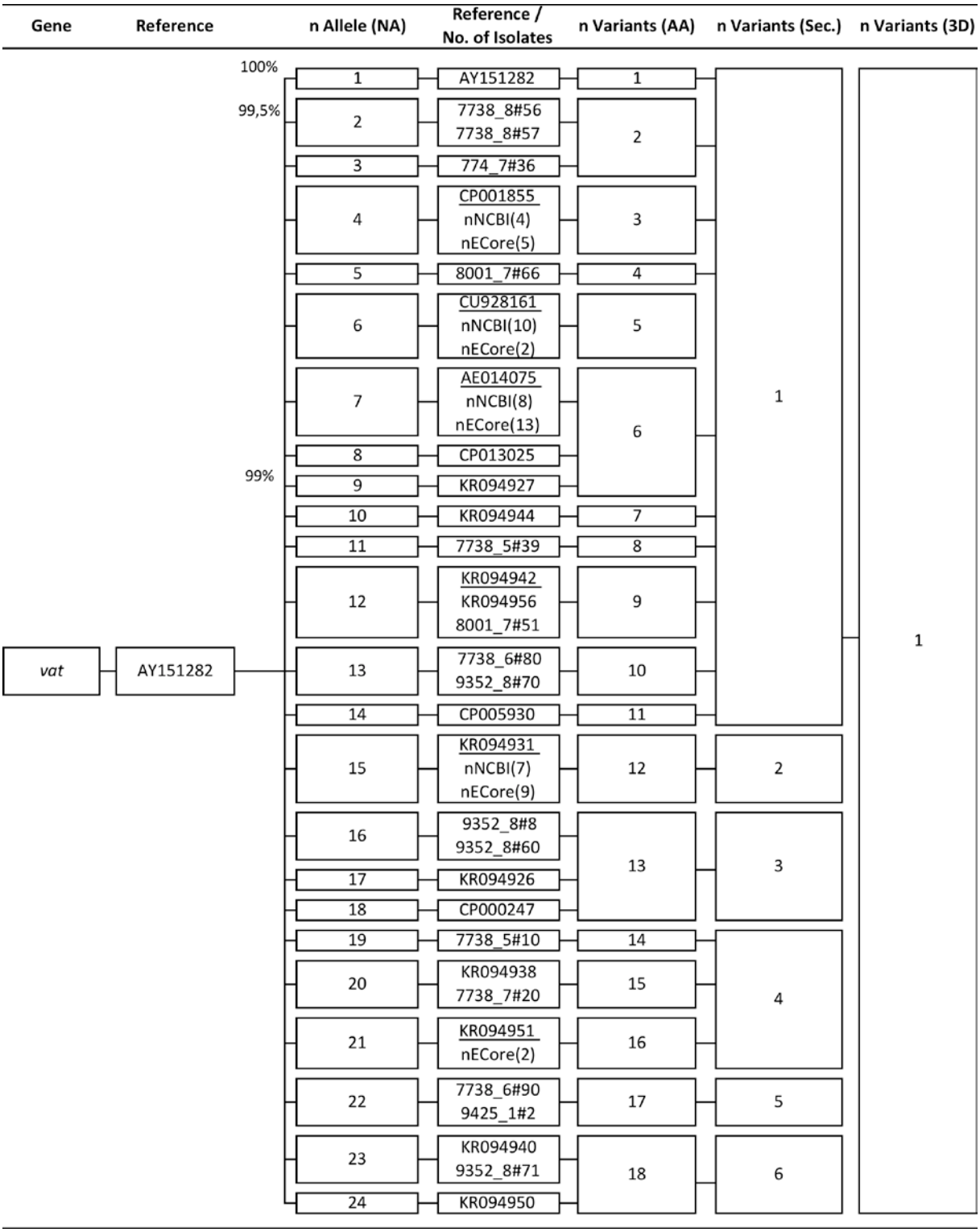

Figure A 11: *vat* (SPATE – Vacuolating autotransporter toxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AY151282 (genetic identity in percent).

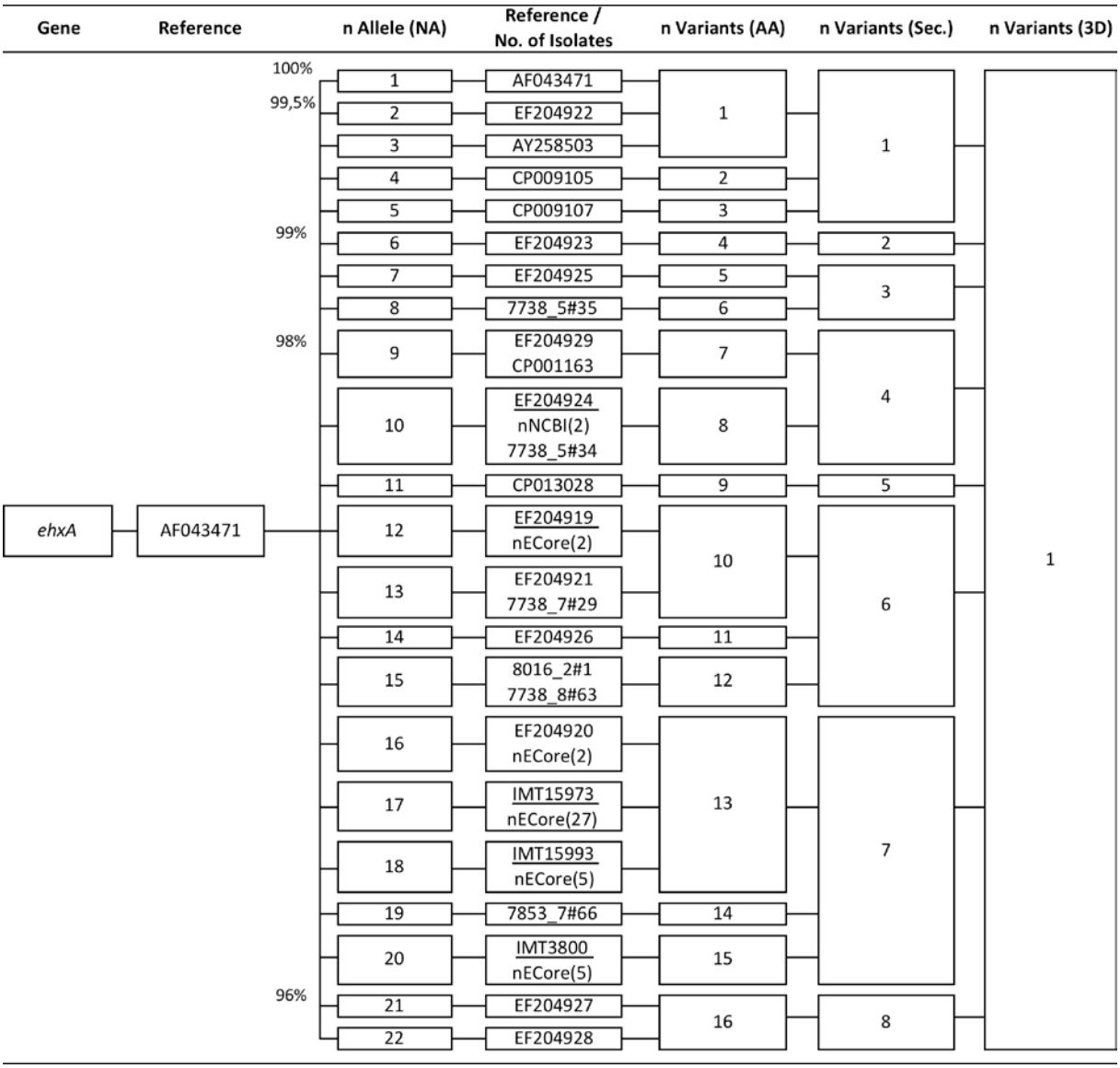

Figure A 12: *ehxA* (EHEC hemolysin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF043471 (genetic identity in percent).

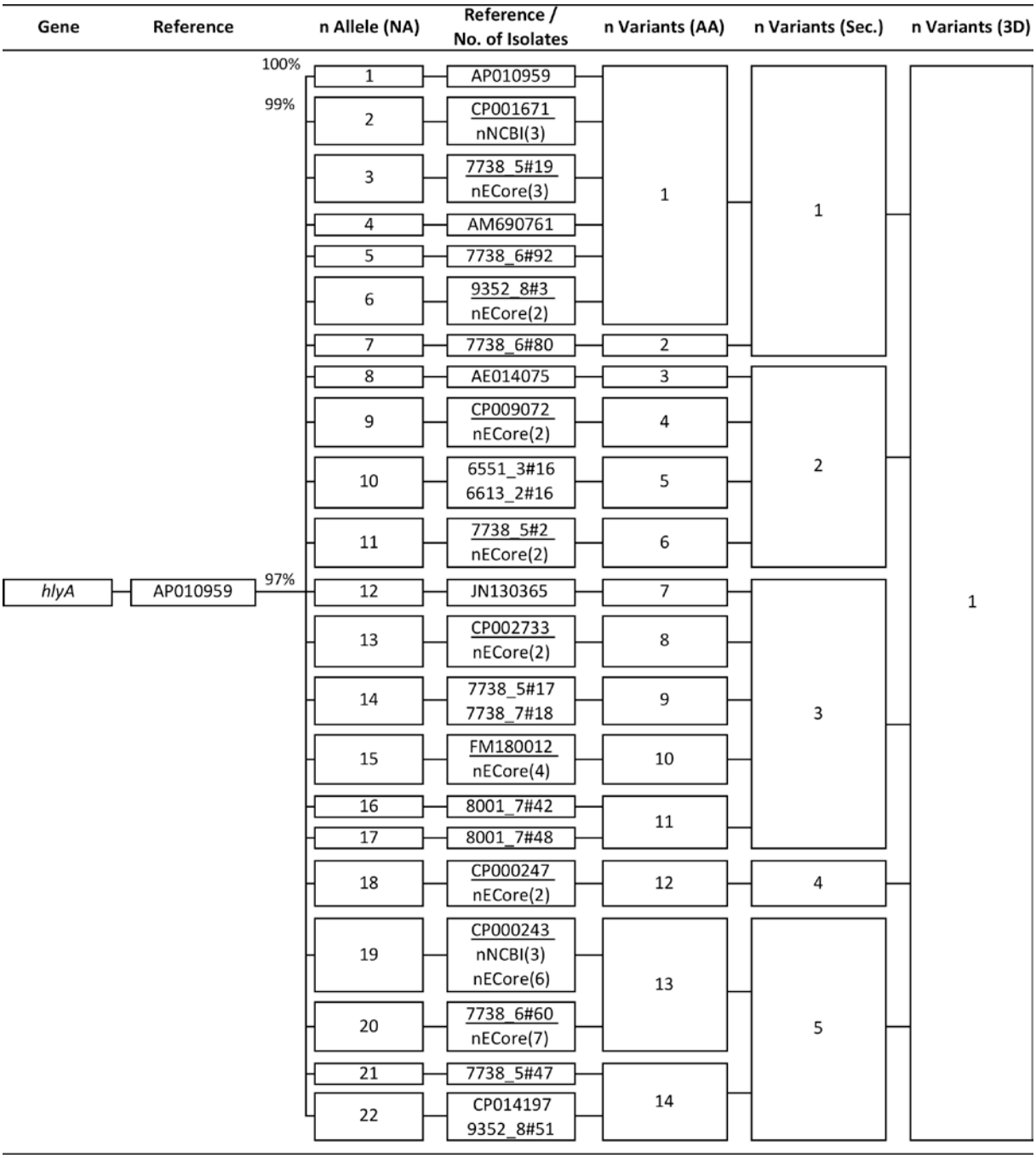

Figure A 13: *hlyA* (Haemolysin A) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AP010959 (genetic identity in percent).

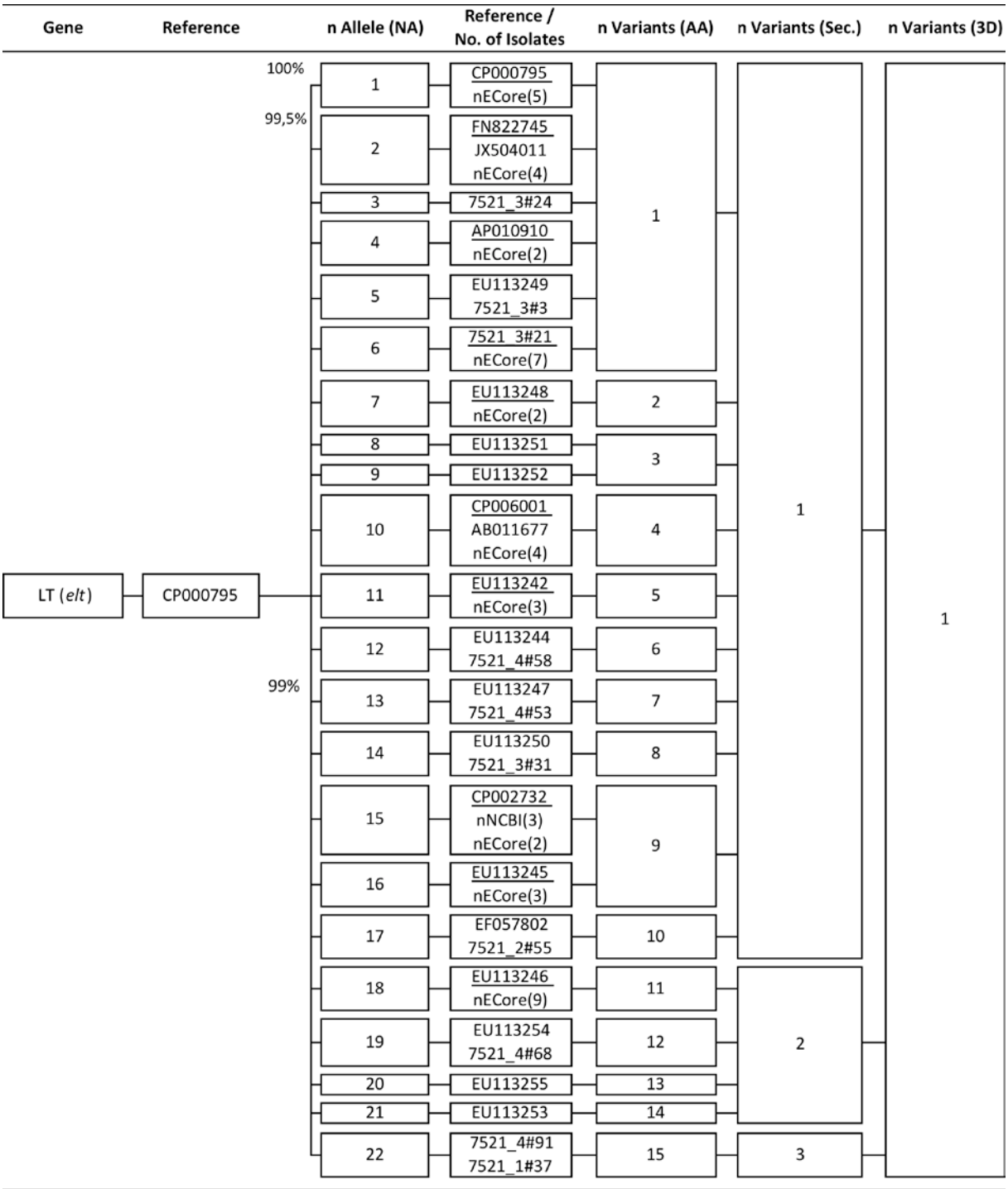

Figure A 14: LT (*elt*) (Heat-labile enterotoxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence CP000795 (genetic identity in percent).

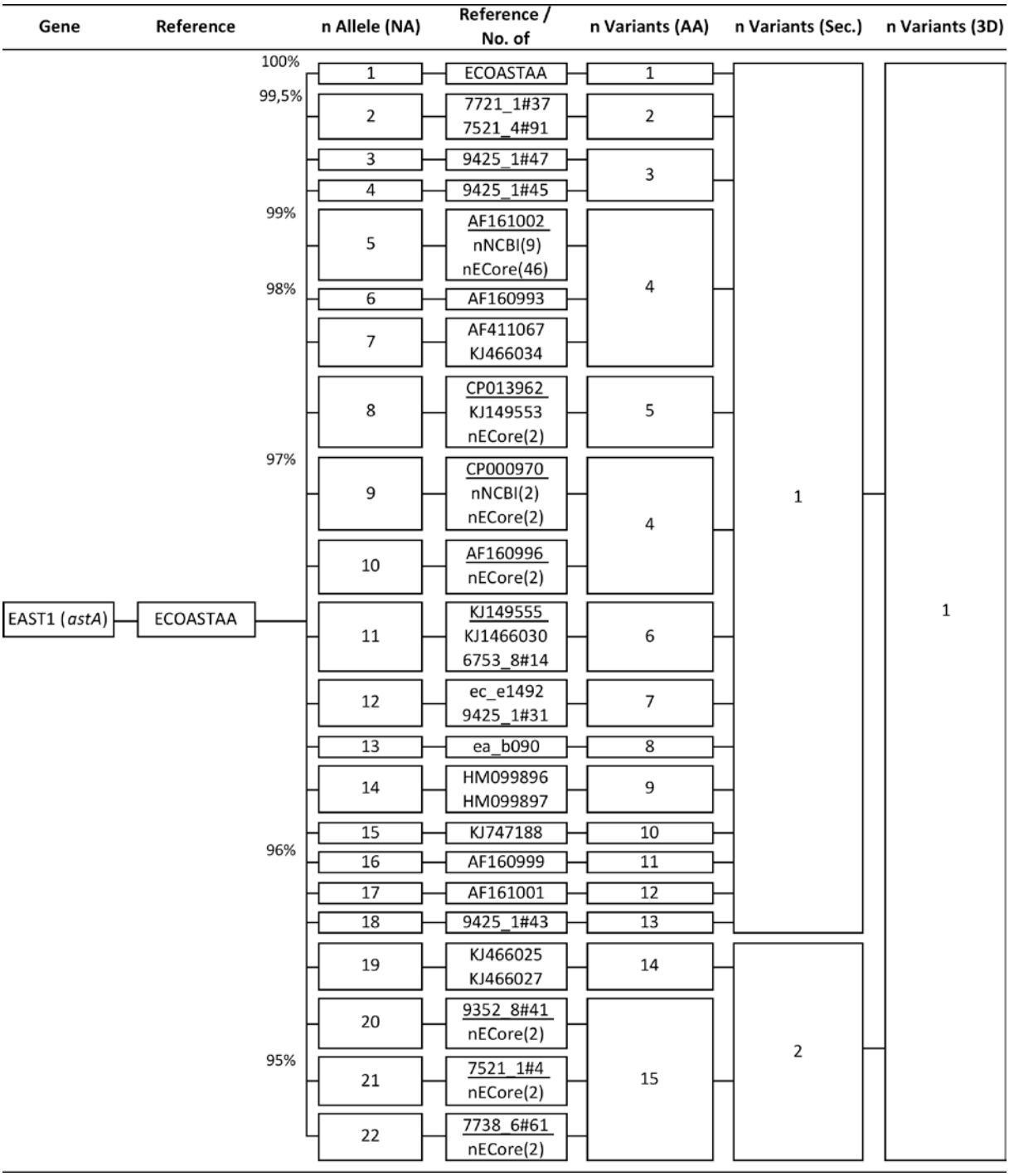

Figure A 15: EAST1 (*astA*) (SPATE – Heat-stabile enterotoxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence ECOASTAA (genetic identity in percent).

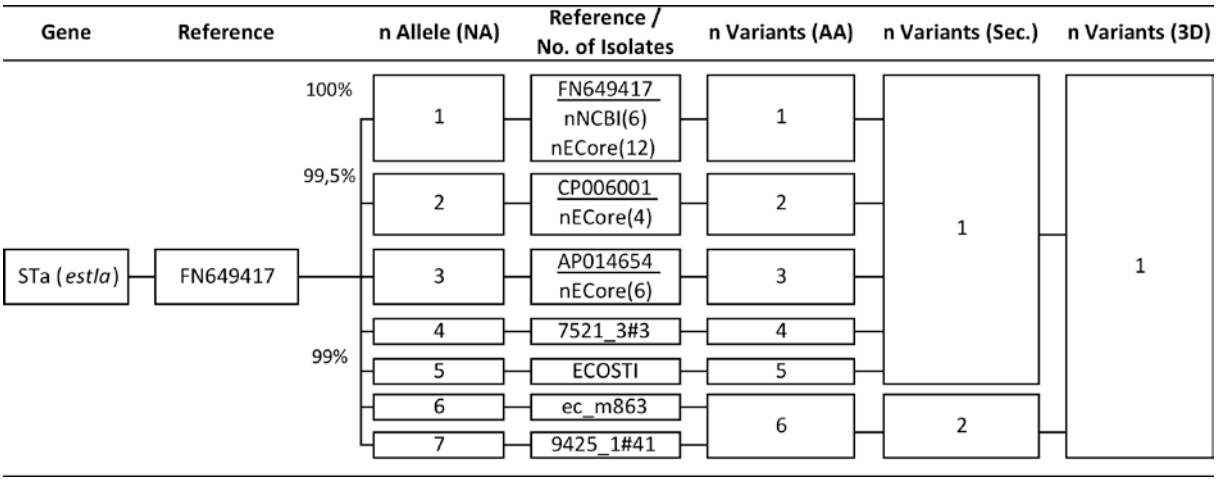

Figure A 16: STa (*estIa*) (Heat-stabile enterotoxin a) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence FN649417 (genetic identity in percent).

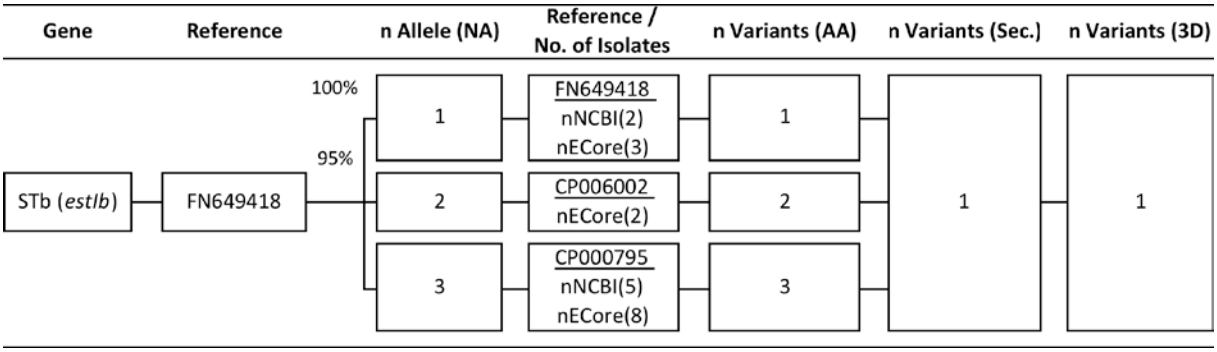

Figure A 17: STb (*estIb*) (Heat-stabile enterotoxin b) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence FN649418 (genetic identity in percent).

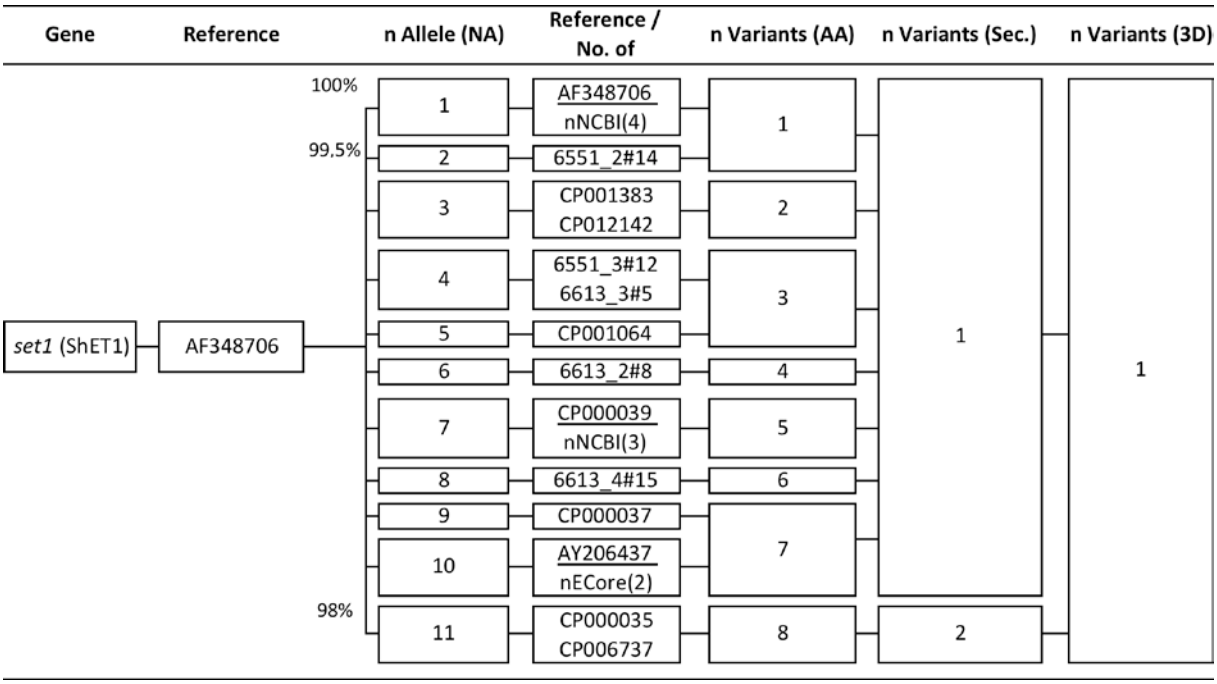

Figure A 18: *set1* (ShET1) (*Shigella* enterotoxin 1) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF348706 (genetic identity in percent).

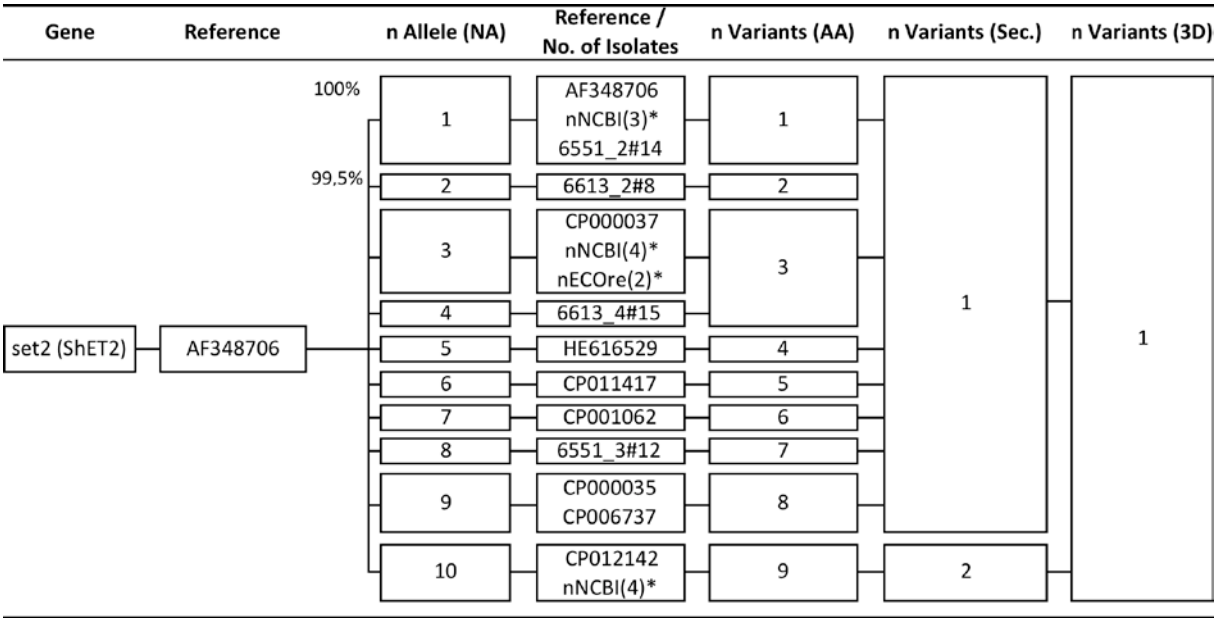

Figure A 19: *set2* (ShET2) (*Shigella* enterotoxin 2) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF348706 (genetic identity in percent).

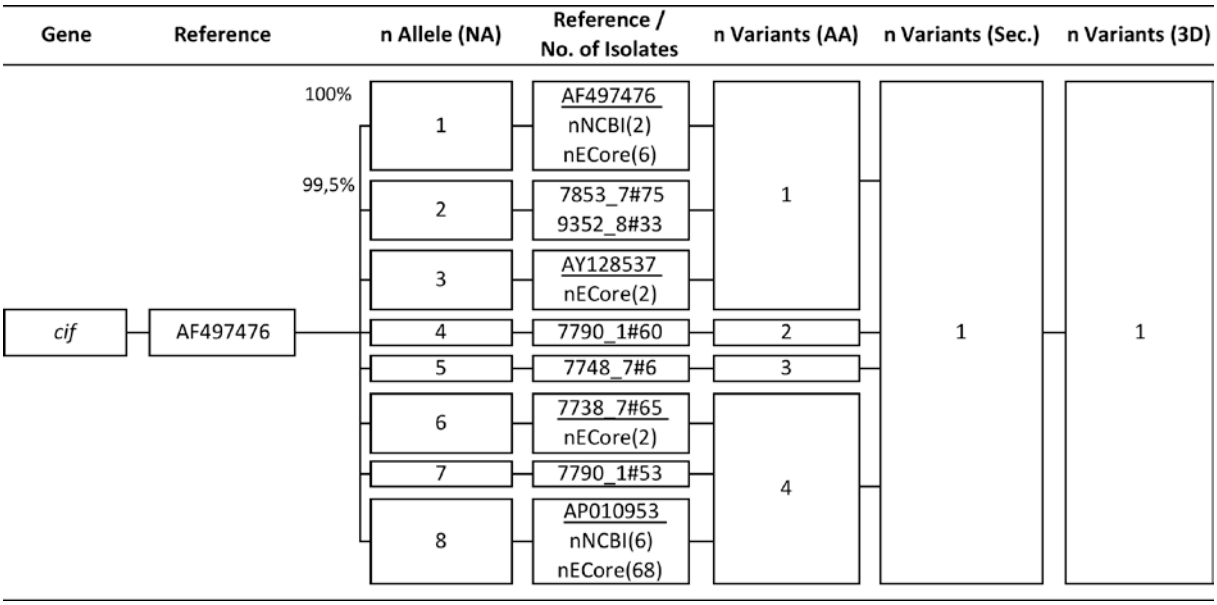

Figure A 20: *cif* (Cycle-inhibiting factor) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF497476 (genetic identity in percent).

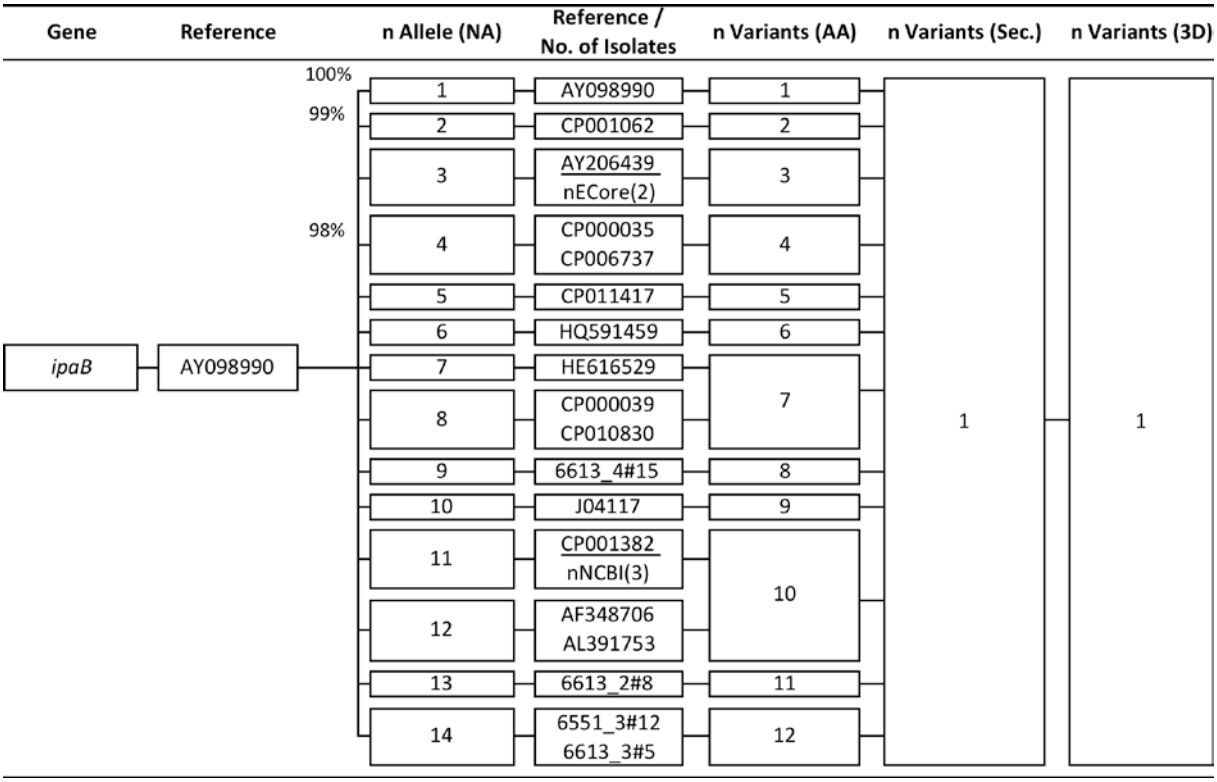

Figure A 21: *ipaB* (Invasion plasmid antigen B) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AY098990 (genetic identity in percent).

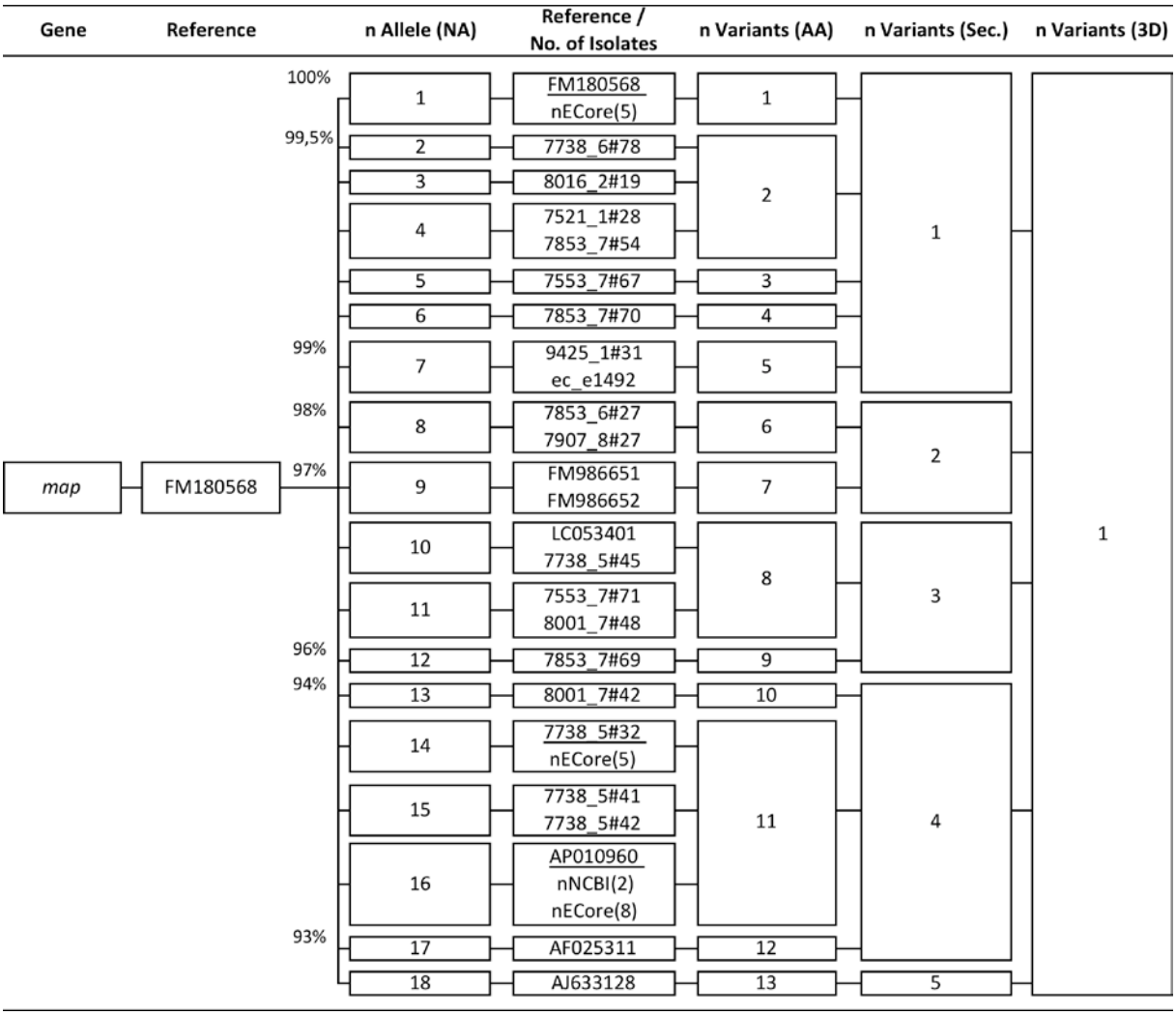

Figure A 22: *map* (Methionine aminopeptidase) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence FM180568 (genetic identity in percent).

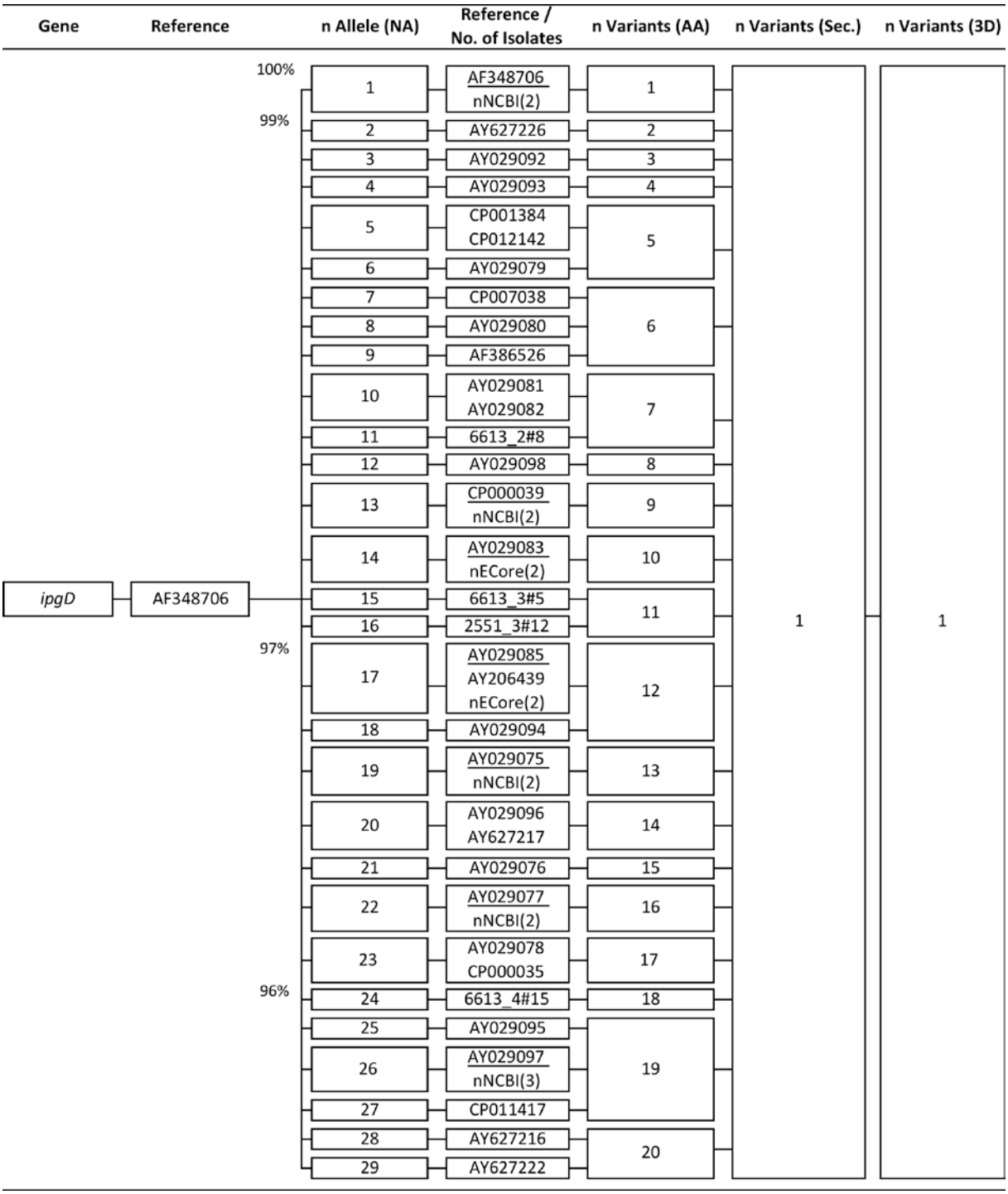

Figure A 23: *ipgD* (Inositol phosphate phosphatase D) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF348706 (genetic identity in percent).

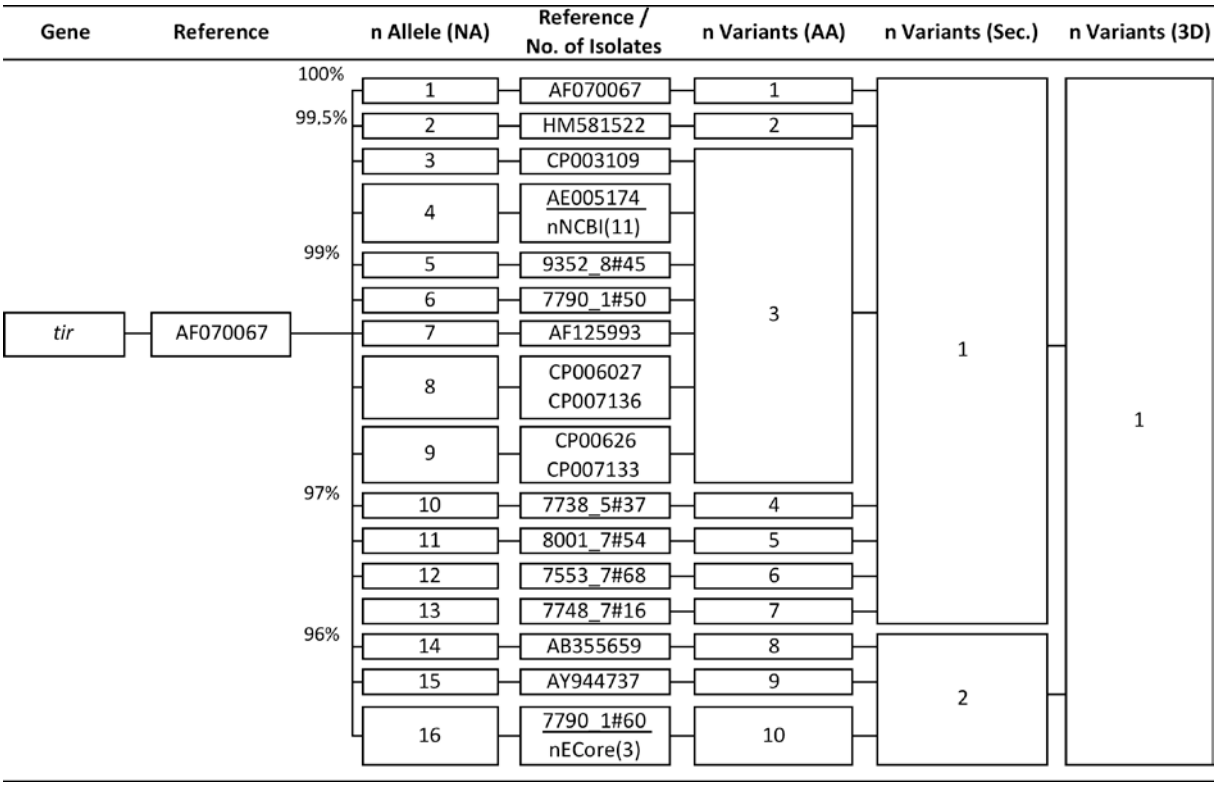

Figure A 24: *tir* (Translocated intimin receptor) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF070067 (genetic identity in percent).

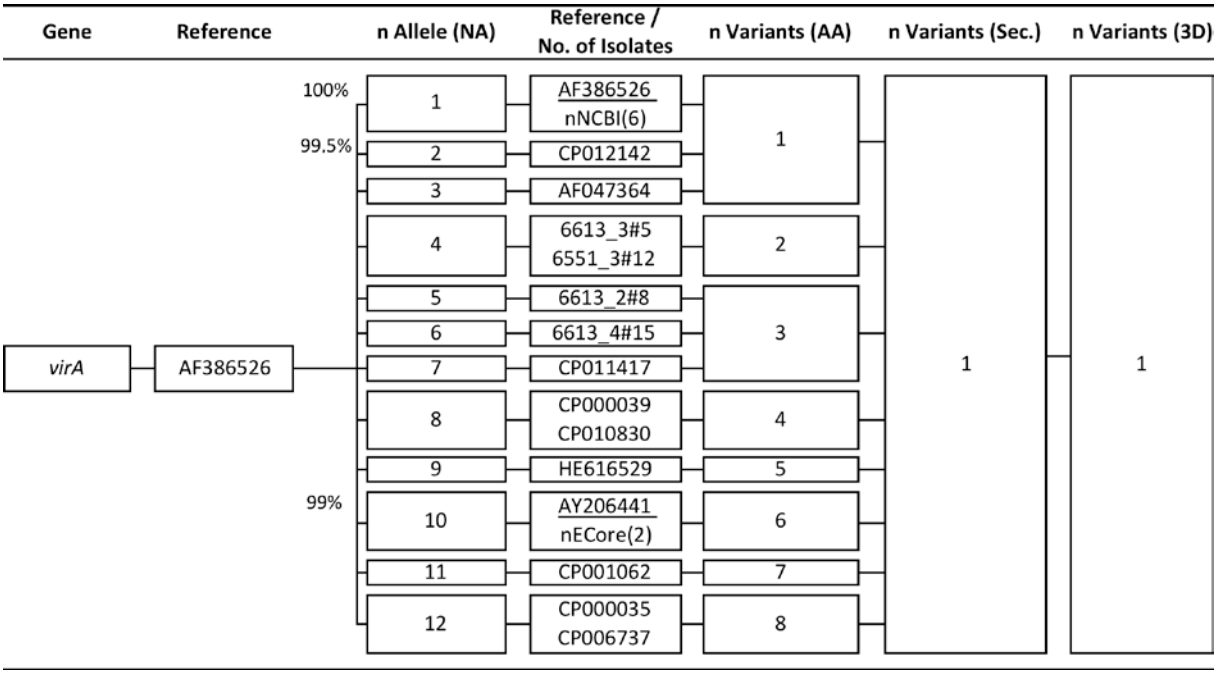

Figure A 25: *virA* (Typ II effector protein) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AF386526 (genetic identity in percent).

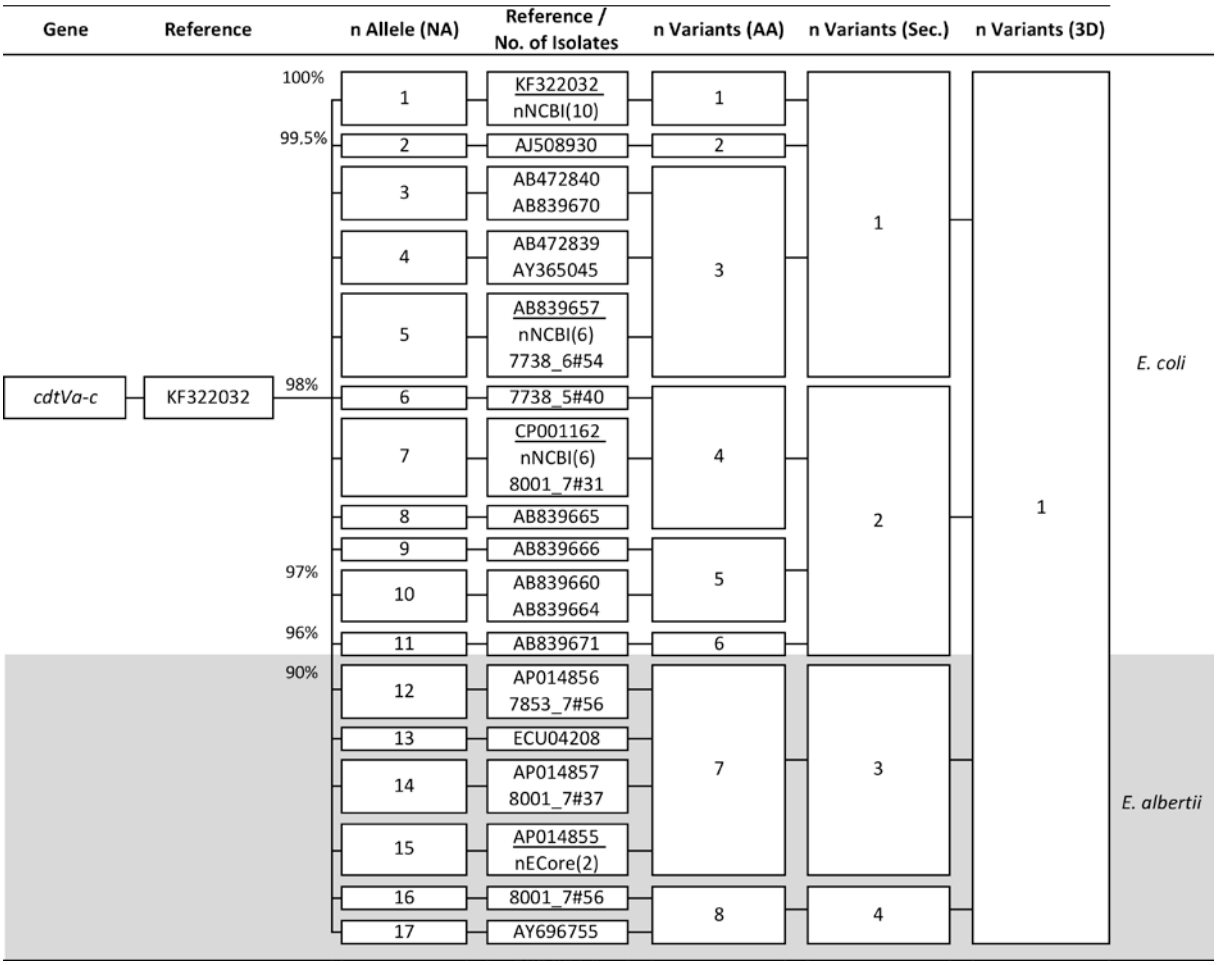

Figure A 26: *cdtVa-c* (Cytoletal distending toxin) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence KF322032 (genetic identity in percent).

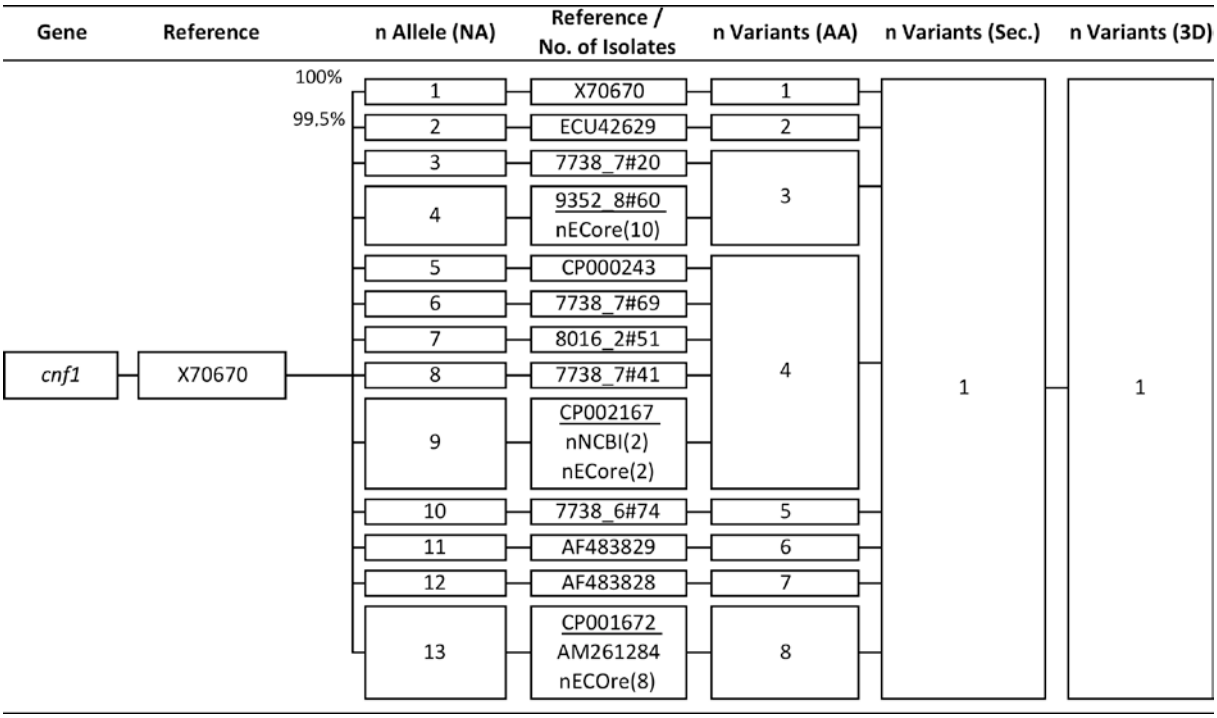

Figure A 27: *cnf1* (Cytotoxin necrotizing factor 1) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence X70670 (genetic identity in percent).

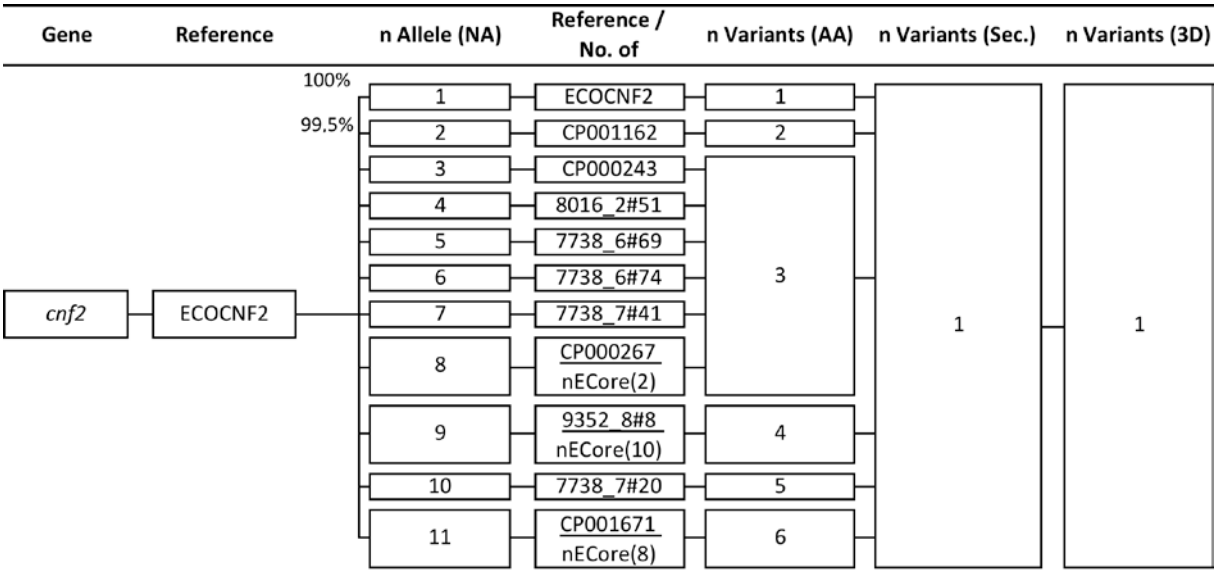

Figure A 28: *cnf2* (Cytotoxin necrotizing factor 2) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence ECOCNF2 (genetic identity in percent).

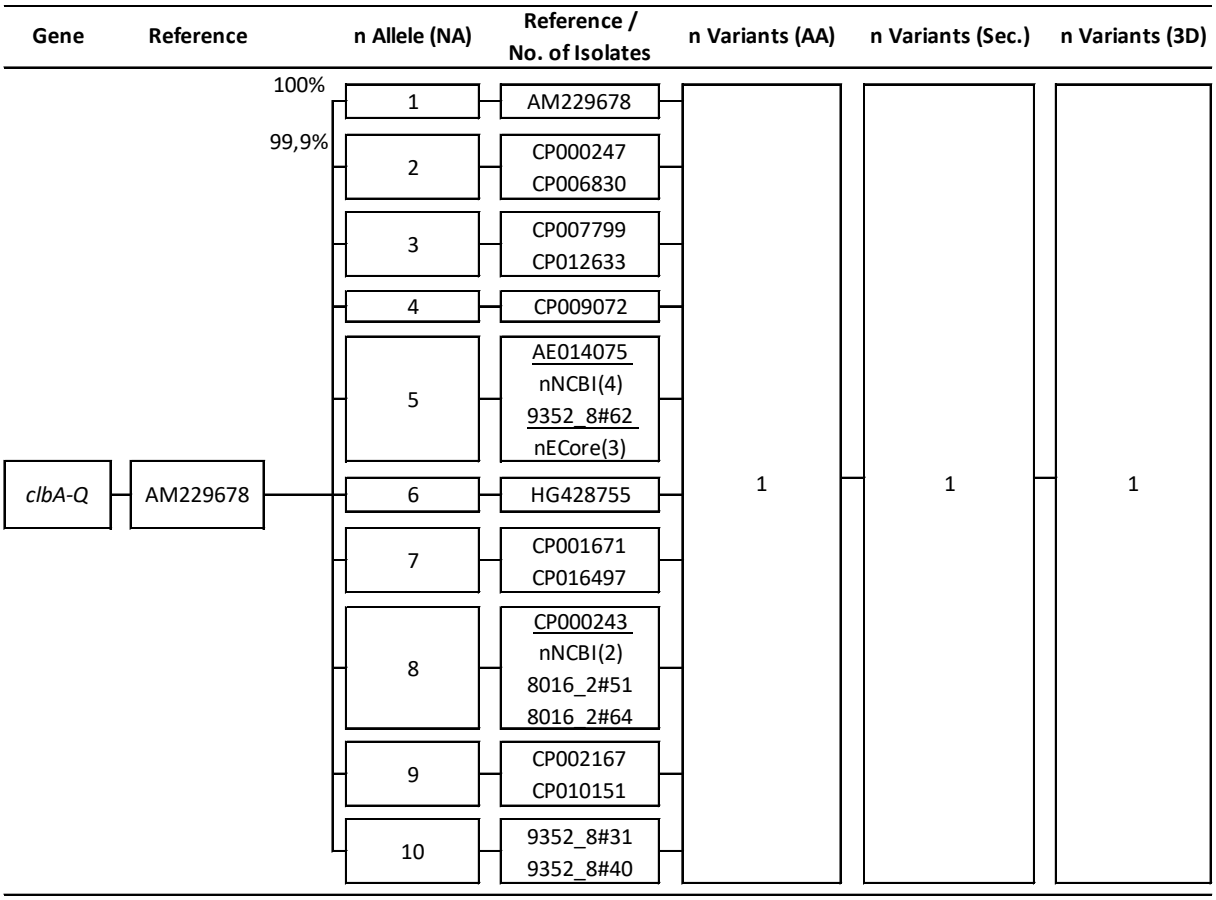

Figure A 29: *clbA-Q* (Colibactin locus) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AM229678 (genetic identity in percent).

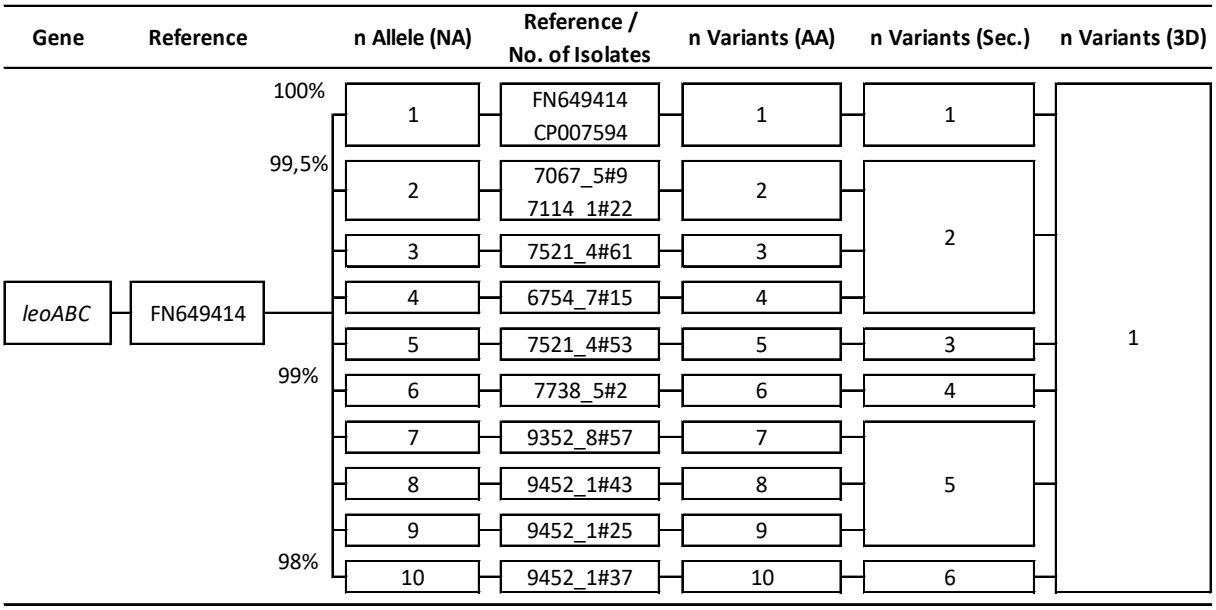

Figure A 30: *leoABC* (Dynamin-like protein) sequence analysis and phylogenetic distribution of the gene (NA: nucleic acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence FN649414 (genetic identity in percent).

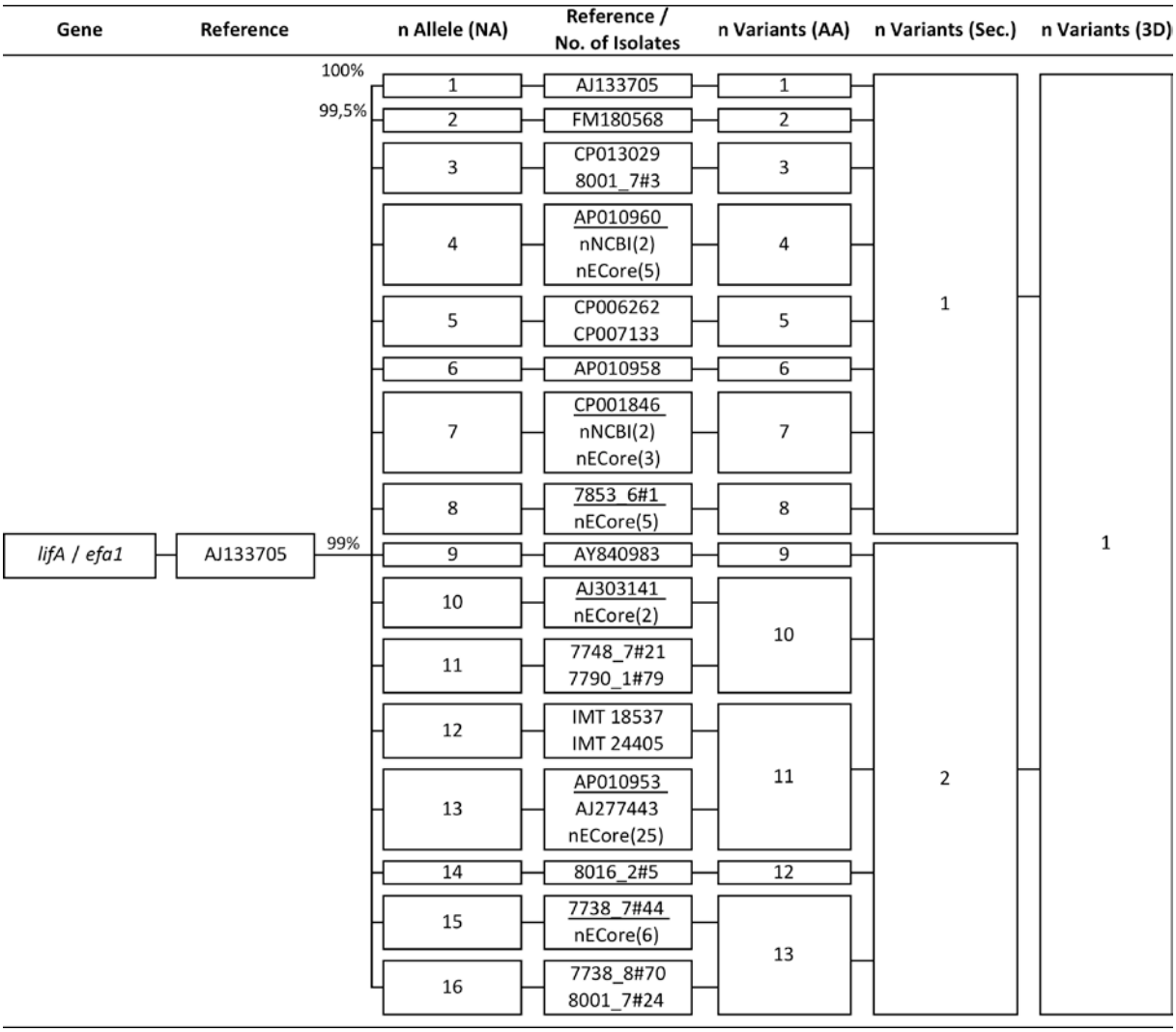

Figure A 31: *lifA / efa1* (Lymphocyte inhibitory factor / EHEC factor for adherence) sequence analysis and phylogenetic distribution of the gene (NA: nucleice acid) and protein (AA: amino acid) sequence as well as predicted secondary (Sec.) and three-dimensional (3D) structure to identify alleles and variants (n: number of alleles / variants) with respect to the reference sequence AJ133705 (genetic identity in percent).