GENOMICS OF PATHOGENIC AND COMMENSAL *ESCHERICHIA COLI*

GENOMIK PATHOGENER UND KOMMENSALER *ESCHERICHIA COLI*

Doctoral thesis for a doctoral degree
at the Graduate School of Life Sciences,
Julius-Maximilians-University of Würzburg,
Section: Infection and Immunity

submitted by
ANDREAS LEIMBACH

from
Schweinfurt

Würzburg 2017

SUBMITTED ON: ....

Office stamp

MEMBERS OF THE *PROMOTIONSKOMITEE*:

Chairperson:	*Prof. Dr. Jörg Schultz*
Primary Supervisor:	*Prof. Dr. Ulrich Dobrindt*
Supervisor (Second):	*Prof. Dr. Rolf Daniel*
Supervisor (Third):	*PD Dr. Knut Ohlsen*

DATE OF PUBLIC DEFENCE: ....

DATE OF RECEIPT OF CERTIFICATES: ....

Dedicated to Dr. Rainer Leimbach
1948 – 2015

SUMMARY

High-throughput sequencing (HTS) has revolutionized bacterial genomics. Its unparalleled sensitivity has opened the door to analyzing bacterial evolution and population genomics, dispersion of mobile genetic elements (MGEs), and within-host adaptation of pathogens, such as *Escherichia coli*.

One of the defining characteristics of intestinal pathogenic *E. coli* (IPEC) pathotypes is a specific repertoire of virulence factors (VFs). Many of these IPEC VFs are used as typing markers in public health laboratories to monitor outbreaks and guide treatment options. Instead, extraintestinal pathogenic *E. coli* (ExPEC) isolates are genotypically diverse and harbor a varied set of VFs – the majority of which also function as fitness factors (FFs) for gastrointestinal colonization.

The aim of this thesis was the genomic characterization of pathogenic and commensal *E. coli* with respect to their virulence- and antibiotic resistance-associated gene content as well as phylogenetic background. In order to conduct the comparative analyses, I created a database of *E. coli* VFs, `ecoli_VF_collection`, with a focus on ExPEC virulence-associated proteins (Leimbach, 2016b). Furthermore, I wrote a suite of scripts and pipelines, `bac-genomics-scripts`, that are useful for bacterial genomics (Leimbach, 2016a). This compilation includes tools for assembly and annotation as well as comparative genomics analyses, like multi-locus sequence typing (MLST), assignment of Clusters of Orthologous Groups (COG) categories, searching for protein homologs, detection of genomic regions of difference (RODs), and calculating pan-genome-wide association statistics.

Using these tools we were able to determine the prevalence of 18 autotransporters (ATs) in a large, phylogenetically heterogeneous strain panel and demonstrate that many AT proteins are not associated with *E. coli* pathotypes. According to multivariate analyses and statistics the distribution of AT variants is instead significantly dependent on phylogenetic lineages. As a consequence, ATs are not suitable to serve as pathotype markers (Zude et al., 2014).

During the German Shiga toxin-producing *E. coli* (STEC) outbreak in 2011, the largest to date, we were one of the teams capable of analyzing the genomic features of two isolates. Based on MLST and detection of orthologous proteins to known *E. coli* reference genomes the close phylogenetic relationship and overall genome similarity to enteroaggregative *E. coli* (EAEC) 55989 was revealed. In particular, we identified VFs of both STEC and EAEC pathotypes, most importantly the prophage-encoded Shiga toxin (Stx) and the pAA-type plasmid harboring aggregative adherence fimbriae. As a result, we could show

that the epidemic was caused by an unusual hybrid pathotype of the O104:H4 serotype. Moreover, we detected the basis of the antibiotic multi-resistant phenotype on an extended-spectrum β-lactamase (ESBL) plasmid through comparisons to reference plasmids. With this information we proposed an evolutionary horizontal gene transfer (HGT) model for the possible emergence of the pathogen (Brzuszkiewicz et al., 2011).

Similarly to ExPEC, *E. coli* isolates of bovine mastitis are genotypically and phenotypically highly diverse and many studies struggled to determine a positive association of putative VFs. Instead the general *E. coli* pathogen-associated molecular pattern (PAMP), lipopolysaccharide (LPS), is implicated as a deciding factor for intramammary inflammation. Nevertheless, a mammary pathogenic *E. coli* (MPEC) pathotype was proposed presumably encompassing strains more adapted to elicit bovine mastitis with virulence traits differentiating them from commensals.

We sequenced eight *E. coli* isolates from udder serous exudate and six fecal commensals (Leimbach et al., 2016). Two mastitis isolate genomes were closed to a finished-grade quality (Leimbach et al., 2015). The genomic sequence of mastitis-associated *E. coli* (MAEC) strain 1303 was used to elucidate the biosynthesis gene cluster of its O70 LPS O-antigen. We analyzed the phylogenetic genealogy of our strain panel plus eleven bovine-associated *E. coli* reference strains and found that commensal or MAEC could not be unambiguously allocated to specific phylogroups within a core genome tree of reference *E. coli*. A thorough gene content analysis could not identify functional convergence of either commensal or MAEC, instead both have only very few gene families enriched in either pathotype. Most importantly, gene content and `ecoli_VF_collection` analyses showed that no virulence determinants are significantly associated with MAEC in comparison to bovine fecal commensals, disproving the MPEC hypothesis. The genetic repertoire of bovine-associated *E. coli*, again, is dominated by phylogenetic background. This is also mostly the case for large virulence-associated *E. coli* gene cluster previously associated with mastitis. Correspondingly, MAEC are facultative and opportunistic pathogens recruited from the bovine commensal gastrointestinal microbiota (Leimbach et al., 2017). Thus, *E. coli* mastitis should be prevented rather than treated, as antibiotics and vaccines have not proven effective.

Although traditional *E. coli* pathotypes serve a purpose for diagnostics and treatment, it is clear that the current typing system is an oversimplification of *E. coli*'s genomic plasticity. Whole genome sequencing (WGS) revealed many nuances of pathogenic *E. coli*, including emerging hybrid or heteropathogenic pathotypes. Diagnostic and public health microbiology need to embrace the future by implementing HTS techniques to target patient care and infection control more efficiently.

ZUSAMMENFASSUNG

Eines der definierenden Charakteristika intestinal pathogener *E. coli* (IPEC) Pathotypen ist ein spezifisches Repertoire an Virulenzfaktoren (VFs). Viele dieser IPEC VFs werden als Typisierungsmarker benutzt. Stattdessen sind Isolate extraintestinal pathogener *E. coli* (ExPEC) genotypisch vielfältig und beherbergen verschiedenartige VF Sets, welche in der Mehrheit auch als Fitnessfaktoren (FFs) für die gastrointestinale Kolonialisierung fungieren.

Das Ziel dieser Dissertation war die genomische Charakterisierung pathogener und kommensaler *E. coli* in Bezug auf ihren Virulenz- und Antibiotikaresistenz-assoziierten Gengehalt sowie ihre phylogenetische Abstammung. Als Voraussetzung für die vergleichenden Analysen erstellte ich eine *E. coli* VF-Datenbank, `ecoli_VF_collection`, mit Fokus auf Virulenz-assoziierte Proteine von ExPEC (Leimbach, 2016b). Darüber hinaus programmierte ich mehrere Skripte und Pipelines zur Anwendung in der bakteriellen Genomik, `bac-genomics-scripts` (Leimbach, 2016a). Diese Sammlung beinhaltet Tools zur Unterstützung von Assemblierung und Annotation sowie komparativer Genomanalysen, wie Multilokus-Sequenztypisierung (MLST), Zuweisung von Clusters of Orthologous Groups (COG) Kategorien, Suche nach homologen Proteinen, Identifizierung von genomisch unterschiedlichen Regionen (RODs) und Berechnung Pan-genomweiter Assoziationsstatistiken.

Mithilfe dieser Tools konnten wir die Prävalenz von 18 Autotransportern (ATs) in einer großen, phylogenetisch heterogenen Stammsammlung bestimmen und nachweisen, dass viele AT-Proteine nicht mit *E. coli* Pathotypen assoziiert sind. Multivariate Analysen und Statistik legten offen, dass die Verteilung von AT-Varianten vielmehr signifikant von phylogenetischen Abstammungslinien abhängt. Deshalb sind ATs nicht als Marker für Pathotypen geeignet (Zude et al., 2014).

Während des bislang größten Ausbruchs von Shiga-Toxin-produzierenden *E. coli* (STEC) im Jahre 2011 in Deutschland waren wir eines der Teams, welches die genomischen Eigenschaften zweier Isolate analysieren konnte. Basierend auf MLST und Detektion orthologer Proteine zu bekannten *E. coli* Referenzgenomen konnte ihre enge phylogenetische Verwandschaft und Ähnlichkeit des gesamten Genoms zum enteroaggregativen *E. coli* (EAEC) 55989 aufgedeckt werden. Im Detail identifizierten wir VFs von STEC und EAEC Pathotypen, vor allem das Prophagen-kodierte Shiga-Toxin (Stx) und ein Plasmid des pAA-Typs kodierend für aggregative Adhärenz-Fimbrien. Die Epidemie wurde demnach durch einen ungewöhnlichen Hybrid-Pathotyp vom O104:H4 Serotyp verursacht. Zusätzlich identifizierten wir die Grundlage für den multiresistenten Phänotyp dieser Ausbruchsstäm-

me auf einem Extended-Spektrum-β-Laktamase (ESBL) Plasmid über Vergleiche mit Referenzplasmiden. Mit diesen Informationen konnten wir ein horizontales Gentransfer-Modell (HGT) zum Auftreten dieses Pathogenen vorschlagen (Brzuszkiewicz et al., 2011).

Ähnlich zu ExPEC sind *E. coli* Isolate boviner Mastitiden genotypisch und phänotypisch sehr divers, und viele Studien scheiterten am Versuch eine positive Assoziation vermeintlicher VFs nachzuweisen. Stattdessen gilt Lipopolysaccharid (LPS) als entscheidender Faktor zur intramammären Entzündung. Gleichwohl wurde ein mammärer pathogener *E. coli* (MPEC) Pathotyp vorgeschlagen, der mutmaßlich Stämme umfasst, welche eher geeignet sind eine bovine Mastitis auszulösen und über Virulenz-Merkmale von Kommensalen abgegrenzt werden können.

Wir sequenzierten acht *E. coli* Isolate aus serösem Eutersekret und sechs fäkale Kommensale (Leimbach et al., 2016). Bei zwei Mastitisisolaten wurden die Genome vollständig geschlossen (Leimbach et al., 2015). Anhand der genomischen Sequenz des Mastitis-assoziierten *E. coli* (MAEC) Stamms 1303 wurde das Gencluster zur Biosynthese seines O70 LPS O-Antigens aufgeklärt. Wir analysierten die phylogenetische Abstammung unserer Stammsammlung plus elf bovin-assoziierter *E. coli* Referenzstämme, aber konnten weder MAEC noch Kommensale bestimmten Phylogruppen innerhalb eines Core-Genom Stammbaums aus Referenz-*E. coli* eindeutig zuordnen. Eine ausführliche Gengehalt-Analyse konnte keine funktionelle Konvergenz innerhalb von Kommensalen oder MAEC identifizieren. Stattdessen besitzen beide nur sehr wenige Genfamilien, die bevorzugt in einer der beiden Pathotypen vorkommen. Weder eine Gengehalt- noch eine `ecoli_VF_collection`-Analyse konnte zeigen, dass eine signifikante Assoziation von bestimmten Virulenzfaktoren mit MAEC, im Vergleich zu bovinen fäkalen Kommensalen, besteht. Damit wurde die MPEC Hypothese widerlegt. Auch das genetische Repertoire von Rinder-assoziierten *E. coli* wird durch die phylogenetische Abstammung bestimmt. Dies ist überwiegend auch bei großen Virulenz-assoziierten Genclustern der Fall, die bisher mit Mastitis in Verbindung gebracht wurden. Dementsprechend sind MAEC fakultative und opportunistische Pathogene, die ihren Ursprung als Kommensale in der bovinen gastrointestinalen Mikrobiota haben (Leimbach et al., 2017).

Obwohl traditionelle *E. coli* Pathotypen in der Diagnostik und Behandlung einen Zweck erfüllen, ist es offensichtlich, dass das derzeitige Typisierungs-System die genomische Plastizität von *E. coli* zu sehr vereinfacht. Die Gesamtgenom-Sequenzierung (WGS) deckte viele Nuancen pathogener *E. coli* auf, einschließlich entstehender hybrider oder heteropathogener Pathotypen. Diagnostische und medizinische Mikrobiologie müssen einen Schritt in Richtung Zukunft gehen und HTS-Technologien anwenden, um Patientenversorgung und Infektionskontrolle effizienter zu unterstützen.

PUBLICATIONS

This is a list of publications which I co-authored in the course of the PhD thesis. Included are also a publicly available dataset of *E. coli* VFs and a collection of bioinformatical scripts for bacterial genomics. First author publications are indicated by a ★ symbol in the margin.

Original research publications and a review article included in this PhD thesis:

1. Duda KA, Lindner B, Brade H, Leimbach A, Brzuszkiewicz E, Dobrindt U, Holst O. 2011. The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type. *Microbiology* 157:1750–1760.
   DOI: 10.1099/mic.0.046912-0

2. Brzuszkiewicz E*, Thürmer A*, Schuldes J*, Leimbach A*, Liesegang H*, Meyer F-D, Boelter J, Petersen H, Gottschalk G, Daniel R. 2011. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol.* 193:883–891. ★
   DOI: 10.1007/s00203-011-0725-6

3. Leimbach A, Hacker J, Dobrindt U. 2013. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. *Curr. Top. Microbiol. Immunol.* 358:3–32. ★
   DOI: 10.1007/82_2012_303
   This is a review article.

4. Zude I*, Leimbach A*, Dobrindt U. 2014. Prevalence of autotransporters in *Escherichia coli*: what is the impact of phylogeny and pathotype? *Int. J. Med. Microbiol.* 304:243–256. ★
   DOI: 10.1016/j.ijmm.2013.10.006

5. Leimbach A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R, Dobrindt U. 2015. Complete genome sequences of *Escherichia coli* strains 1303 and ECC-1470 isolated from bovine mastitis. *Genome Announc.* 3:e00182-15. ★
   DOI: 10.1128/genomeA.00182-15

6. Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, Zerbe H, Daniel R, Dobrindt U. 2016. Whole-genome draft ★

* These authors contributed equally.

sequences of six commensal fecal and six mastitis-associated *Escherichia coli* strains of bovine origin. *Genome Announc.* 4:e00753-16.
DOI: 10.1128/genomeA.00753-16

★ 7. Leimbach A, Poehlein A, Vollmers J, Görlich D, Daniel R, Dobrindt U. 2017. No evidence for a bovine mastitis *Escherichia coli* pathotype. *BMC Genomics* 18:359.
DOI: 10.1186/s12864-017-3739-x

Publicly available dataset and bioinformatical scripts included in this PhD thesis:

★ 8. Leimbach A. 2016. `ecoli_VF_collection`: v0.1. *Zenodo*.
DOI: 10.5281/zenodo.56686

★ 9. Leimbach A. 2016. `bac-genomics-scripts`: Bovine *E. coli* mastitis comparative genomics edition. *Zenodo*.
DOI: 10.5281/zenodo.215824

Original research publications outside the scope of this PhD thesis:

10. Oehler D, Poehlein A, Leimbach A, Müller N, Daniel R, Gottschalk G, Schink B. 2012. Genome-guided analysis of physiological and morphological traits of the fermentative acetate oxidizer *Thermacetogenium phaeum*. *BMC Genomics* 13:723.
DOI: 10.1186/1471-2164-13-723

11. Eidam C, Poehlein A, Leimbach A, Michael GB, Kadlec K, Liesegang H, Daniel R, Sweeney MT, Murray RW, Watts JL, Schwarz S. 2015. Analysis and comparative genomics of ICE*Mh1*, a novel integrative and conjugative element (ICE) of *Mannheimia haemolytica*. *J. Antimicrob. Chemother.* 70:93–97.
DOI: 10.1093/jac/dku361

12. Djukic M, Poehlein A, Strauß J, Tann F, Leimbach A, Hoppert M, Daniel R. 2015. High quality draft genome of *Lactobacillus kunkeei* EFB6, isolated from a German European foulbrood outbreak of honeybees. *Stand. Genomic Sci.* 10:16.
DOI: 10.1186/1944-3277-10-16

13. Hertel R, Rodríguez DP, Hollensteiner J, Dietrich S, Leimbach A, Hoppert M, Liesegang H, Volland S. 2015. Genome-based identification of active prophage regions by next generation sequencing in *Bacillus licheniformis* DSM13. *PLoS One* 10:e0120759.
DOI: 10.1371/journal.pone.0120759

14. Poehlein A*, Riegel K*, König SM, Leimbach A, Daniel R, Dürre P. 2015. Genome sequence of *Clostridium sporogenes* DSM 795^T,

* These authors contributed equally.

an amino acid-degrading, nontoxic surrogate of neurotoxin-producing *Clostridium botulinum*. *Stand. Genomic Sci.* 10:40.
DOI: 10.1186/s40793-015-0016-y

15. Ullrich SR*, Poehlein A*, Voget S, Hoppert M, Daniel R, Leimbach A, Tischler JS, Schlömann M, Mühling M. 2015. Permanent draft genome sequence of *Acidiphilium* sp. JA12-A1. *Stand. Genomic Sci.* 10:56.
DOI: 10.1186/s40793-015-0040-y

16. Wildeman P, Brüggemann H, Scholz CF, Leimbach A, Söderquist B. 2016. *Propionibacterium avidum* as an etiological agent of prosthetic hip joint infection. *PLoS One* 11:e0158164.
DOI: 10.1371/journal.pone.0158164

It was the best of times, it was the worst of times,
it was the age of wisdom, it was the age of foolishness,
it was the epoch of belief, it was the epoch of incredulity,
it was the season of Light, it was the season of Darkness,
it was the spring of hope, it was the winter of despair . . .

— A Tale of Two Cities (1859), Charles Dickens

ACKNOWLEDGMENTS

I owe many thanks to colleagues, advisors, open scientists, friends, and family:

Ulrich Dobrindt for the opportunity to undertake my doctoral studies with his advise, feedback, inspiration, and guidance.

Rolf Daniel for accepting me into his laboratory and taking care of me as one of his own. Knut Ohlsen for officiating as a member in my thesis committee.

John Vollmers, Marvin Djukic, Anja Poehlein, Sandra Wiegand, Florian Lenk, Katrin Lorenzen, Dominik Schneider, and Sascha Dietrich for being awesome colleagues. Specifically Anja Poehlein for additional supervision. Jörg Schuldes and Andrea Thürmer for providing me a place to stay in between one of my many needed relocations. Finally, all past and current members of the G_2L.

Roswitha Schiller, Ingmar Zude, and Christian-Daniel Köhler for being great companions from afar.

Many open scientists sharing their experience and wisdom online, providing amazing open source software, and pushing science into new directions. In particular, thanks goes to Konrad Förstner, Kathryn Holt, Titus Brown, Mick Watson, Torsten Seemann, Nick Loman, David Studholme, and countless more. Also a big thank you to Alan McNally for hosting me – it was the best week of my thesis. Last but not least, Daniel Gerlach for advise on LaTeX quirks.

The podcasts Methodisch Inkorrekt and Open Science Radio for the entertainment during long evenings.

My family for everything!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AAF	aggregative adherence fimbriae
AIEC	adherent invasive *E. coli*
AT	autotransporter
BAM	β-barrel assembly machinery
BGI	Beijing Genomics Institute
CDS	coding sequence
COG	Clusters of Orthologous Groups
CU	chaperone-usher
DAEC	diffusely adherent *E. coli*
DOI	Digital Object Identifier
EAEC	enteroaggregative *E. coli*
EAHEC	entero-aggregative-haemorrhagic *E. coli*
EHEC	enterohaemorrhagic *E. coli*
EIEC	enteroinvasive *E. coli*
EPEC	enteropathogenic *E. coli*

ESBL	extended-spectrum β-lactamase
ETEC	enterotoxigenic *E. coli*
ETT2	*E. coli* type III secretion system 2
ExPEC	extraintestinal pathogenic *E. coli*
Fec	ferric iron(III)-dicitrate uptake system
FF	fitness factor
Flag-2	*E. coli* peritrichous flagella 2 gene cluster
G2L	Göttingen Genomics Laboratory
GI	genomic island
HGT	horizontal gene transfer
HPA	Health Protection Agency
HTS	high-throughput sequencing
HUS	haemolytic uraemic syndrome
IL	interleukin
IM	inner membrane
IMI	intramammary infection
INSDC	International Nucleotide Sequence Database Collaboration
IPEC	intestinal pathogenic *E. coli*
LEE	locus of enterocyte effacement
LPS	lipopolysaccharide
MAEC	mastitis-associated *E. coli*
MGE	mobile genetic element
MLST	multi-locus sequence typing
MNEC	newborn meningitis-associated *E. coli*
MPEC	mammary pathogenic *E. coli*
NCBI	National Center for Biotechnology Information
NET	neutrophil extracellular trap
NGS	next-generation sequencing
OG	orthologous group

OM	outer membrane
OMP	outer membrane protein
ORF	open reading frame
PAMP	pathogen-associated molecular pattern
PCR	polymerase chain reaction
PGM	Personal Genome Machine
PMN	polymorphonuclear neutrophil
RKI	Robert Koch Institute
ROD	region of difference
rRNA	ribosomal RNA
SCC	somatic cell count
SNP	single nucleotide polymorphism
SP	signal peptide
SPATE	serine protease autotransporters of *Enterobacteriaceae*
SRA	Sequence Read Archive
ST	sequence type
STEC	Shiga toxin-producing *E. coli*
Stx	Shiga toxin
T2SS	type II secretion system
T3SS	type III secretion system
T4SS	type IV secretion system
T5SS	type V secretion system
T6SS	type VI secretion system
TLR	Toll-like receptor
TNF	Tumor necrosis factor
UKE	University Medical Center Hamburg-Eppendorf
UKM	Münster University Hospital
UPEC	uropathogenic *E. coli*
UTI	urinary tract infection

VF	virulence factor
VFDB	virulence factor database
WGS	whole-genome (shotgun) sequencing
WHO	World Health Organization

Part I

GENERAL INTRODUCTION

1 INTRODUCTION

The genome of an organism contains its entire hereditary information. The corresponding DNA encodes for all RNA and protein molecules of a cell, needed for maintaining its functionality. From these genes it is possible to infer the capabilities of cells and organisms, like metabolic potential, motility, and virulence. Comparing the genomes of organisms can be used to detect differences in the nucleotides and calculate phylogenetic relationships. Bacterial genomes were the first DNA sequences of cellular life forms to be fully decoded and this information used to answer biological questions.

1.1 BACTERIAL GENOMICS

whole-genome shotgun sequencing

Analyzing the genomes of bacteria has come a long way: from the beginnings of examining short stretches of DNA to sequencing complete bacterial genomes. In 1995 the first bacterial genome was fully sequenced[1]. Fleischmann et al. (1995) applied a technique called whole-genome (shotgun) sequencing (WGS) to randomly sequence small pieces of the genomic nucleotides *en masse*, instead of "walking" along the genome in a known region. These DNA snippets were later assembled computationally into larger contiguous fragments (contigs) (Figure 1 on the following page). Because it is difficult to assemble repeat regions, genomes remain in a fragmented/draft form unless other molecular biology techniques are applied to subsequently stitch the contigs together into the closed genome – a process called "gap closure" (Section 1.1.3 on page 11) (Koren and Phillippy, 2015; Nagarajan et al., 2010; Phillippy, 2017). WGS was a revolution in the genomics/sequencing field.

Single complete genomes of bacterial species already had huge impact on bacteriology, especially on molecular biology applications. But not until several bacterial genomes were available was it possible to conduct *comparative bioinformatical analyses*. These analyses showed, that bacteria have an unexpected high genomic plasticity, not only inter- but also intra-species (Bentley and Parkhill, 2015; Loman and Pallen, 2015). Comparative genomics also revolutionized bacterial taxonomy[2] and as a consequence questioned the bacterial species concept. A species is traditionally defined as a coherent, discrete, and individual group. A large amount of strain-to-strain diversity is introduced by

[1] Non-pathogenic *Haemophilus influenzae* strain Rd was sequenced with a relatively small genome size of 1,830,137 bp (Fleischmann et al., 1995).

[2] Bacterial taxonomy classically relies on phenotypic characteristics for classification (e. g. metabolism and Gram stain).

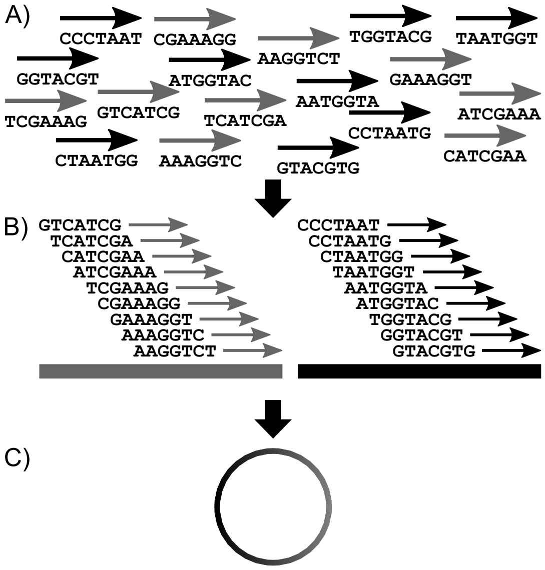


Figure 1: Assembly of overlapping shotgun sequencing reads (A) into contigs (B) and the contigs into a complete circular genome sequence (C). Figure created with `Inkscape` (v0.91).

horizontal gene transfer (HGT) within many bacterial species, but also between bacterial species and even between different domains of life (Dobrindt et al., 2004; Land et al., 2015; Loman and Pallen, 2015). Thus, the traditional bifurcating phylogenetic tree of the three domains of life is pervaded by HGT and homologous recombination, and bacterial species rather represent an interconnected network structure (Figure 2) (Doolittle and Zhaxybayeva, 2009; Martin and Embley, 2004). Nevertheless, a uniform species nomenclature serves as a useful tool for clinical microbiology and other indicator species (Chan et al., 2012a).

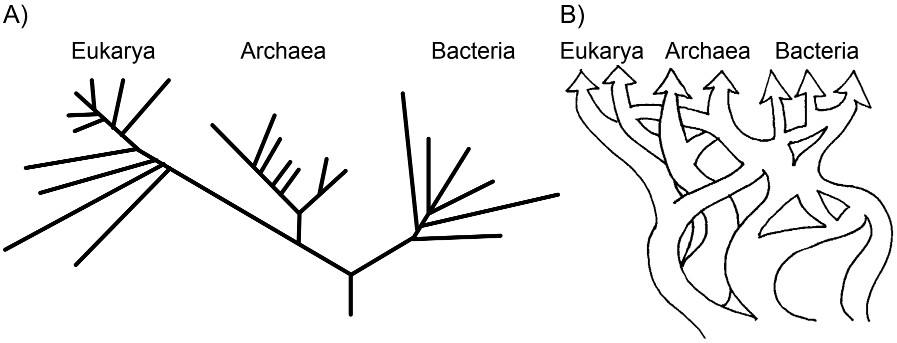


Figure 2: (A) Three-domain phylogenetic tree based on ribosomal RNA (rRNA) gene sequences. (B) The three-domain tree with examples of pervasive HGT between taxa. Adapted from Martin and Embley (2004) and Doolittle and Zhaxybayeva (2009) using `Inkscape`.

Bacteria are the evolutionary most diverse domain of life and much is still to be discovered as was recently exemplary shown by sequencing bacterial genomes from unexamined environmental samples with uncultivable[3] bacterial genera (Hanage, 2016; Hug et al., 2016).

With the turn of the century several genomes from individual bacterial species became available. The study of intra-species diversity birthed the field of bacterial population genomics. It became clear that single bacterial genomes are inadequate to describe bacterial species, because of the exceptional genomic diversity (Medini et al., 2008). Sequencing the genomes of bacterial populations allows the investigation of population structures and, in the case of host-associated bacteria like pathogens, within-host evolution and transmission history. This expands our understanding of the evolution and virulence of pathogens (Wilson, 2012). New concepts emerged like the *core genome* defined by Lan and Reeves (2000), as comprising those genes present in almost all individuals of a species. Genes in this category function mostly in housekeeping, like replication, transcription, translation, and essential metabolic pathways. The *flexible genome* includes genes with a variable presence/absence in individual isolates of a bacterial species plus genes unique to each strain (singletons). These genes account for the phenotypic diversity within bacterial populations and permit niche adaptation with specific fitness traits, like pathogenicity and antimicrobial resistances (Halachev et al., 2011; Medini et al., 2008). Flexible genome genes are mostly contained on mobile genetic elements (MGEs) that enable HGT, like plasmids, phages, and genomic islands (GIs) (Dobrindt et al., 2004).

core genome

flexible genome, alias accessory/dispensable/variable genome

Tettelin et al. (2005) coined the term *pan-genome*, which is the total gene repertoire of a bacterial species, including both the core and the flexible genome (Bentley and Parkhill, 2015; Medini et al., 2005). Bacterial species with an *open* pan-genome have a dynamic (basically infinite) genomic content, i. e. the species' gene repertoire increases every time a new genome is sequenced. An open pan-genome is characteristic for bacterial species that live in very different habitats, and thus require high adaptability, like *Escherichia coli* (Figure 6 on page 14). On the contrary, *closed* pan-genomes do not change by the addition of new genomes of a species. Such species have a static genomic content and it is possible to determine the full gene repertoire by sequencing enough genomes (e. g. the niche-restricted *Bacillus anthracis*) (Halachev et al., 2011; McInerney et al., 2017; Rouli et al., 2015; Tettelin et al., 2008).

pan-genome

1.1.1 *Current sequencing technologies (in microbiology)*

The genomic revolution is tightly coupled with developments in sequencing technologies and bioinformatical algorithms/tools. The sequencing technique that is used the longest was developed by Frederick

[3] Metagenomics (direct sequencing of environmental DNA from all living microorganisms isolated in a specific habitat) and single-cell genomics with new bioinformatical techniques are suitable for this purpose (Section 7.3 on page 222).

"first-generation" sequencing technique

Sanger in the 1970s, the dideoxy chain terminator technique (Sanger et al., 1977). This "first-generation" sequencing strategy involved amplified DNA templates. Single molecules can easily be amplified by polymerase chain reaction (PCR). For bacterial WGS (where the sequence is not known) the onerous and expensive approach has to be taken by introducing recombinant DNA fragments into bacterial clones (plasmids or bacterial artificial chromosomes, BACs) for subsequent sequencing. Because of several improvements, especially in automation, Sanger sequencing in the form of capillary sequencers is still in use today and will continue to be so for the foreseeable future (Land et al., 2015; Loman and Pallen, 2015).

1.1.1.1 *High-throughput sequencing*

also known as next-generation sequencing (NGS)

The rise of *high-throughput sequencing (HTS)* revolutionized sequencing of bacterial genomes as well as more complicated genomes from other organisms. The highly superior throughput and time improvements provided by these techniques (Figure 3 on the facing page), now made it possible to complete bacterial sequencing projects, that used to take years and hundreds of thousands of dollars, in merely hours or days and for less than 100$ (Loman and Pallen, 2015). Furthermore, the development of smaller HTS machines with a lower initial investment moved bacterial genome sequencing from large genome centers into the hands of individual researchers (McPherson, 2009). Thus, HTS quickly replaced first-generation Sanger sequencing in genomics, although it is still in use for low-throughput amplicon sequencing and gap closure of draft genomes.

"SECOND-GENERATION" SEQUENCING TECHNIQUES

Second-generation sequencing techniques replaced the cloning strategy for Sanger genomic sequencing with clonal parallel amplification of fragmented DNA *in vitro* (via PCR or cluster amplification) and advanced nucleotide detection methods. The result was a massive increase in throughput, rendering the sequencing of several bacterial genomes in a few hours possible (Loman and Pallen, 2015; Loman et al., 2012a; Medini et al., 2008). The first commercially HTS technology became available in 2005 with the 454 Genome Sequencer GS20, which employed emulsion PCR with subsequent *pyrosequencing* (Margulies et al., 2005). The pyrosequencing method detects emitted light powered by the release of pyrophosphates during DNA elongation of each base. The 454 company was later bought by Roche, and the machine was improved upon several times, but sale of hardware and chemistry was discontinued in 2016. A similar platform in terms of PCR amplification procedure and read length/error characteristics was developed by Ion Torrent and released 2011 (PGM, Proton . . .). The Ion Torrent machines act like a *pH meter* by detecting the release of H^+ during

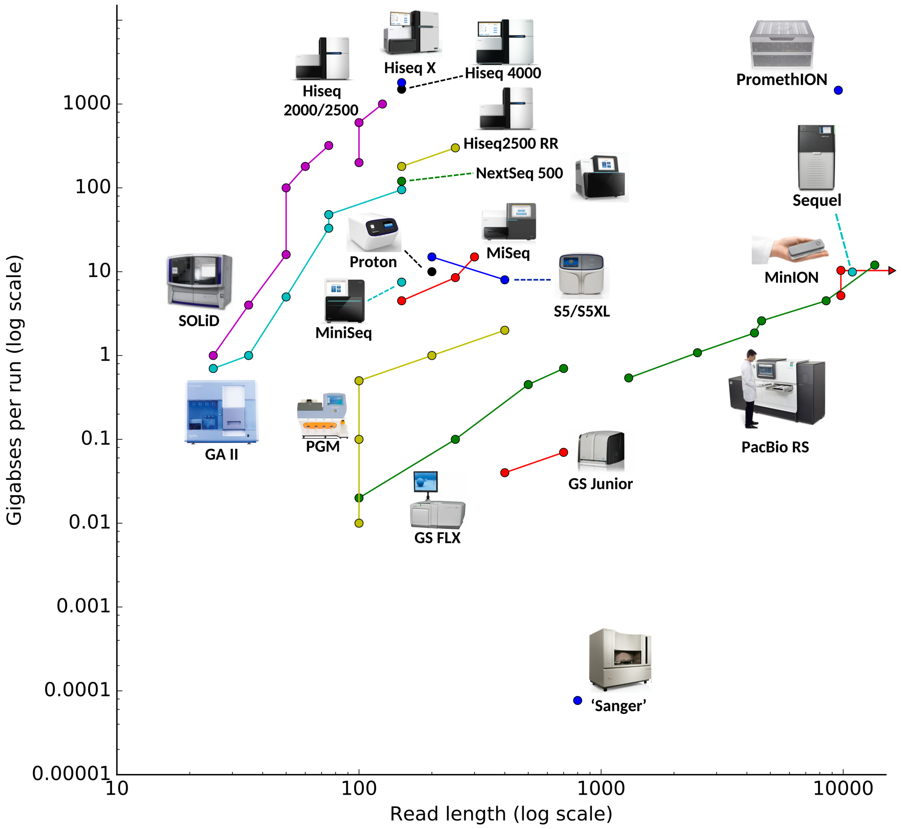


Figure 3: Read lengths and throughputs of HTS machines: Sanger = ABI 3730xl, SOLiD (© ABI); GS Junior, GS FLX (454, © Roche); MiniSeq, GA II, MiSeq, NextSeq, HiSeq (© Illumina); PGM, Proton, S5/S5XL (© Ion Torrent); PacBio RS, Sequel (© Pacific Biosciences); MinION, PromethION (© Oxford Nanopore Technologies). Each connected data point represents a further development in chemistry or hardware. Adapted from Nederbragt (2016) using Inkscape.

base incorporation. The currently most successful vendor[4] is Illumina (formerly Solexa) with its sequencing-by-synthesis method with either four- or two-colored *fluorophore-labelled reversible nucleotide terminators* (Goodwin et al., 2016; Loman et al., 2012a; Medini et al., 2008). Illumina released its first HTS machine 2006 and currently offers the widest array of different machines from small-scale benchtop platforms (like MiniSeq and MiSeq), to medium throughput (NextSeq), and to the highest throughput giants currently available (HiSeq in different models and the newest NovaSeq). There are other commercially available second-generation techniques (e. g. ABI's SOLiD), but are not as popular. All second-generation techniques have the ability in common to quickly generate large amounts of short sequencing reads (in the range of 50–700 bp) for which new bioinformatical *de novo* assembly or read alignment (mapping) algorithms and tools had to be developed (Section 1.1.2 on the next page) (Loman et al., 2012a; Pop and Salzberg, 2008).

short sequencing reads

"THIRD-GENERATION" SEQUENCING TECHNIQUES The newest addition to the HTS congregation are *single-molecule* sequencing methods without initial DNA amplification, also termed third-generation sequencing techniques. Two vendors currently compete in this market: Pacific Biosciences (PacBio) with its *single-molecule real-time* (SMRT) sequencing (first machine released 2011) and the most recent and probably most disruptive technology from Oxford Nanopore Technologies with its protein *nanopore* sequencing approach (first access in 2014). Pacific Biosciences (PacBio RS II and Sequel) uses a zero-mode waveguide flow cell to detect incorporated bases directly during DNA synthesis by an immobilized DNA polymerase in the bottom of a well. Oxford Nanopore Technologies (MinION[5], PromethION, and the upcoming SmidgION[6] and GridION X5[7]) developed a technology that threads a single DNA molecule through a nanopore[8] in an insulating lipid bilayer and detects changes in electrical current, which are characteristic for a particular DNA sequence in the pore. Because of its unique DNA sequence detection, it can stream the data during a sequencing run directly on a computer and analysis can be updated simultaneously. Nanopore sequencing is also able to directly sequence RNA without the traditional complementary DNA (cDNA) detour (Goodwin et al., 2016; Koren and Phillippy, 2015; Lannoy et al., 2017; Loman and Pallen,

[4] Mostly because of its to date unparalleled throughput and resulting low costs.

[5] About the size of a small "office stapler" (Risse et al., 2015) that plugs into (and is powered by) a Universal Serial Bus (USB) port. Because of their portability MinIONs have been used in the field and even on the International Space Station. Also, the small size makes the initial investment in hardware minimal in comparison to the other machines.

[6] Developed to be run plugged into smartphones.

[7] A system that can run and analyze five MinION flow cells independently.

[8] Currently, a variant of the *E. coli* curli fibre pore-forming protein CsgG serves as the *nanopore* (Barnhart and Chapman, 2006; Lannoy et al., 2017).

2015). The major advantage of these third-generation techniques is the superior read length: PacBio reaches read lengths in excess of 50 kb and on average 10 – 15 kb, while with MinION read lengths close to 1 Mbp have been achieved[9] and can now routinely reach ~150 kb with an average read length of 10 kb (Figure 3 on page 7) (Goodwin et al., 2016). Because of these long reads, the third generation sparked a rejuvenation of finishing bacterial genomes by simplifying the assembly problem, possibly even eliminating the concept of draft genomes (Chain et al., 2009; Koren and Phillippy, 2015; Koren et al., 2013; Land et al., 2015; Loman and Pallen, 2015; Loman et al., 2015). The biggest caveat, however, is a higher error rate in comparison to second-generation machines, because single DNA molecules without prior amplification are sequenced (Jain et al., 2017; Lannoy et al., 2017). Last but not least, both techniques can also detect methylation and other chemical base epigenetic modifications that e. g. control expression levels in bacteria and affect pathogen behavior (Jain et al., 2016; Loman and Pallen, 2015; Simpson et al., 2017; Stoiber et al., 2016).

read lengths up to 1 Mbp

For further information on HTS have a look at the detailed overview article of new sequencing techniques written by Goodwin et al. (2016). However, the quick pace of hardware and chemistry development in the field challenges the sluggishness of traditional scientific publishing. More importantly, the rapid advancement of the field increasingly strains the data analysis capabilities of microbiologists.

1.1.2 *High-throughput data deluge and resulting challenges*

HTS technologies and ever decreasing costs of cheap genomic data generation[10] also brought the problem of dealing with huge amounts of digital data to bacteriology (Pallen, 2016). This has resulted in a cost shift from sequencing to data management and analysis (Land et al., 2015). Thus, biology is currently undergoing the same "big data"-driven changes as several physics disciplines did for the last decade, albeit with more complex/heterogeneous systems. Although it is cheap to sequence bacterial genomes nowadays, massive additional costs for data (computing power and storage) and bioinformatical expertise are often underestimated (Loman et al., 2012a; Sboner et al., 2011).

BIG DATA

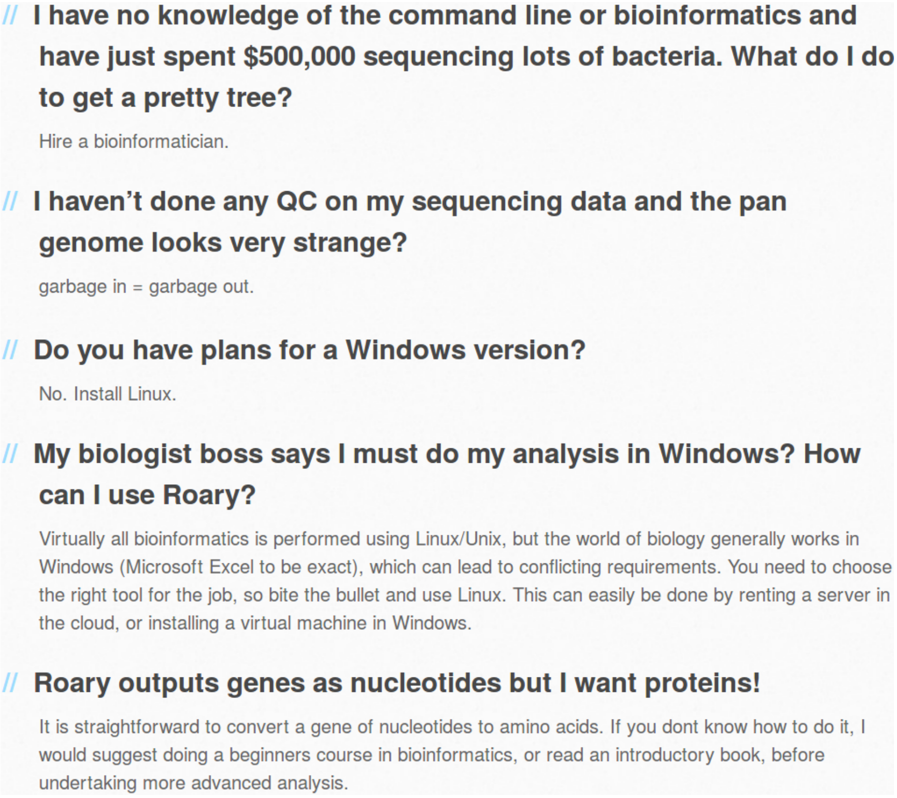
An amusing example of the current problems is shown in Figure 4 on the next page from the manual of the pan-genome software tool `ROARY`[11] developed by Andrew J. Page from the Sanger institute (Page

[9] The sequenced *E. coli* K-12 MG1655 genome was covered with just *seven reads*: http://lab.loman.net/2017/03/09/ultrareads-for-nanopore/

[10] Sequence data generation is undercutting Moore's law (the doubling of transistor fittings on integrated circuits every year) by a factor of more than 10^3. Thus, without advancements in computer sciences and algorithms the facilities to store, process, analyze, and maintain the data are not sustainable (Sboner et al., 2011). See https://www.genome.gov/sequencingcostsdata/.

[11] https://sanger-pathogens.github.io/Roary/

et al., 2015). It illustrates the current challenges with the data flood in bacterial genomics, the shortage of trained microbiologists, and exaggerated expectations.



// I have no knowledge of the command line or bioinformatics and have just spent $500,000 sequencing lots of bacteria. What do I do to get a pretty tree?

Hire a bioinformatician.

// I haven't done any QC on my sequencing data and the pan genome looks very strange?

garbage in = garbage out.

// Do you have plans for a Windows version?

No. Install Linux.

// My biologist boss says I must do my analysis in Windows? How can I use Roary?

Virtually all bioinformatics is performed using Linux/Unix, but the world of biology generally works in Windows (Microsoft Excel to be exact), which can lead to conflicting requirements. You need to choose the right tool for the job, so bite the bullet and use Linux. This can easily be done by renting a server in the cloud, or installing a virtual machine in Windows.

// Roary outputs genes as nucleotides but I want proteins!

It is straightforward to convert a gene of nucleotides to amino acids. If you dont know how to do it, I would suggest doing a beginners course in bioinformatics, or read an introductory book, before undertaking more advanced analysis.

Figure 4: An excerpt from the ROARY manual FAQ, illustrating the problems in analyzing the large amounts of data obtained by modern bacterial genomics (Page et al., 2015). Source: https://sanger-pathogens.github.io/Roary/ accessed on the 29th of May 2017.

open source development

Although several companies were able to launch commercial HTS technologies and corresponding workflows, this was not as successful for downstream data analysis pipelines. Because of the immense speed of HTS technology advancement commercial software suites have difficulties in standardization and keeping up. Algorithm and tool improvements are thus driven via *open source* development by dedicated and enthusiastic (microbial) bioinformaticians (Pallen, 2016).

Microbial bioinformatics/computational biology, like microbiology, is a maturing research field, which needs funding to develop tools, databases, and a suitable career structure[12] (Pallen, 2016). Especially training biology students, that can handle, analyze (particularly statis-

[12] See a series of blog posts by Mick Watson on the lonely/pet bioinformatician situation: http://www.opiniomics.org/a-guide-for-the-lonely-bioinformatician/, http://www.opiniomics.org/the-lonely-bioinformatician-revisited-clinical-labs/, http://www.opiniomics.org/how-to-recruit-a-good-bioinformatician/, and http://www.opiniomics.org/youre-not-allowed-bioinformatics-anymore/

tical inferences), and get biological insights out of the acquired data, has been one of the biggest challenges for several years and will be in the future (Pevzner and Shamir, 2009). Applied computational biology also challenges the current publishing system, as many essential tasks for modern bacterial genetics, like tool/database advancement and support, cannot be supported by additional high profile publications in contrast to the development of a new (and maybe obsolete) tool (Section 3.1 on page 65). HTS changed and is changing the way all microbiologists work, while genome sequencing will become a routine approach in research laboratories (Pallen, 2016). After all, as Jorgensen (2011) eloquently put it

> "we are all computational biologists now"

and we better be ready for it!

1.1.3 *Bacterial (whole-genome shotgun) sequencing*

Although, several hundred bacterial genomes can be sequenced in the matter of a single day on a single second-generation HTS machine, short sequencing reads are not suitable to close and finish[13] whole bacterial genomes without gaps. Repeats that are longer than the read length cannot be assembled unambiguously (the assembler starts to assemble a new contig), which resulted in a large increase of bacterial "draft" genomes (Figure 5 on the next page) (Chain et al., 2009; Pop and Salzberg, 2008; Treangen and Salzberg, 2011). If a finished genome was desired, the gaps in draft genomes were closed with individual PCR amplification and traditional Sanger sequencing (primer walking) under significant resource investments and teams of people. An alternative approach to span most genomic repeats is to sequence *paired reads*, which are reads from the ends of size selected-inserts separated by a known distance (Koren and Phillippy, 2015; Nagarajan et al., 2010; Phillippy, 2017). However, the increased read length of third-generation sequencing techniques confers the ability to close bacterial genomes without manual intervention by spanning large repeat stretches (alone or in combination with short read HTS). With up-to-date sequencing techniques of the second and third generation a bacterial draft genome can cost close to 1$ (Land et al., 2015; Loman and Pallen, 2015) and most finished genomes can be completed for under 1,000$ (Koren and Phillippy, 2015).

Draft genomes are sufficient for many analyses but naturally lack in the elucidation of large-scale structural analysis of genomes (like genome architecture, rearrangements, and synteny). Additionally, the accuracy of studies is enhanced with finished genomes, by eliminating

[13] A *finished* genome sequence has nearly no gaps, i. e. each replicon in a single contiguous sequence, and is of high quality (Chain et al., 2009; Koren and Phillippy, 2015; MacLean et al., 2009).

mapping artifacts, missed gene calls, and inaccurate repeat assembly (Koren and Phillippy, 2015; Koren et al., 2013).

Currently there are 8,890 finished and 90,023 draft prokaryotic (archaeal and bacterial) genomes stored in the Genbank repository[14] of the National Center for Biotechnology Information (NCBI) (Figure 5 on this page). The impact of HTS on the number of prokaryotic genomes sequenced is tremendous, especially apparent in the explosion of the number of draft genomes available. However, Figure 5 does not even take the huge amounts of unassembled HTS bacterial genome and metagenomic sequencing reads stored in NCBI's Sequence Read Archive (SRA) into account, which has grown even faster in the last decade and was 8,000 times bigger than the Genbank database in 2015.

previously known as the "Short Read Archive"

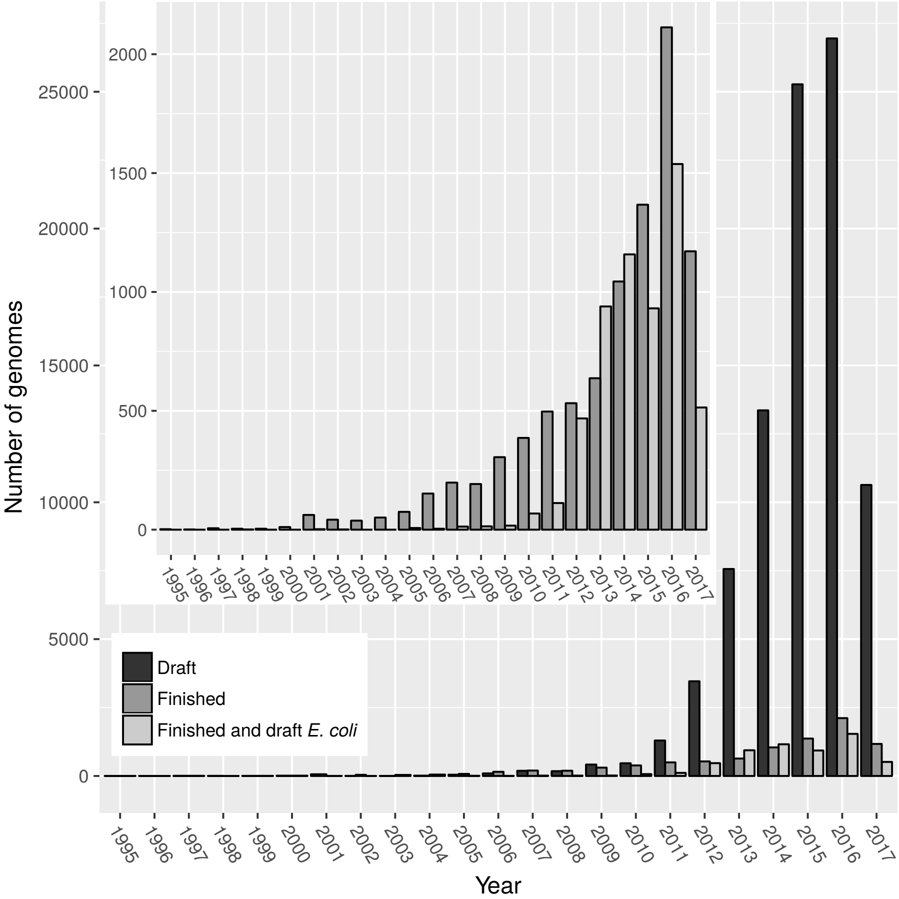


Figure 5: Number of prokaryotic (archaeal and bacterial) draft and finished genome sequences, and the sum of draft and finished *E. coli* genomes stored each year in NCBI's Genbank[14] database. The inset shows the same numbers zoomed in for the finished and *E. coli* genomes. Both bar charts were plotted with `R` (v3.4.0) (R Core Team, 2017) and package `ggplot2` (v2.2.1) (Wickham, 2009), and merged with `Inkscape`.

[14] Source: `ftp://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt` downloaded on the 25th of May 2017.

The sequencing revolution in bacterial genomics with its accompanying high resolution and discriminatory power has lead to novel insights (Bentley and Parkhill, 2015; Klemm and Dougan, 2016; McAdam et al., 2014) into

- bacterial evolution and population dynamics
- HGT mechanisms and dispersion
- within host genetic diversification
- expanding our knowledge in the emergence, adaptation, and transmission chains of pathogenic lineages

Close to 50% of available bacterial genomes are Proteobacteria (Land et al., 2015), with currently 350 finished and 5,485 draft *E. coli* genomes in Genbank[14] (Figure 5).

1.2 PHYLOGENETIC HISTORY, POPULATION GENETICS, AND GENOMIC EVOLUTION OF COMMENSAL AND PATHOGENIC *E. COLI*

Escherichia coli is not only the foremost molecular biology workhorse, but also an exemplary organism for bacterial genomics. *E. coli* was one of the earliest microbial species to be completely sequenced in 1997[15] (Blattner et al., 1997). The species includes extremely diverse isolates in regard to phenotype, that are mostly host-associated with several animals but also able to persist in abiotic conditions. Additionally, *E. coli* is famous for including both commensal and the alphabet soup of pathogenic strains, "pathotypes"[16]. Because of this diversity *E. coli* has an open pan-genome (Figure 6 on the following page) and only about 60% of each *E. coli* genome is considered to belong to its core genome (Kaas et al., 2012; Land et al., 2015). HGT mediated by MGEs plays a deciding role in the vast plasticity in *E. coli* genomes. Both the phenotypic diversity (especially as a human pathogen) and the high genome plasticity tempted to study the genome of the *E. coli* species in great detail.

E. coli *pathotypes*

1.2.1 E. coli *as an all-rounder: the thin line between commensalism and pathogenicity*

The following review article by Leimbach et al. (2013) describes the current knowledge on the phylogenetic history, population genetics, and evolution of extant *E. coli*. Emphasis is put on the relationship

[15] Laboratory *E. coli* K-12 strain MG1655 was sequenced.

[16] E. g. enterohaemorrhagic *E. coli* (EHEC), enteroaggregative *E. coli* (EAEC), extraintestinal pathogenic *E. coli* (ExPEC) . . . , see our Leimbach et al. (2013) review article in Section 1.2.1.2 on page 15

[17] https://www.genoscope.cns.fr/agc/microscope/home/index.php

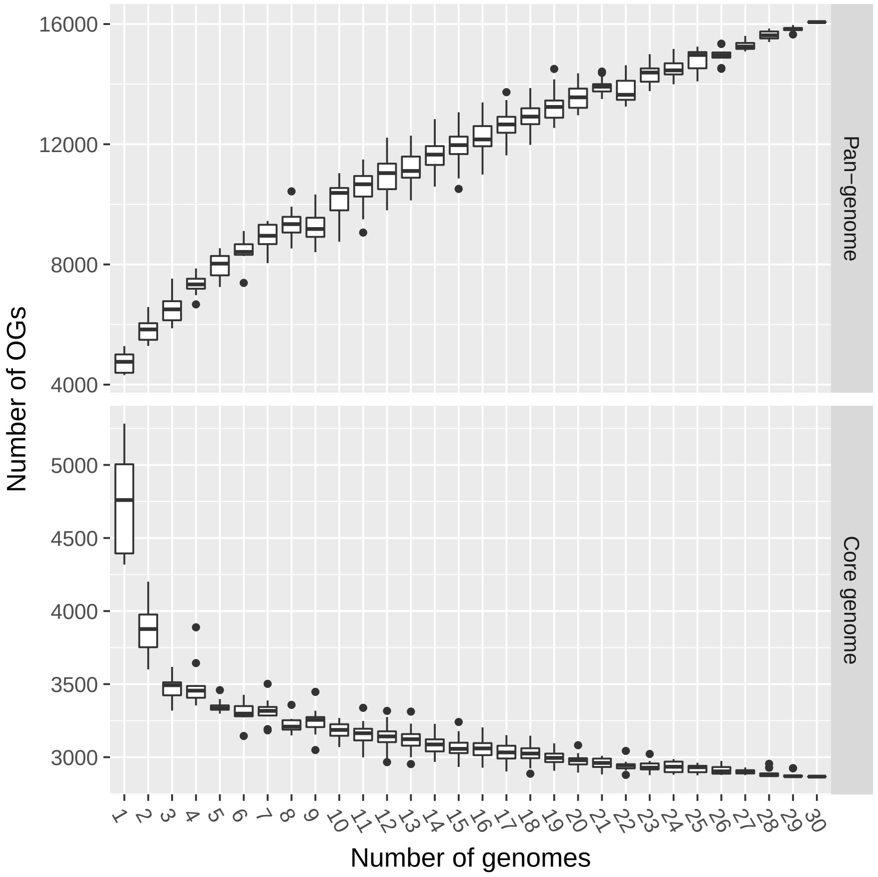


Figure 6: *E. coli* core and pan-genome boxplots with 30 selected *E. coli* genomes: *E. coli* EAEC 042 (phylogroup D1), EAEC 101-1 (A), UPEC 536 (B2), EAEC 55989 (B1), commensal B REL606 (A), ETEC B7A (B1), UPEC CFT073 (B2), commensal DH1 (A), ETEC E24377A (B1), commensal ED1a (B2), ETEC H10407 (A), commensal HS (A), commensal IAI1 (B1), UPEC IAI39 (D2), commensal K-12 (A), AIEC LF82 (B2), UPEC NA114 (B2), commensal Nissle 1917 (B2), EHEC O103:H2 12009 (B1), EHEC O157:H7 EDL933 (E), EHEC O157:H7 Sakai (E), EHEC O26:H11 11368 (B1), MNEC S88 (B2), commensal SE11 (B1), commensal SE15 (B2), environmental isolate SMS-3-5 (D2), UPEC UMN026 (D1), ETEC UMNF18 (A), UPEC UTI89 (B2), and commensal W3110 (A). The number of orthologous groups (OGs) in common to all respective genomes (core genome) and the total number of OGs present in the genomes (pan-genome) are shown. The boxplots represent the distribution of OGs with about 1,000 random different input order combinations of the genomes. With these 30 genomes the core genome had 2,876 OGs and the pan-genome a total of 16,065 OGs. OGs were calculated with 80% amino acid identity and coverage cutoffs to cluster the gene protein sequences of the included *E. coli* genomes with the MicroScope platform[17] (Vallenet et al., 2017). Figure plotted with `R` (v3.4.0) (R Core Team, 2017) and package `ggplot2` (v2.2.1) (Wickham, 2009).

of commensal and pathogenic isolates, specifically the difficulties in differentiating commensal *E. coli* from ExPEC. The original publisher PDF pages are indicated by enclosing frames.

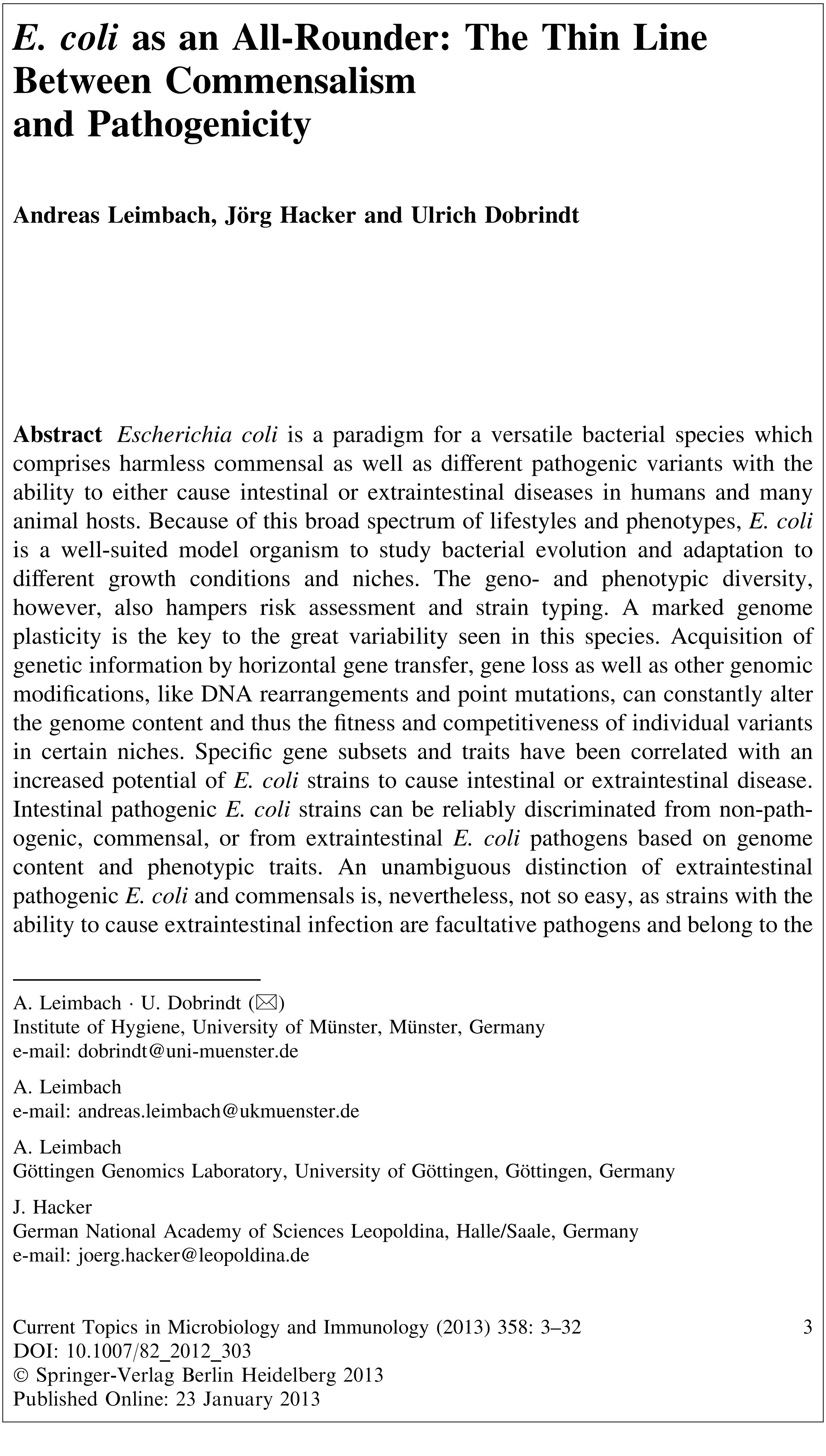
> Leimbach A, Hacker J, Dobrindt U. 2013. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. *Curr. Top. Microbiol. Immunol.* 358:3–32.
> DOI: 10.1007/82_2012_303

1.2.1.1 *Contributions*

I contributed the introduction, all figures, and the chapters on *E. coli* population genetics, genome plasticity, and the outlook to this review article. Ulrich Dobrindt wrote the chapter on genomic differences between commensal *E. coli* and ExPEC. Detailed individual author contributions for each part of the review article and each figure/table can be found in Table 8 and Table 9 on page 229, respectively.

1.2.1.2 *Main article*

Reprinted from Leimbach et al. (2013) with permission of Springer. The article can be found on pages 16–45 or at:
`https://link.springer.com/chapter/10.1007%2F82_2012_303`



E. coli as an All-Rounder: The Thin Line Between Commensalism and Pathogenicity

Andreas Leimbach, Jörg Hacker and Ulrich Dobrindt

Abstract *Escherichia coli* is a paradigm for a versatile bacterial species which comprises harmless commensal as well as different pathogenic variants with the ability to either cause intestinal or extraintestinal diseases in humans and many animal hosts. Because of this broad spectrum of lifestyles and phenotypes, *E. coli* is a well-suited model organism to study bacterial evolution and adaptation to different growth conditions and niches. The geno- and phenotypic diversity, however, also hampers risk assessment and strain typing. A marked genome plasticity is the key to the great variability seen in this species. Acquisition of genetic information by horizontal gene transfer, gene loss as well as other genomic modifications, like DNA rearrangements and point mutations, can constantly alter the genome content and thus the fitness and competitiveness of individual variants in certain niches. Specific gene subsets and traits have been correlated with an increased potential of *E. coli* strains to cause intestinal or extraintestinal disease. Intestinal pathogenic *E. coli* strains can be reliably discriminated from non-pathogenic, commensal, or from extraintestinal *E. coli* pathogens based on genome content and phenotypic traits. An unambiguous distinction of extraintestinal pathogenic *E. coli* and commensals is, nevertheless, not so easy, as strains with the ability to cause extraintestinal infection are facultative pathogens and belong to the

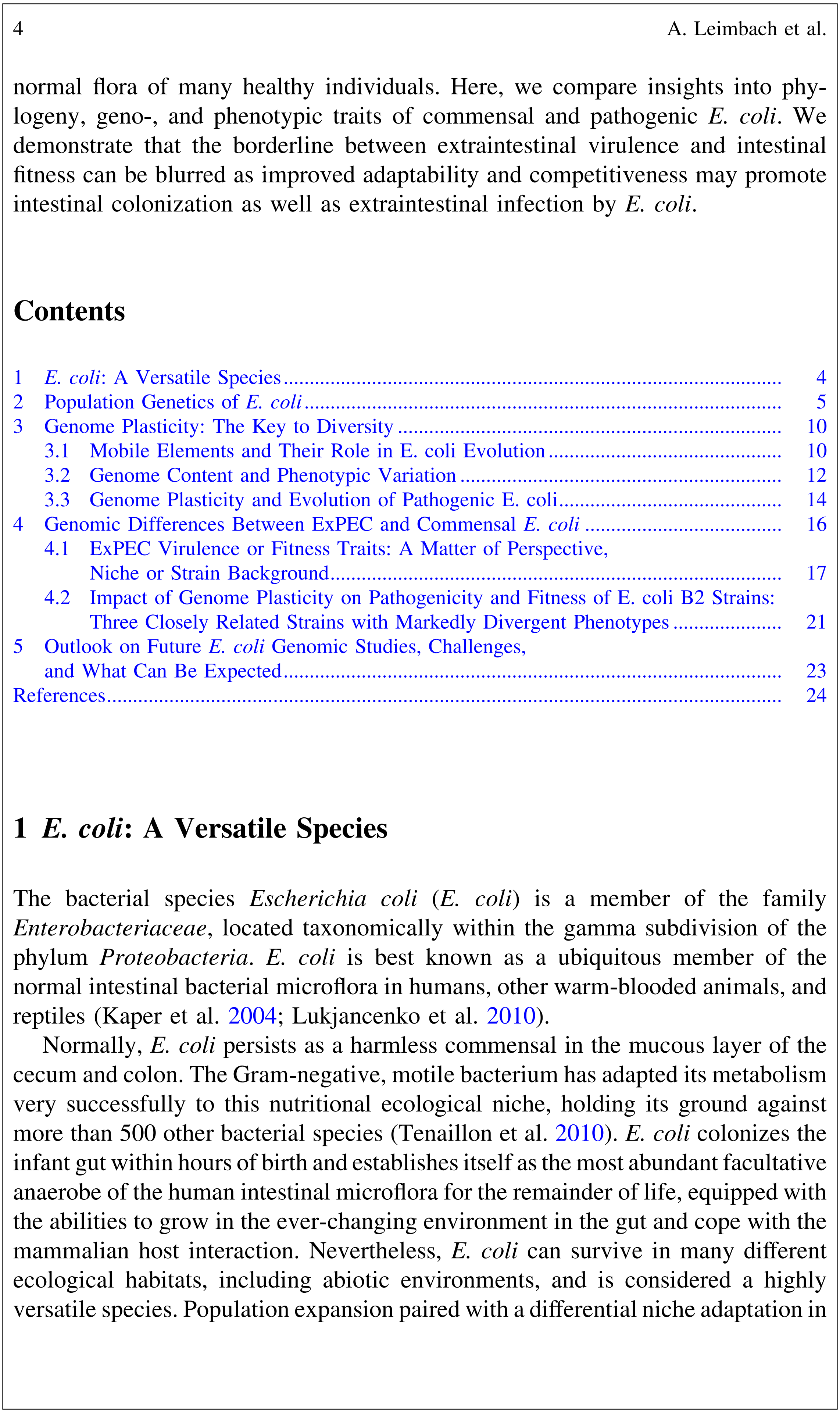
A. Leimbach · U. Dobrindt (✉)
Institute of Hygiene, University of Münster, Münster, Germany
e-mail: dobrindt@uni-muenster.de

A. Leimbach
e-mail: andreas.leimbach@ukmuenster.de

A. Leimbach
Göttingen Genomics Laboratory, University of Göttingen, Göttingen, Germany

J. Hacker
German National Academy of Sciences Leopoldina, Halle/Saale, Germany
e-mail: joerg.hacker@leopoldina.de

Current Topics in Microbiology and Immunology (2013) 358: 3–32 3
DOI: 10.1007/82_2012_303
© Springer-Verlag Berlin Heidelberg 2013
Published Online: 23 January 2013



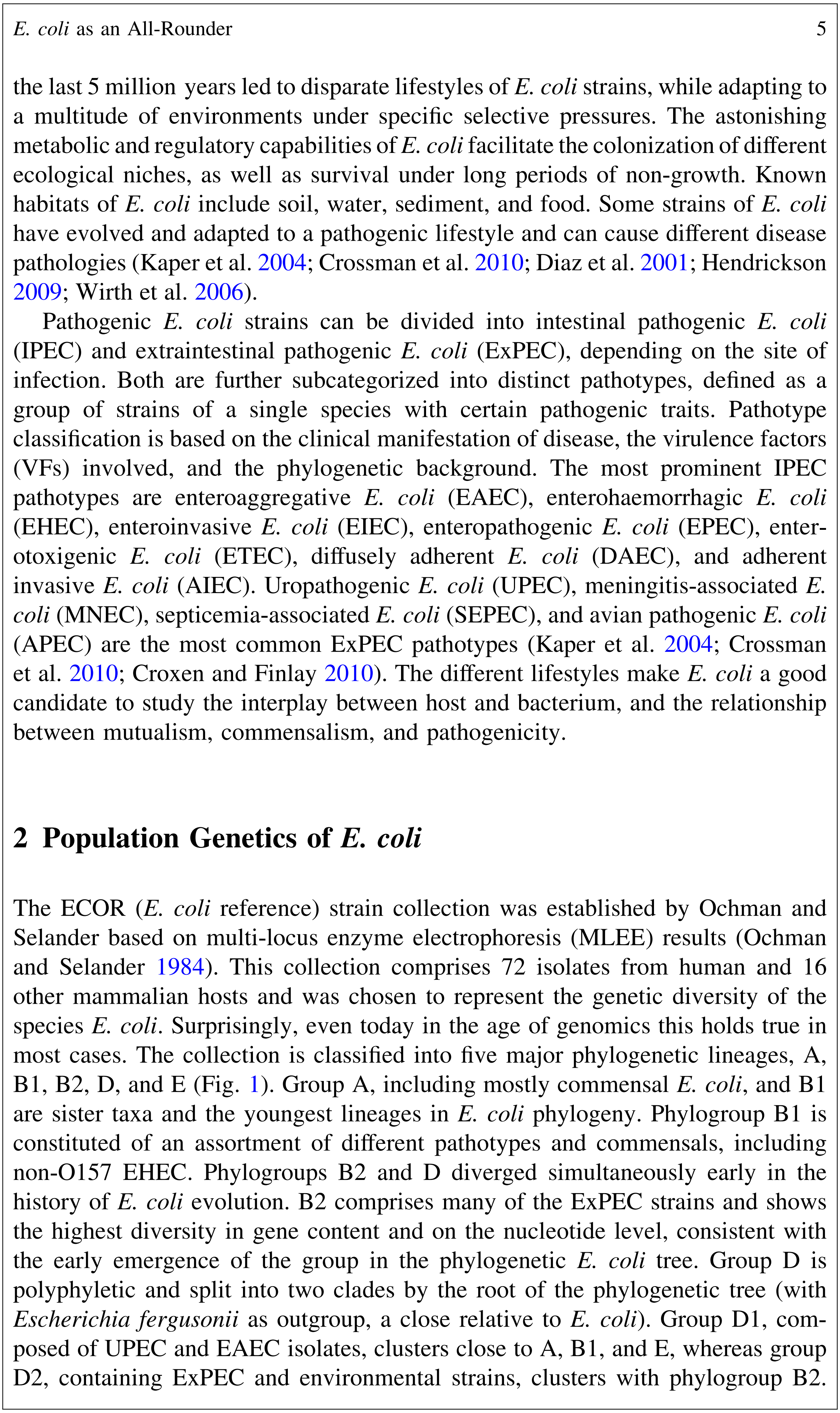
normal flora of many healthy individuals. Here, we compare insights into phylogeny, geno-, and phenotypic traits of commensal and pathogenic *E. coli*. We demonstrate that the borderline between extraintestinal virulence and intestinal fitness can be blurred as improved adaptability and competitiveness may promote intestinal colonization as well as extraintestinal infection by *E. coli*.

Contents

1 *E. coli*: A Versatile Species

The bacterial species *Escherichia coli* (*E. coli*) is a member of the family *Enterobacteriaceae*, located taxonomically within the gamma subdivision of the phylum *Proteobacteria*. *E. coli* is best known as a ubiquitous member of the normal intestinal bacterial microflora in humans, other warm-blooded animals, and reptiles (Kaper et al. 2004; Lukjancenko et al. 2010).

Normally, *E. coli* persists as a harmless commensal in the mucous layer of the cecum and colon. The Gram-negative, motile bacterium has adapted its metabolism very successfully to this nutritional ecological niche, holding its ground against more than 500 other bacterial species (Tenaillon et al. 2010). *E. coli* colonizes the infant gut within hours of birth and establishes itself as the most abundant facultative anaerobe of the human intestinal microflora for the remainder of life, equipped with the abilities to grow in the ever-changing environment in the gut and cope with the mammalian host interaction. Nevertheless, *E. coli* can survive in many different ecological habitats, including abiotic environments, and is considered a highly versatile species. Population expansion paired with a differential niche adaptation in

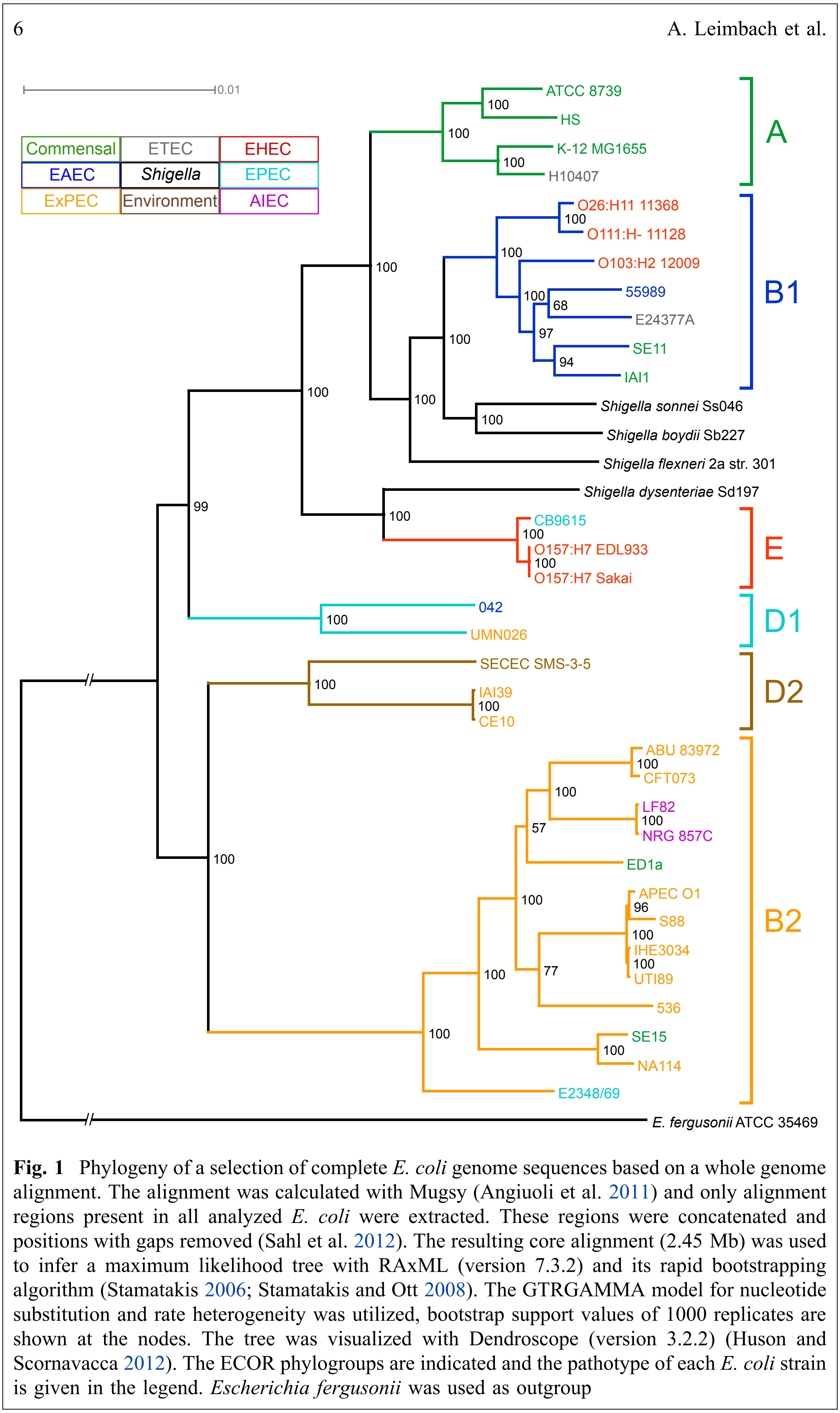


the last 5 million years led to disparate lifestyles of *E. coli* strains, while adapting to a multitude of environments under specific selective pressures. The astonishing metabolic and regulatory capabilities of *E. coli* facilitate the colonization of different ecological niches, as well as survival under long periods of non-growth. Known habitats of *E. coli* include soil, water, sediment, and food. Some strains of *E. coli* have evolved and adapted to a pathogenic lifestyle and can cause different disease pathologies (Kaper et al. 2004; Crossman et al. 2010; Diaz et al. 2001; Hendrickson 2009; Wirth et al. 2006).

Pathogenic *E. coli* strains can be divided into intestinal pathogenic *E. coli* (IPEC) and extraintestinal pathogenic *E. coli* (ExPEC), depending on the site of infection. Both are further subcategorized into distinct pathotypes, defined as a group of strains of a single species with certain pathogenic traits. Pathotype classification is based on the clinical manifestation of disease, the virulence factors (VFs) involved, and the phylogenetic background. The most prominent IPEC pathotypes are enteroaggregative *E. coli* (EAEC), enterohaemorrhagic *E. coli* (EHEC), enteroinvasive *E. coli* (EIEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), diffusely adherent *E. coli* (DAEC), and adherent invasive *E. coli* (AIEC). Uropathogenic *E. coli* (UPEC), meningitis-associated *E. coli* (MNEC), septicemia-associated *E. coli* (SEPEC), and avian pathogenic *E. coli* (APEC) are the most common ExPEC pathotypes (Kaper et al. 2004; Crossman et al. 2010; Croxen and Finlay 2010). The different lifestyles make *E. coli* a good candidate to study the interplay between host and bacterium, and the relationship between mutualism, commensalism, and pathogenicity.

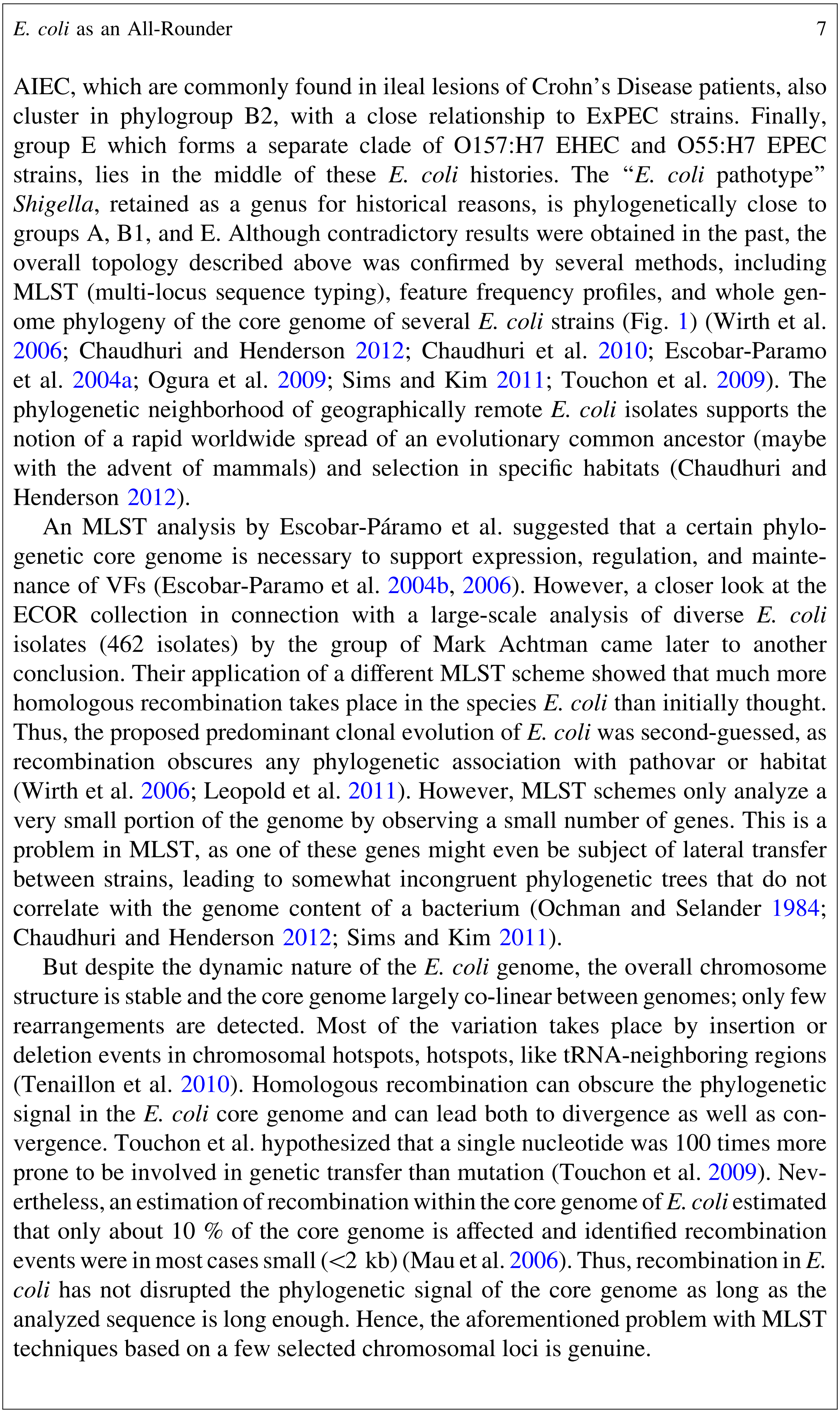
2 Population Genetics of *E. coli*

The ECOR (*E. coli* reference) strain collection was established by Ochman and Selander based on multi-locus enzyme electrophoresis (MLEE) results (Ochman and Selander 1984). This collection comprises 72 isolates from human and 16 other mammalian hosts and was chosen to represent the genetic diversity of the species *E. coli*. Surprisingly, even today in the age of genomics this holds true in most cases. The collection is classified into five major phylogenetic lineages, A, B1, B2, D, and E (Fig. 1). Group A, including mostly commensal *E. coli*, and B1 are sister taxa and the youngest lineages in *E. coli* phylogeny. Phylogroup B1 is constituted of an assortment of different pathotypes and commensals, including non-O157 EHEC. Phylogroups B2 and D diverged simultaneously early in the history of *E. coli* evolution. B2 comprises many of the ExPEC strains and shows the highest diversity in gene content and on the nucleotide level, consistent with the early emergence of the group in the phylogenetic *E. coli* tree. Group D is polyphyletic and split into two clades by the root of the phylogenetic tree (with *Escherichia fergusonii* as outgroup, a close relative to *E. coli*). Group D1, composed of UPEC and EAEC isolates, clusters close to A, B1, and E, whereas group D2, containing ExPEC and environmental strains, clusters with phylogroup B2.



6 A. Leimbach et al.

Fig. 1 Phylogeny of a selection of complete *E. coli* genome sequences based on a whole genome alignment. The alignment was calculated with Mugsy (Angiuoli et al. 2011) and only alignment regions present in all analyzed *E. coli* were extracted. These regions were concatenated and positions with gaps removed (Sahl et al. 2012). The resulting core alignment (2.45 Mb) was used to infer a maximum likelihood tree with RAxML (version 7.3.2) and its rapid bootstrapping algorithm (Stamatakis 2006; Stamatakis and Ott 2008). The GTRGAMMA model for nucleotide substitution and rate heterogeneity was utilized, bootstrap support values of 1000 replicates are shown at the nodes. The tree was visualized with Dendroscope (version 3.2.2) (Huson and Scornavacca 2012). The ECOR phylogroups are indicated and the pathotype of each *E. coli* strain is given in the legend. *Escherichia fergusonii* was used as outgroup



AIEC, which are commonly found in ileal lesions of Crohn's Disease patients, also cluster in phylogroup B2, with a close relationship to ExPEC strains. Finally, group E which forms a separate clade of O157:H7 EHEC and O55:H7 EPEC strains, lies in the middle of these *E. coli* histories. The "*E. coli* pathotype" *Shigella*, retained as a genus for historical reasons, is phylogenetically close to groups A, B1, and E. Although contradictory results were obtained in the past, the overall topology described above was confirmed by several methods, including MLST (multi-locus sequence typing), feature frequency profiles, and whole genome phylogeny of the core genome of several *E. coli* strains (Fig. 1) (Wirth et al. 2006; Chaudhuri and Henderson 2012; Chaudhuri et al. 2010; Escobar-Paramo et al. 2004a; Ogura et al. 2009; Sims and Kim 2011; Touchon et al. 2009). The phylogenetic neighborhood of geographically remote *E. coli* isolates supports the notion of a rapid worldwide spread of an evolutionary common ancestor (maybe with the advent of mammals) and selection in specific habitats (Chaudhuri and Henderson 2012).

An MLST analysis by Escobar-Páramo et al. suggested that a certain phylogenetic core genome is necessary to support expression, regulation, and maintenance of VFs (Escobar-Paramo et al. 2004b, 2006). However, a closer look at the ECOR collection in connection with a large-scale analysis of diverse *E. coli* isolates (462 isolates) by the group of Mark Achtman came later to another conclusion. Their application of a different MLST scheme showed that much more homologous recombination takes place in the species *E. coli* than initially thought. Thus, the proposed predominant clonal evolution of *E. coli* was second-guessed, as recombination obscures any phylogenetic association with pathovar or habitat (Wirth et al. 2006; Leopold et al. 2011). However, MLST schemes only analyze a very small portion of the genome by observing a small number of genes. This is a problem in MLST, as one of these genes might even be subject of lateral transfer between strains, leading to somewhat incongruent phylogenetic trees that do not correlate with the genome content of a bacterium (Ochman and Selander 1984; Chaudhuri and Henderson 2012; Sims and Kim 2011).

But despite the dynamic nature of the *E. coli* genome, the overall chromosome structure is stable and the core genome largely co-linear between genomes; only few rearrangements are detected. Most of the variation takes place by insertion or deletion events in chromosomal hotspots, hotspots, like tRNA-neighboring regions (Tenaillon et al. 2010). Homologous recombination can obscure the phylogenetic signal in the *E. coli* core genome and can lead both to divergence as well as convergence. Touchon et al. hypothesized that a single nucleotide was 100 times more prone to be involved in genetic transfer than mutation (Touchon et al. 2009). Nevertheless, an estimation of recombination within the core genome of *E. coli* estimated that only about 10 % of the core genome is affected and identified recombination events were in most cases small (<2 kb) (Mau et al. 2006). Thus, recombination in *E. coli* has not disrupted the phylogenetic signal of the core genome as long as the analyzed sequence is long enough. Hence, the aforementioned problem with MLST techniques based on a few selected chromosomal loci is genuine.

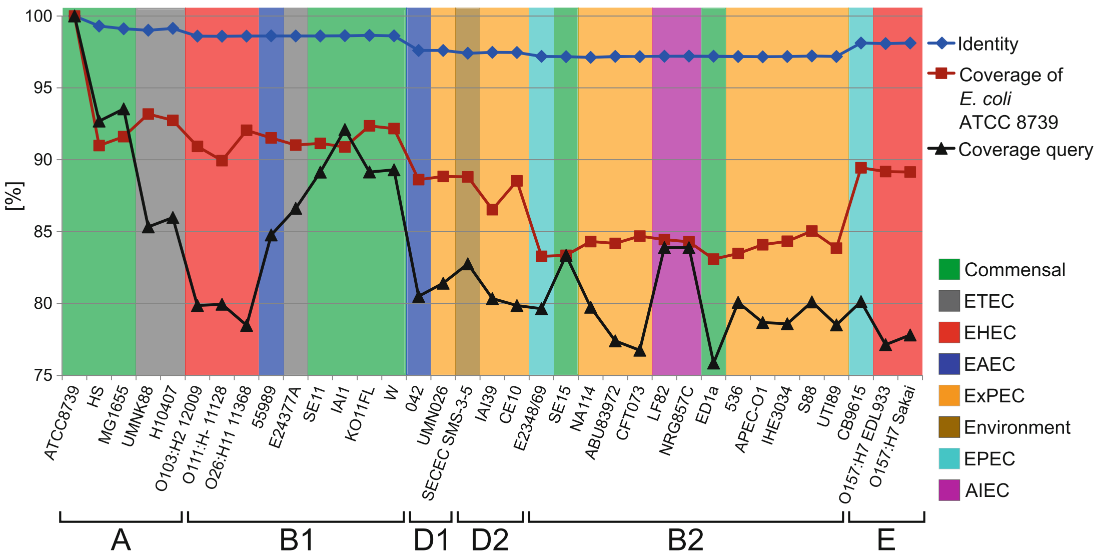
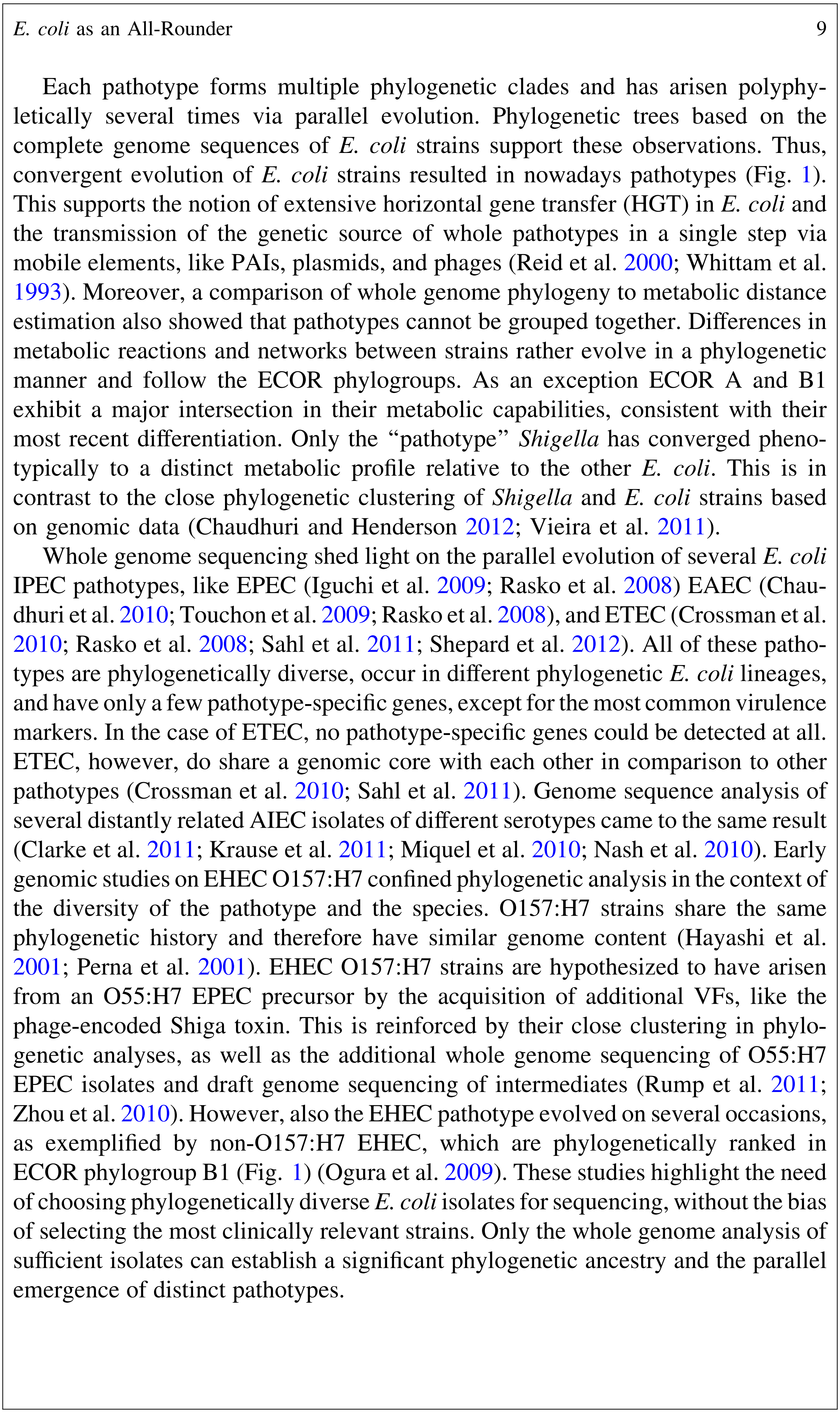


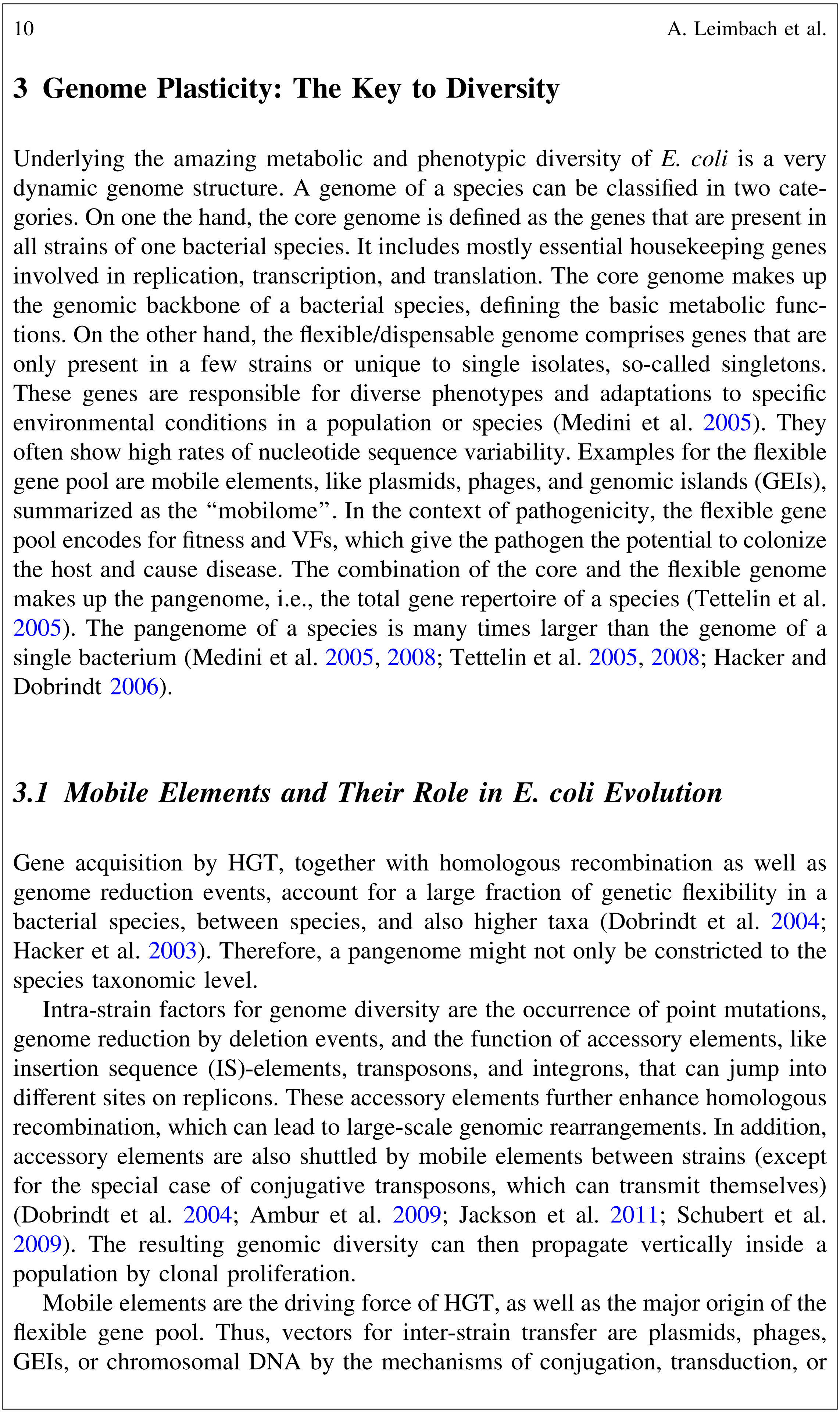
Fig. 2 Genome content comparison of completely sequenced *E. coli* strains. The commensal *E. coli* isolate ATCC 8739 was chosen as a reference. All the other *E. coli* genomes were aligned against the reference with the dnadiff script of the MUMmer package (version 3.22) (Kurtz et al. 2004). Coverage of *E. coli* ATCC 8739 is indicated in *red*, coverage of the respective query in *black*, and identity of the aligned sequences in *blue*. No correlation between pathotype and genome coverage of the reference or the query can be seen, but rather correlation between phylogroups

As solution emerged the new "gold standard" of phylogenetic analysis, whole genome phylogeny. The method leads to clear phylogenetic signals in *E. coli*, as exemplified by robust tree phylogenies calculated with different methods. As a result, the predominantly clonal population structure of *E. coli* can be used to delineate the major phylogenetic groups described above (Fig. 1) (Tenaillon et al. 2010; Chaudhuri and Henderson 2012; Sims and Kim 2011; Leopold et al. 2011). With this information, several genomic features can be related to the major phylogenetic groups outlined by the ECOR collection. For example, Rhs elements are arranged according to the phylogroups (Hill et al. 1995), but also VFs like the *Yersisina* 'high pathogenicity island' (HPI) (Clermont et al. 2001) and the putative type III secretion system ETT2 (Ren et al. 2004). Additionally, a connection can also be found in the distribution of extraintestinal pathotypes and phylogenetic ancestry. While strains from ECOR phylogroups A and B1 usually do not exhibit ExPEC phenotypes and lack ExPEC VFs, ECOR B2 and D cluster the majority of ExPEC strains (Boyd and Hartl 1998). The five major phylogenetic groups might even represent diverse ecological niches, as they have a different distribution in humans, domesticated animals, and wild animals (Tenaillon et al. 2010). There is, however, no direct correlation between the pathotype and phylogenetic lineage. Comparison of genome coverage and nucleotide identity of selected IPEC, ExPEC and commensal *E. coli* relative to non-pathogenic *E. coli* ATCC8739 could not reveal marked overlaps. No correlation among pathotype and genome coverage of the reference or the query can be seen, but rather correlation between phylogroups (Fig. 2).



Each pathotype forms multiple phylogenetic clades and has arisen polyphyletically several times via parallel evolution. Phylogenetic trees based on the complete genome sequences of *E. coli* strains support these observations. Thus, convergent evolution of *E. coli* strains resulted in nowadays pathotypes (Fig. 1). This supports the notion of extensive horizontal gene transfer (HGT) in *E. coli* and the transmission of the genetic source of whole pathotypes in a single step via mobile elements, like PAIs, plasmids, and phages (Reid et al. 2000; Whittam et al. 1993). Moreover, a comparison of whole genome phylogeny to metabolic distance estimation also showed that pathotypes cannot be grouped together. Differences in metabolic reactions and networks between strains rather evolve in a phylogenetic manner and follow the ECOR phylogroups. As an exception ECOR A and B1 exhibit a major intersection in their metabolic capabilities, consistent with their most recent differentiation. Only the "pathotype" *Shigella* has converged phenotypically to a distinct metabolic profile relative to the other *E. coli*. This is in contrast to the close phylogenetic clustering of *Shigella* and *E. coli* strains based on genomic data (Chaudhuri and Henderson 2012; Vieira et al. 2011).

Whole genome sequencing shed light on the parallel evolution of several *E. coli* IPEC pathotypes, like EPEC (Iguchi et al. 2009; Rasko et al. 2008) EAEC (Chaudhuri et al. 2010; Touchon et al. 2009; Rasko et al. 2008), and ETEC (Crossman et al. 2010; Rasko et al. 2008; Sahl et al. 2011; Shepard et al. 2012). All of these pathotypes are phylogenetically diverse, occur in different phylogenetic *E. coli* lineages, and have only a few pathotype-specific genes, except for the most common virulence markers. In the case of ETEC, no pathotype-specific genes could be detected at all. ETEC, however, do share a genomic core with each other in comparison to other pathotypes (Crossman et al. 2010; Sahl et al. 2011). Genome sequence analysis of several distantly related AIEC isolates of different serotypes came to the same result (Clarke et al. 2011; Krause et al. 2011; Miquel et al. 2010; Nash et al. 2010). Early genomic studies on EHEC O157:H7 confined phylogenetic analysis in the context of the diversity of the pathotype and the species. O157:H7 strains share the same phylogenetic history and therefore have similar genome content (Hayashi et al. 2001; Perna et al. 2001). EHEC O157:H7 strains are hypothesized to have arisen from an O55:H7 EPEC precursor by the acquisition of additional VFs, like the phage-encoded Shiga toxin. This is reinforced by their close clustering in phylogenetic analyses, as well as the additional whole genome sequencing of O55:H7 EPEC isolates and draft genome sequencing of intermediates (Rump et al. 2011; Zhou et al. 2010). However, also the EHEC pathotype evolved on several occasions, as exemplified by non-O157:H7 EHEC, which are phylogenetically ranked in ECOR phylogroup B1 (Fig. 1) (Ogura et al. 2009). These studies highlight the need of choosing phylogenetically diverse *E. coli* isolates for sequencing, without the bias of selecting the most clinically relevant strains. Only the whole genome analysis of sufficient isolates can establish a significant phylogenetic ancestry and the parallel emergence of distinct pathotypes.



3 Genome Plasticity: The Key to Diversity

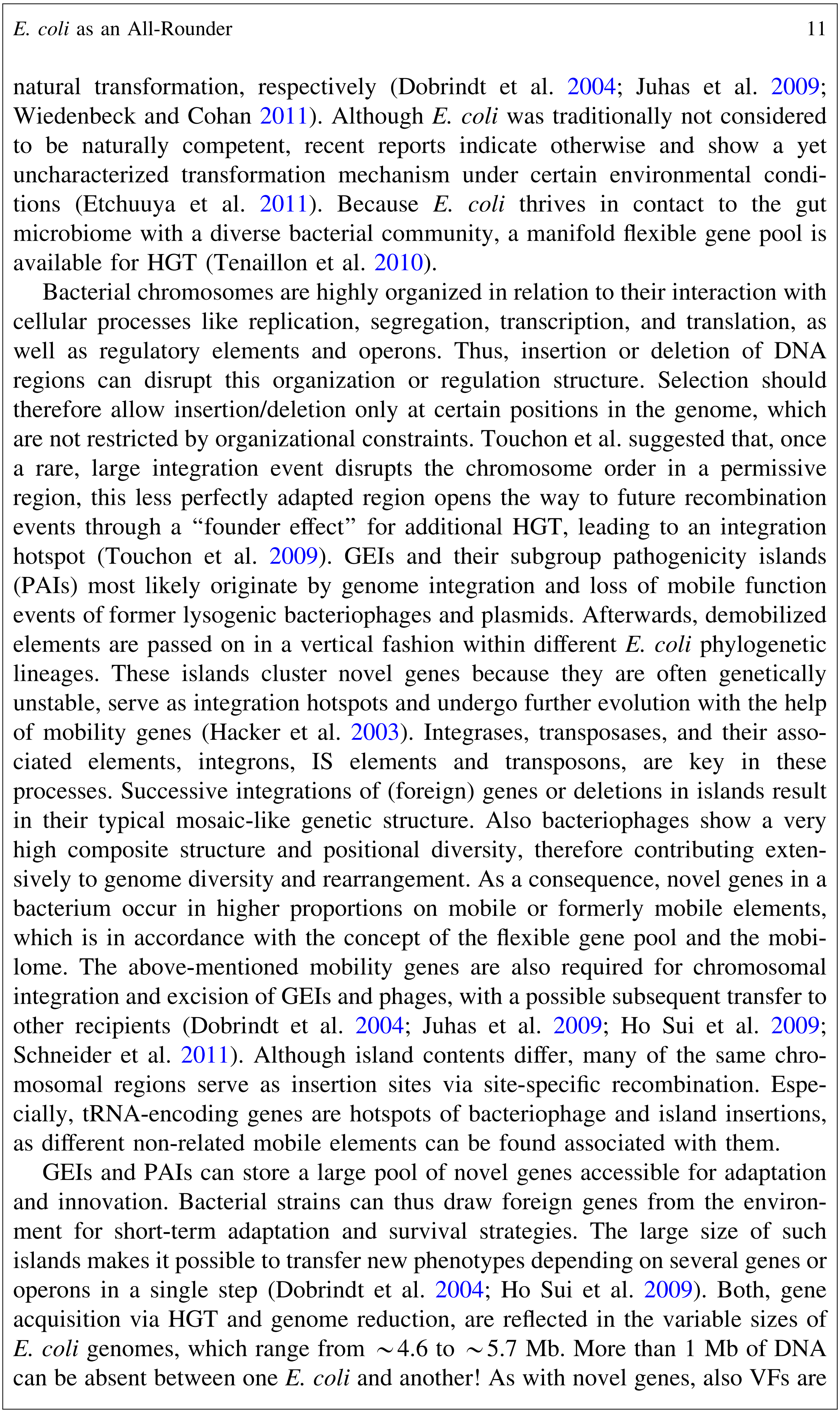
Underlying the amazing metabolic and phenotypic diversity of *E. coli* is a very dynamic genome structure. A genome of a species can be classified in two categories. On one the hand, the core genome is defined as the genes that are present in all strains of one bacterial species. It includes mostly essential housekeeping genes involved in replication, transcription, and translation. The core genome makes up the genomic backbone of a bacterial species, defining the basic metabolic functions. On the other hand, the flexible/dispensable genome comprises genes that are only present in a few strains or unique to single isolates, so-called singletons. These genes are responsible for diverse phenotypes and adaptations to specific environmental conditions in a population or species (Medini et al. 2005). They often show high rates of nucleotide sequence variability. Examples for the flexible gene pool are mobile elements, like plasmids, phages, and genomic islands (GEIs), summarized as the “mobilome”. In the context of pathogenicity, the flexible gene pool encodes for fitness and VFs, which give the pathogen the potential to colonize the host and cause disease. The combination of the core and the flexible genome makes up the pangenome, i.e., the total gene repertoire of a species (Tettelin et al. 2005). The pangenome of a species is many times larger than the genome of a single bacterium (Medini et al. 2005, 2008; Tettelin et al. 2005, 2008; Hacker and Dobrindt 2006).

3.1 Mobile Elements and Their Role in E. coli Evolution

Gene acquisition by HGT, together with homologous recombination as well as genome reduction events, account for a large fraction of genetic flexibility in a bacterial species, between species, and also higher taxa (Dobrindt et al. 2004; Hacker et al. 2003). Therefore, a pangenome might not only be constricted to the species taxonomic level.

Intra-strain factors for genome diversity are the occurrence of point mutations, genome reduction by deletion events, and the function of accessory elements, like insertion sequence (IS)-elements, transposons, and integrons, that can jump into different sites on replicons. These accessory elements further enhance homologous recombination, which can lead to large-scale genomic rearrangements. In addition, accessory elements are also shuttled by mobile elements between strains (except for the special case of conjugative transposons, which can transmit themselves) (Dobrindt et al. 2004; Ambur et al. 2009; Jackson et al. 2011; Schubert et al. 2009). The resulting genomic diversity can then propagate vertically inside a population by clonal proliferation.

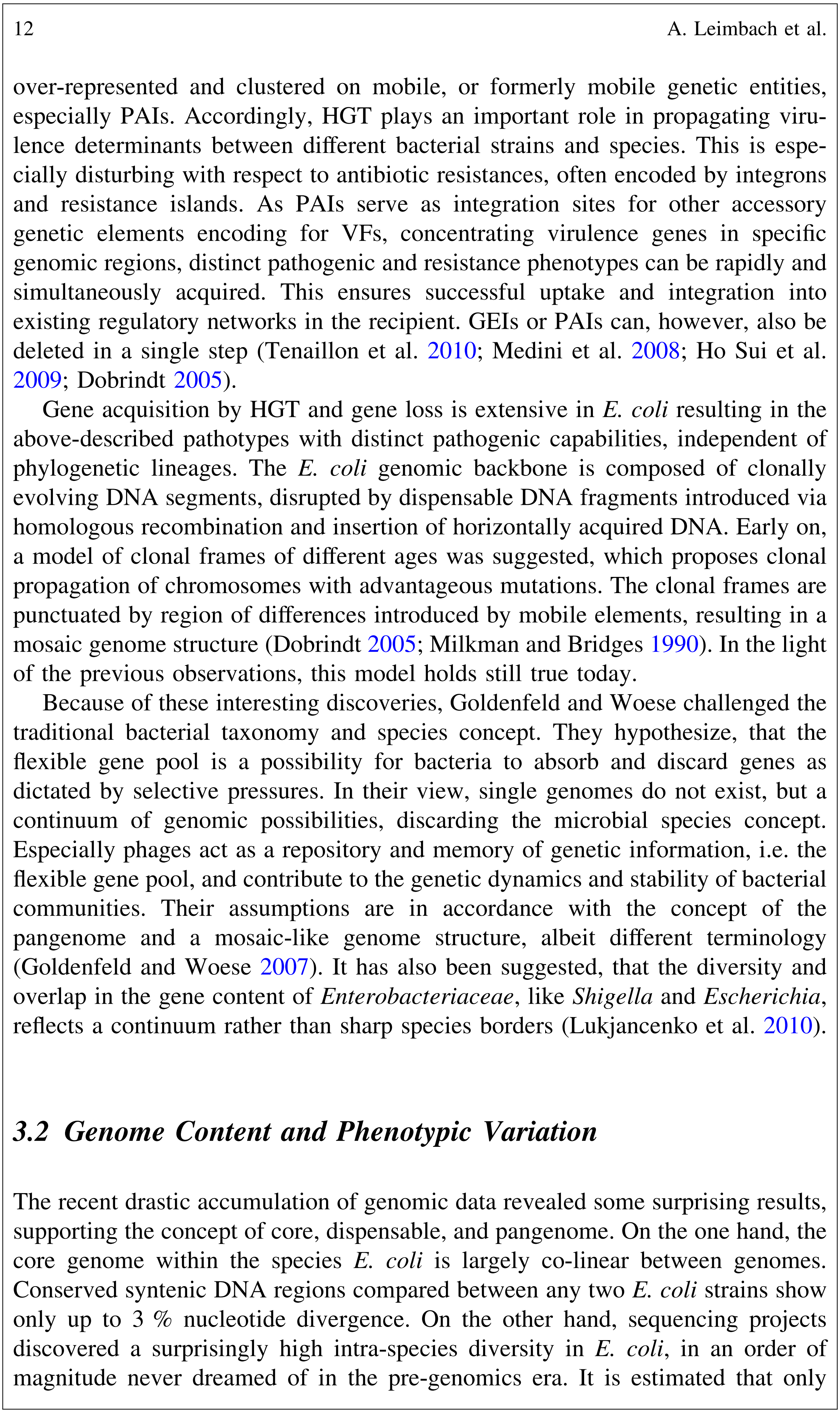
Mobile elements are the driving force of HGT, as well as the major origin of the flexible gene pool. Thus, vectors for inter-strain transfer are plasmids, phages, GEIs, or chromosomal DNA by the mechanisms of conjugation, transduction, or



natural transformation, respectively (Dobrindt et al. 2004; Juhas et al. 2009; Wiedenbeck and Cohan 2011). Although *E. coli* was traditionally not considered to be naturally competent, recent reports indicate otherwise and show a yet uncharacterized transformation mechanism under certain environmental conditions (Etchuuya et al. 2011). Because *E. coli* thrives in contact to the gut microbiome with a diverse bacterial community, a manifold flexible gene pool is available for HGT (Tenaillon et al. 2010).

Bacterial chromosomes are highly organized in relation to their interaction with cellular processes like replication, segregation, transcription, and translation, as well as regulatory elements and operons. Thus, insertion or deletion of DNA regions can disrupt this organization or regulation structure. Selection should therefore allow insertion/deletion only at certain positions in the genome, which are not restricted by organizational constraints. Touchon et al. suggested that, once a rare, large integration event disrupts the chromosome order in a permissive region, this less perfectly adapted region opens the way to future recombination events through a "founder effect" for additional HGT, leading to an integration hotspot (Touchon et al. 2009). GEIs and their subgroup pathogenicity islands (PAIs) most likely originate by genome integration and loss of mobile function events of former lysogenic bacteriophages and plasmids. Afterwards, demobilized elements are passed on in a vertical fashion within different *E. coli* phylogenetic lineages. These islands cluster novel genes because they are often genetically unstable, serve as integration hotspots and undergo further evolution with the help of mobility genes (Hacker et al. 2003). Integrases, transposases, and their associated elements, integrons, IS elements and transposons, are key in these processes. Successive integrations of (foreign) genes or deletions in islands result in their typical mosaic-like genetic structure. Also bacteriophages show a very high composite structure and positional diversity, therefore contributing extensively to genome diversity and rearrangement. As a consequence, novel genes in a bacterium occur in higher proportions on mobile or formerly mobile elements, which is in accordance with the concept of the flexible gene pool and the mobilome. The above-mentioned mobility genes are also required for chromosomal integration and excision of GEIs and phages, with a possible subsequent transfer to other recipients (Dobrindt et al. 2004; Juhas et al. 2009; Ho Sui et al. 2009; Schneider et al. 2011). Although island contents differ, many of the same chromosomal regions serve as insertion sites via site-specific recombination. Especially, tRNA-encoding genes are hotspots of bacteriophage and island insertions, as different non-related mobile elements can be found associated with them.

GEIs and PAIs can store a large pool of novel genes accessible for adaptation and innovation. Bacterial strains can thus draw foreign genes from the environment for short-term adaptation and survival strategies. The large size of such islands makes it possible to transfer new phenotypes depending on several genes or operons in a single step (Dobrindt et al. 2004; Ho Sui et al. 2009). Both, gene acquisition via HGT and genome reduction, are reflected in the variable sizes of *E. coli* genomes, which range from ~4.6 to ~5.7 Mb. More than 1 Mb of DNA can be absent between one *E. coli* and another! As with novel genes, also VFs are



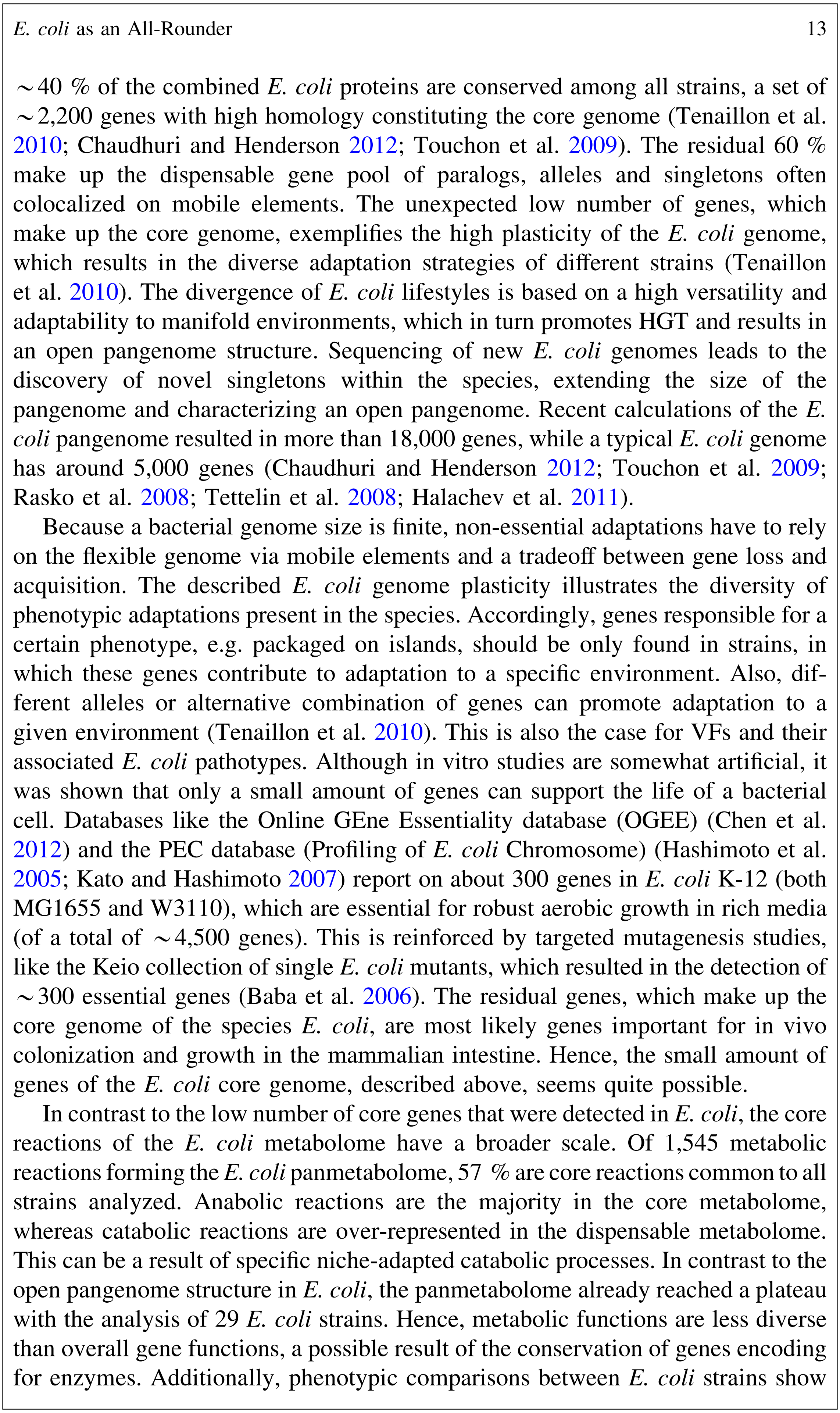
over-represented and clustered on mobile, or formerly mobile genetic entities, especially PAIs. Accordingly, HGT plays an important role in propagating virulence determinants between different bacterial strains and species. This is especially disturbing with respect to antibiotic resistances, often encoded by integrons and resistance islands. As PAIs serve as integration sites for other accessory genetic elements encoding for VFs, concentrating virulence genes in specific genomic regions, distinct pathogenic and resistance phenotypes can be rapidly and simultaneously acquired. This ensures successful uptake and integration into existing regulatory networks in the recipient. GEIs or PAIs can, however, also be deleted in a single step (Tenaillon et al. 2010; Medini et al. 2008; Ho Sui et al. 2009; Dobrindt 2005).

Gene acquisition by HGT and gene loss is extensive in *E. coli* resulting in the above-described pathotypes with distinct pathogenic capabilities, independent of phylogenetic lineages. The *E. coli* genomic backbone is composed of clonally evolving DNA segments, disrupted by dispensable DNA fragments introduced via homologous recombination and insertion of horizontally acquired DNA. Early on, a model of clonal frames of different ages was suggested, which proposes clonal propagation of chromosomes with advantageous mutations. The clonal frames are punctuated by region of differences introduced by mobile elements, resulting in a mosaic genome structure (Dobrindt 2005; Milkman and Bridges 1990). In the light of the previous observations, this model holds still true today.

Because of these interesting discoveries, Goldenfeld and Woese challenged the traditional bacterial taxonomy and species concept. They hypothesize, that the flexible gene pool is a possibility for bacteria to absorb and discard genes as dictated by selective pressures. In their view, single genomes do not exist, but a continuum of genomic possibilities, discarding the microbial species concept. Especially phages act as a repository and memory of genetic information, i.e. the flexible gene pool, and contribute to the genetic dynamics and stability of bacterial communities. Their assumptions are in accordance with the concept of the pangenome and a mosaic-like genome structure, albeit different terminology (Goldenfeld and Woese 2007). It has also been suggested, that the diversity and overlap in the gene content of *Enterobacteriaceae*, like *Shigella* and *Escherichia*, reflects a continuum rather than sharp species borders (Lukjancenko et al. 2010).

3.2 Genome Content and Phenotypic Variation

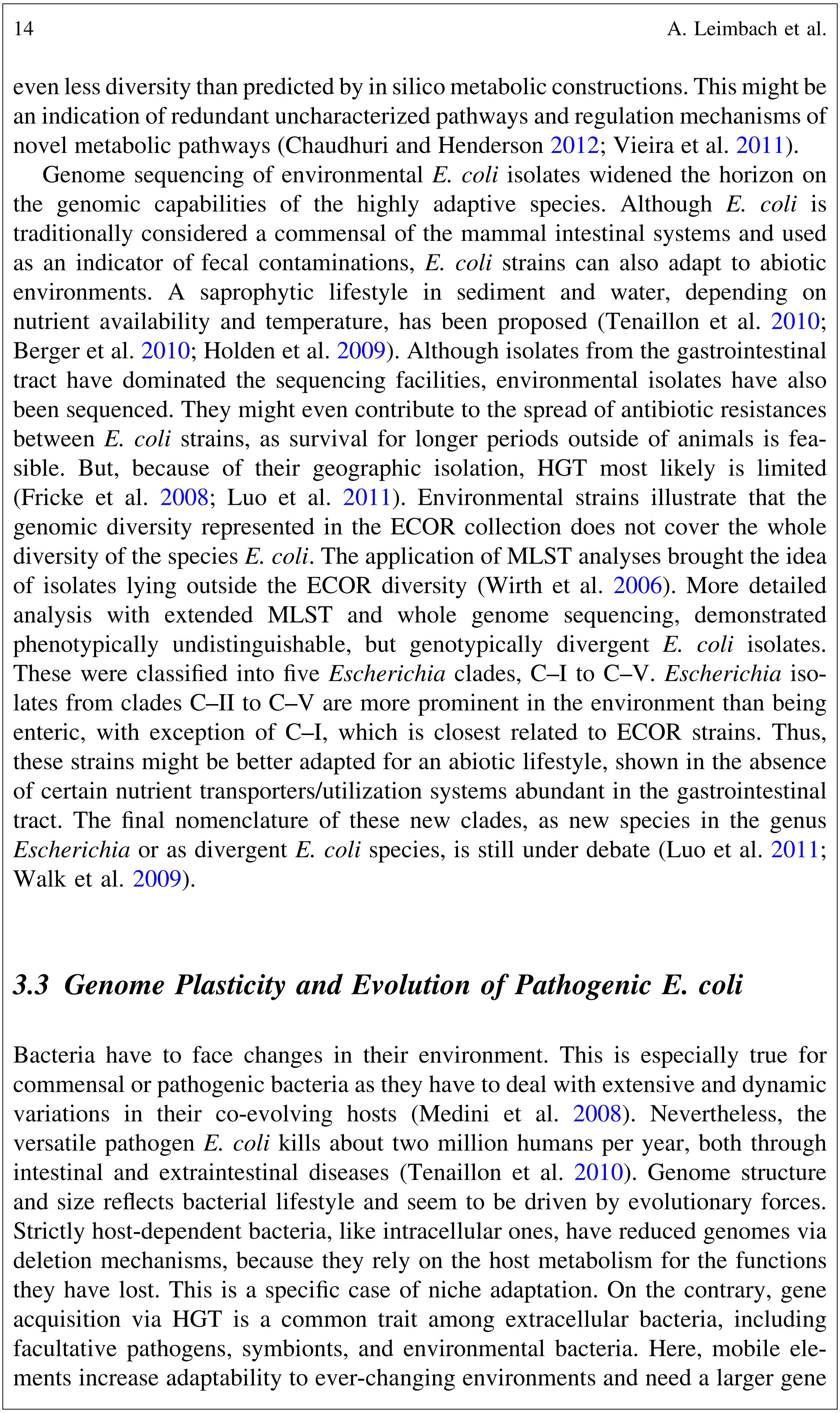
The recent drastic accumulation of genomic data revealed some surprising results, supporting the concept of core, dispensable, and pangenome. On the one hand, the core genome within the species *E. coli* is largely co-linear between genomes. Conserved syntenic DNA regions compared between any two *E. coli* strains show only up to 3 % nucleotide divergence. On the other hand, sequencing projects discovered a surprisingly high intra-species diversity in *E. coli*, in an order of magnitude never dreamed of in the pre-genomics era. It is estimated that only



~40 % of the combined *E. coli* proteins are conserved among all strains, a set of ~2,200 genes with high homology constituting the core genome (Tenaillon et al. 2010; Chaudhuri and Henderson 2012; Touchon et al. 2009). The residual 60 % make up the dispensable gene pool of paralogs, alleles and singletons often colocalized on mobile elements. The unexpected low number of genes, which make up the core genome, exemplifies the high plasticity of the *E. coli* genome, which results in the diverse adaptation strategies of different strains (Tenaillon et al. 2010). The divergence of *E. coli* lifestyles is based on a high versatility and adaptability to manifold environments, which in turn promotes HGT and results in an open pangenome structure. Sequencing of new *E. coli* genomes leads to the discovery of novel singletons within the species, extending the size of the pangenome and characterizing an open pangenome. Recent calculations of the *E. coli* pangenome resulted in more than 18,000 genes, while a typical *E. coli* genome has around 5,000 genes (Chaudhuri and Henderson 2012; Touchon et al. 2009; Rasko et al. 2008; Tettelin et al. 2008; Halachev et al. 2011).

Because a bacterial genome size is finite, non-essential adaptations have to rely on the flexible genome via mobile elements and a tradeoff between gene loss and acquisition. The described *E. coli* genome plasticity illustrates the diversity of phenotypic adaptations present in the species. Accordingly, genes responsible for a certain phenotype, e.g. packaged on islands, should be only found in strains, in which these genes contribute to adaptation to a specific environment. Also, different alleles or alternative combination of genes can promote adaptation to a given environment (Tenaillon et al. 2010). This is also the case for VFs and their associated *E. coli* pathotypes. Although in vitro studies are somewhat artificial, it was shown that only a small amount of genes can support the life of a bacterial cell. Databases like the Online GEne Essentiality database (OGEE) (Chen et al. 2012) and the PEC database (Profiling of *E. coli* Chromosome) (Hashimoto et al. 2005; Kato and Hashimoto 2007) report on about 300 genes in *E. coli* K-12 (both MG1655 and W3110), which are essential for robust aerobic growth in rich media (of a total of ~4,500 genes). This is reinforced by targeted mutagenesis studies, like the Keio collection of single *E. coli* mutants, which resulted in the detection of ~300 essential genes (Baba et al. 2006). The residual genes, which make up the core genome of the species *E. coli*, are most likely genes important for in vivo colonization and growth in the mammalian intestine. Hence, the small amount of genes of the *E. coli* core genome, described above, seems quite possible.

In contrast to the low number of core genes that were detected in *E. coli*, the core reactions of the *E. coli* metabolome have a broader scale. Of 1,545 metabolic reactions forming the *E. coli* panmetabolome, 57 % are core reactions common to all strains analyzed. Anabolic reactions are the majority in the core metabolome, whereas catabolic reactions are over-represented in the dispensable metabolome. This can be a result of specific niche-adapted catabolic processes. In contrast to the open pangenome structure in *E. coli*, the panmetabolome already reached a plateau with the analysis of 29 *E. coli* strains. Hence, metabolic functions are less diverse than overall gene functions, a possible result of the conservation of genes encoding for enzymes. Additionally, phenotypic comparisons between *E. coli* strains show

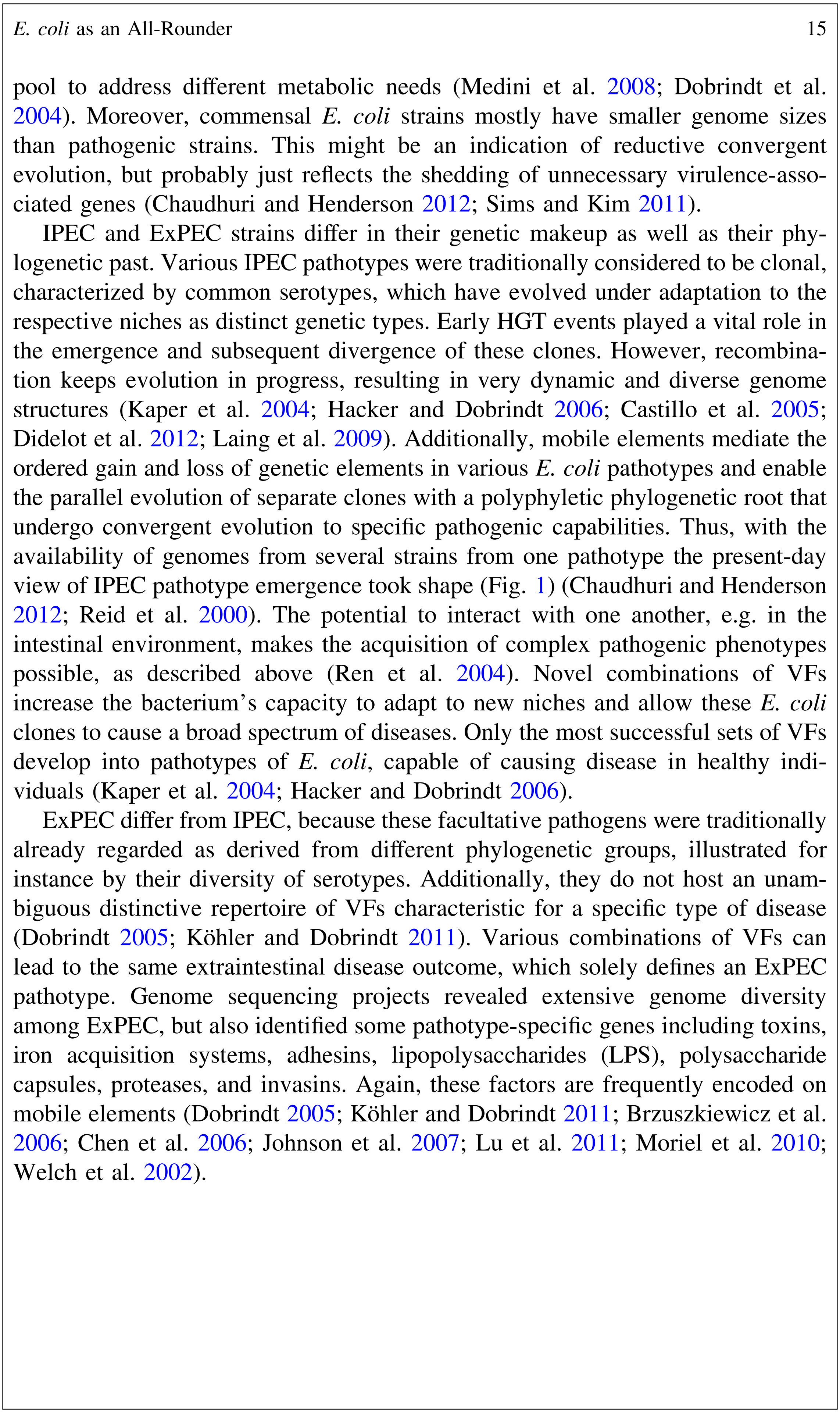


even less diversity than predicted by in silico metabolic constructions. This might be an indication of redundant uncharacterized pathways and regulation mechanisms of novel metabolic pathways (Chaudhuri and Henderson 2012; Vieira et al. 2011).

Genome sequencing of environmental *E. coli* isolates widened the horizon on the genomic capabilities of the highly adaptive species. Although *E. coli* is traditionally considered a commensal of the mammal intestinal systems and used as an indicator of fecal contaminations, *E. coli* strains can also adapt to abiotic environments. A saprophytic lifestyle in sediment and water, depending on nutrient availability and temperature, has been proposed (Tenaillon et al. 2010; Berger et al. 2010; Holden et al. 2009). Although isolates from the gastrointestinal tract have dominated the sequencing facilities, environmental isolates have also been sequenced. They might even contribute to the spread of antibiotic resistances between *E. coli* strains, as survival for longer periods outside of animals is feasible. But, because of their geographic isolation, HGT most likely is limited (Fricke et al. 2008; Luo et al. 2011). Environmental strains illustrate that the genomic diversity represented in the ECOR collection does not cover the whole diversity of the species *E. coli*. The application of MLST analyses brought the idea of isolates lying outside the ECOR diversity (Wirth et al. 2006). More detailed analysis with extended MLST and whole genome sequencing, demonstrated phenotypically undistinguishable, but genotypically divergent *E. coli* isolates. These were classified into five *Escherichia* clades, C–I to C–V. *Escherichia* isolates from clades C–II to C–V are more prominent in the environment than being enteric, with exception of C–I, which is closest related to ECOR strains. Thus, these strains might be better adapted for an abiotic lifestyle, shown in the absence of certain nutrient transporters/utilization systems abundant in the gastrointestinal tract. The final nomenclature of these new clades, as new species in the genus *Escherichia* or as divergent *E. coli* species, is still under debate (Luo et al. 2011; Walk et al. 2009).

3.3 Genome Plasticity and Evolution of Pathogenic *E. coli*

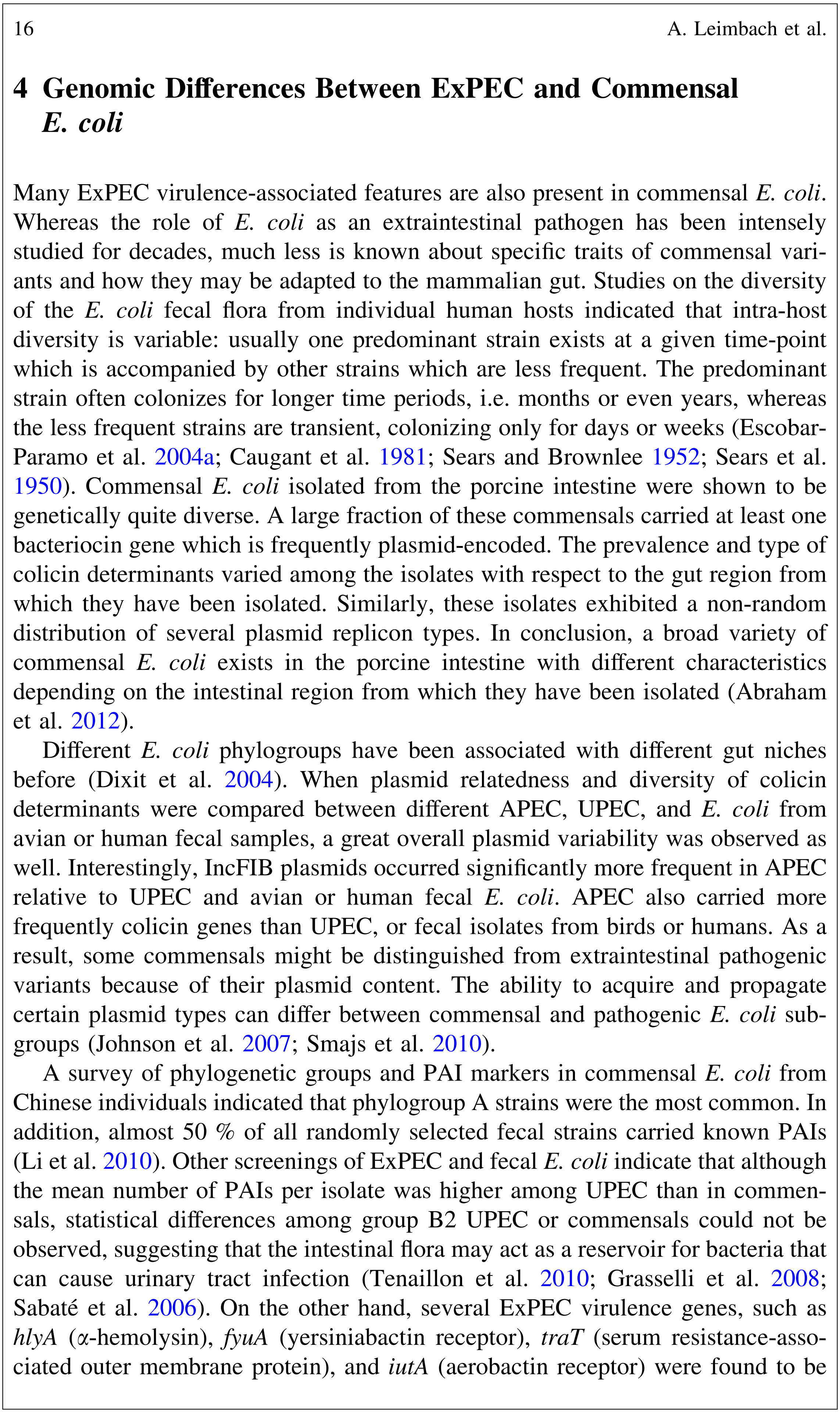
Bacteria have to face changes in their environment. This is especially true for commensal or pathogenic bacteria as they have to deal with extensive and dynamic variations in their co-evolving hosts (Medini et al. 2008). Nevertheless, the versatile pathogen *E. coli* kills about two million humans per year, both through intestinal and extraintestinal diseases (Tenaillon et al. 2010). Genome structure and size reflects bacterial lifestyle and seem to be driven by evolutionary forces. Strictly host-dependent bacteria, like intracellular ones, have reduced genomes via deletion mechanisms, because they rely on the host metabolism for the functions they have lost. This is a specific case of niche adaptation. On the contrary, gene acquisition via HGT is a common trait among extracellular bacteria, including facultative pathogens, symbionts, and environmental bacteria. Here, mobile elements increase adaptability to ever-changing environments and need a larger gene



pool to address different metabolic needs (Medini et al. 2008; Dobrindt et al. 2004). Moreover, commensal *E. coli* strains mostly have smaller genome sizes than pathogenic strains. This might be an indication of reductive convergent evolution, but probably just reflects the shedding of unnecessary virulence-associated genes (Chaudhuri and Henderson 2012; Sims and Kim 2011).

IPEC and ExPEC strains differ in their genetic makeup as well as their phylogenetic past. Various IPEC pathotypes were traditionally considered to be clonal, characterized by common serotypes, which have evolved under adaptation to the respective niches as distinct genetic types. Early HGT events played a vital role in the emergence and subsequent divergence of these clones. However, recombination keeps evolution in progress, resulting in very dynamic and diverse genome structures (Kaper et al. 2004; Hacker and Dobrindt 2006; Castillo et al. 2005; Didelot et al. 2012; Laing et al. 2009). Additionally, mobile elements mediate the ordered gain and loss of genetic elements in various *E. coli* pathotypes and enable the parallel evolution of separate clones with a polyphyletic phylogenetic root that undergo convergent evolution to specific pathogenic capabilities. Thus, with the availability of genomes from several strains from one pathotype the present-day view of IPEC pathotype emergence took shape (Fig. 1) (Chaudhuri and Henderson 2012; Reid et al. 2000). The potential to interact with one another, e.g. in the intestinal environment, makes the acquisition of complex pathogenic phenotypes possible, as described above (Ren et al. 2004). Novel combinations of VFs increase the bacterium's capacity to adapt to new niches and allow these *E. coli* clones to cause a broad spectrum of diseases. Only the most successful sets of VFs develop into pathotypes of *E. coli*, capable of causing disease in healthy individuals (Kaper et al. 2004; Hacker and Dobrindt 2006).

ExPEC differ from IPEC, because these facultative pathogens were traditionally already regarded as derived from different phylogenetic groups, illustrated for instance by their diversity of serotypes. Additionally, they do not host an unambiguous distinctive repertoire of VFs characteristic for a specific type of disease (Dobrindt 2005; Köhler and Dobrindt 2011). Various combinations of VFs can lead to the same extraintestinal disease outcome, which solely defines an ExPEC pathotype. Genome sequencing projects revealed extensive genome diversity among ExPEC, but also identified some pathotype-specific genes including toxins, iron acquisition systems, adhesins, lipopolysaccharides (LPS), polysaccharide capsules, proteases, and invasins. Again, these factors are frequently encoded on mobile elements (Dobrindt 2005; Köhler and Dobrindt 2011; Brzuszkiewicz et al. 2006; Chen et al. 2006; Johnson et al. 2007; Lu et al. 2011; Moriel et al. 2010; Welch et al. 2002).

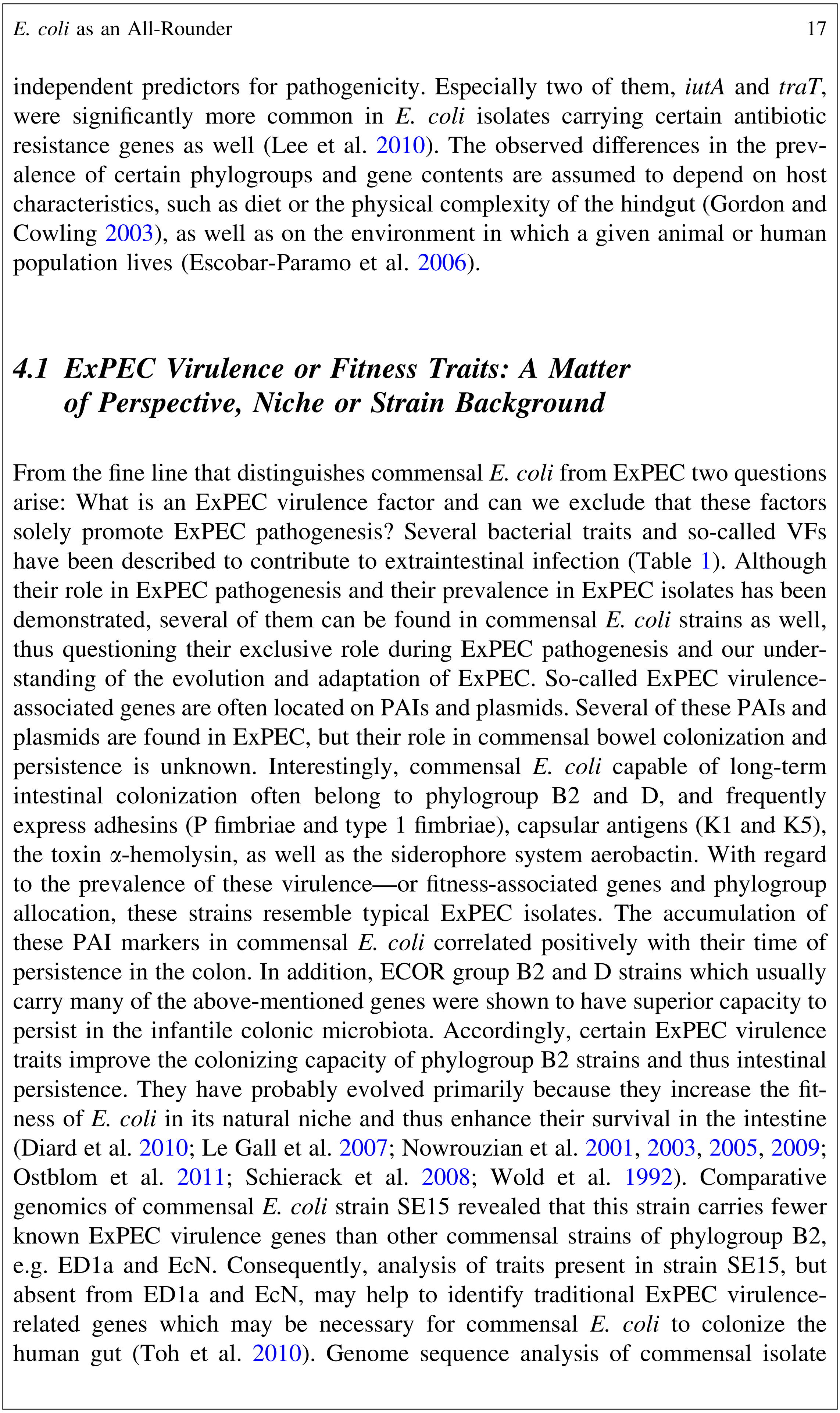


4 Genomic Differences Between ExPEC and Commensal *E. coli*

Many ExPEC virulence-associated features are also present in commensal *E. coli*. Whereas the role of *E. coli* as an extraintestinal pathogen has been intensely studied for decades, much less is known about specific traits of commensal variants and how they may be adapted to the mammalian gut. Studies on the diversity of the *E. coli* fecal flora from individual human hosts indicated that intra-host diversity is variable: usually one predominant strain exists at a given time-point which is accompanied by other strains which are less frequent. The predominant strain often colonizes for longer time periods, i.e. months or even years, whereas the less frequent strains are transient, colonizing only for days or weeks (Escobar-Paramo et al. 2004a; Caugant et al. 1981; Sears and Brownlee 1952; Sears et al. 1950). Commensal *E. coli* isolated from the porcine intestine were shown to be genetically quite diverse. A large fraction of these commensals carried at least one bacteriocin gene which is frequently plasmid-encoded. The prevalence and type of colicin determinants varied among the isolates with respect to the gut region from which they have been isolated. Similarly, these isolates exhibited a non-random distribution of several plasmid replicon types. In conclusion, a broad variety of commensal *E. coli* exists in the porcine intestine with different characteristics depending on the intestinal region from which they have been isolated (Abraham et al. 2012).

Different *E. coli* phylogroups have been associated with different gut niches before (Dixit et al. 2004). When plasmid relatedness and diversity of colicin determinants were compared between different APEC, UPEC, and *E. coli* from avian or human fecal samples, a great overall plasmid variability was observed as well. Interestingly, IncFIB plasmids occurred significantly more frequent in APEC relative to UPEC and avian or human fecal *E. coli*. APEC also carried more frequently colicin genes than UPEC, or fecal isolates from birds or humans. As a result, some commensals might be distinguished from extraintestinal pathogenic variants because of their plasmid content. The ability to acquire and propagate certain plasmid types can differ between commensal and pathogenic *E. coli* subgroups (Johnson et al. 2007; Smajs et al. 2010).

A survey of phylogenetic groups and PAI markers in commensal *E. coli* from Chinese individuals indicated that phylogroup A strains were the most common. In addition, almost 50 % of all randomly selected fecal strains carried known PAIs (Li et al. 2010). Other screenings of ExPEC and fecal *E. coli* indicate that although the mean number of PAIs per isolate was higher among UPEC than in commensals, statistical differences among group B2 UPEC or commensals could not be observed, suggesting that the intestinal flora may act as a reservoir for bacteria that can cause urinary tract infection (Tenaillon et al. 2010; Grasselli et al. 2008; Sabaté et al. 2006). On the other hand, several ExPEC virulence genes, such as *hlyA* (α-hemolysin), *fyuA* (yersiniabactin receptor), *traT* (serum resistance-associated outer membrane protein), and *iutA* (aerobactin receptor) were found to be



independent predictors for pathogenicity. Especially two of them, *iutA* and *traT*, were significantly more common in *E. coli* isolates carrying certain antibiotic resistance genes as well (Lee et al. 2010). The observed differences in the prevalence of certain phylogroups and gene contents are assumed to depend on host characteristics, such as diet or the physical complexity of the hindgut (Gordon and Cowling 2003), as well as on the environment in which a given animal or human population lives (Escobar-Paramo et al. 2006).

4.1 ExPEC Virulence or Fitness Traits: A Matter of Perspective, Niche or Strain Background

From the fine line that distinguishes commensal *E. coli* from ExPEC two questions arise: What is an ExPEC virulence factor and can we exclude that these factors solely promote ExPEC pathogenesis? Several bacterial traits and so-called VFs have been described to contribute to extraintestinal infection (Table 1). Although their role in ExPEC pathogenesis and their prevalence in ExPEC isolates has been demonstrated, several of them can be found in commensal *E. coli* strains as well, thus questioning their exclusive role during ExPEC pathogenesis and our understanding of the evolution and adaptation of ExPEC. So-called ExPEC virulence-associated genes are often located on PAIs and plasmids. Several of these PAIs and plasmids are found in ExPEC, but their role in commensal bowel colonization and persistence is unknown. Interestingly, commensal *E. coli* capable of long-term intestinal colonization often belong to phylogroup B2 and D, and frequently express adhesins (P fimbriae and type 1 fimbriae), capsular antigens (K1 and K5), the toxin α-hemolysin, as well as the siderophore system aerobactin. With regard to the prevalence of these virulence—or fitness-associated genes and phylogroup allocation, these strains resemble typical ExPEC isolates. The accumulation of these PAI markers in commensal *E. coli* correlated positively with their time of persistence in the colon. In addition, ECOR group B2 and D strains which usually carry many of the above-mentioned genes were shown to have superior capacity to persist in the infantile colonic microbiota. Accordingly, certain ExPEC virulence traits improve the colonizing capacity of phylogroup B2 strains and thus intestinal persistence. They have probably evolved primarily because they increase the fitness of *E. coli* in its natural niche and thus enhance their survival in the intestine (Diard et al. 2010; Le Gall et al. 2007; Nowrouzian et al. 2001, 2003, 2005, 2009; Ostblom et al. 2011; Schierack et al. 2008; Wold et al. 1992). Comparative genomics of commensal *E. coli* strain SE15 revealed that this strain carries fewer known ExPEC virulence genes than other commensal strains of phylogroup B2, e.g. ED1a and EcN. Consequently, analysis of traits present in strain SE15, but absent from ED1a and EcN, may help to identify traditional ExPEC virulence-related genes which may be necessary for commensal *E. coli* to colonize the human gut (Toh et al. 2010). Genome sequence analysis of commensal isolate

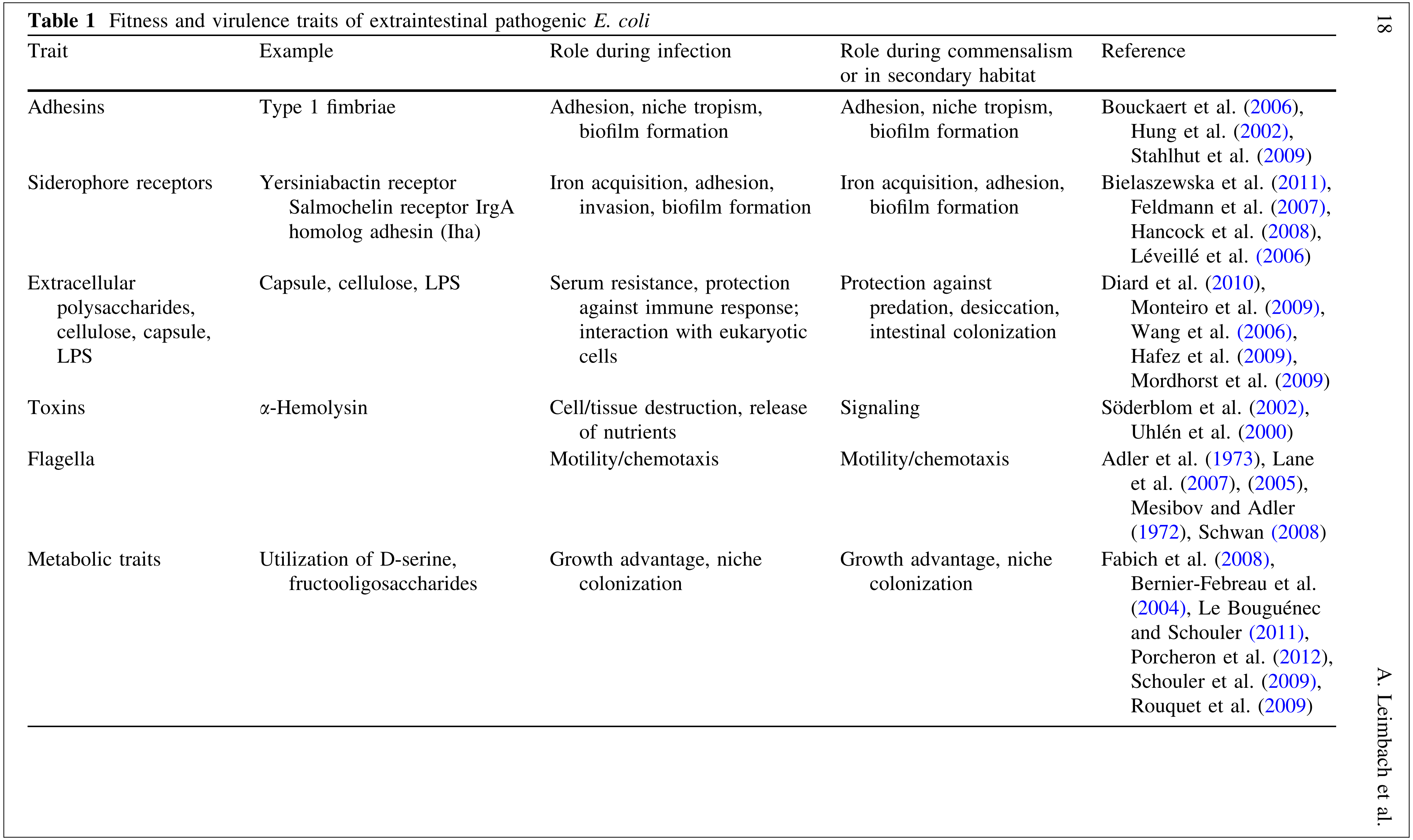
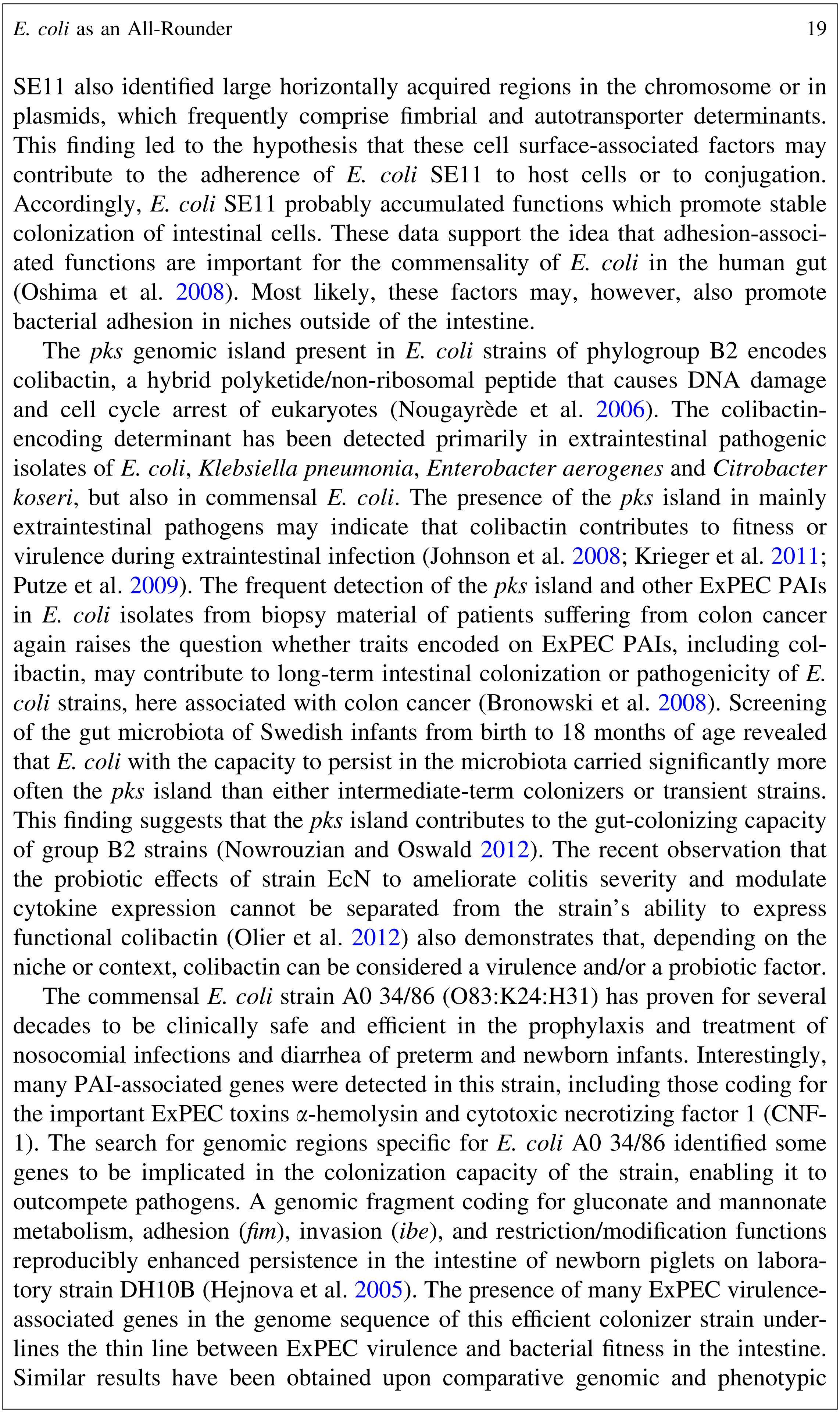


Table 1 Fitness and virulence traits of extraintestinal pathogenic *E. coli*

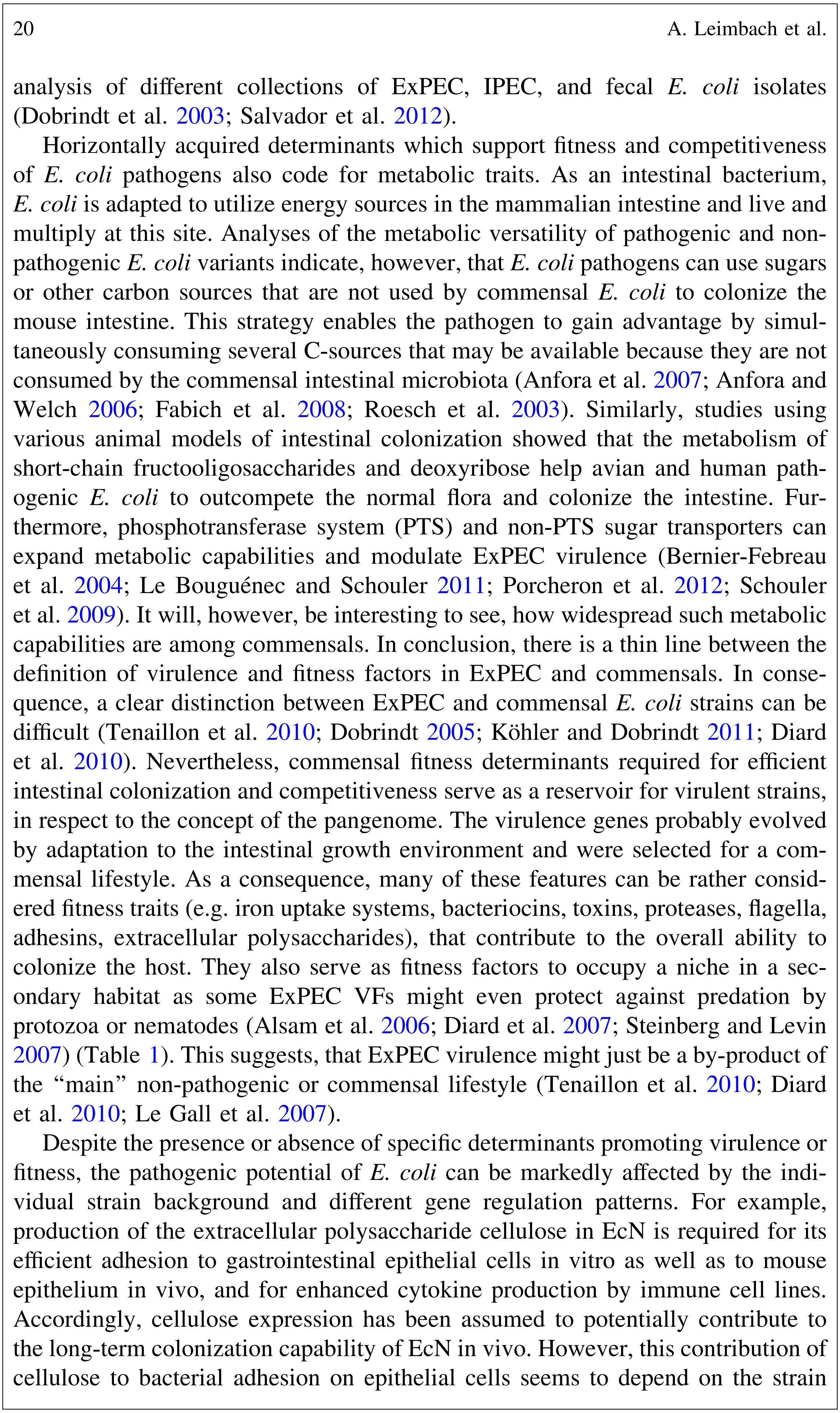
Trait	Example	Role during infection	Role during commensalism or in secondary habitat	Reference
Adhesins	Type 1 fimbriae	Adhesion, niche tropism, biofilm formation	Adhesion, niche tropism, biofilm formation	Bouckaert et al. (2006), Hung et al. (2002), Stahlhut et al. (2009)
Siderophore receptors	Yersiniabactin receptor Salmochelin receptor IrgA homolog adhesin (Iha)	Iron acquisition, adhesion, invasion, biofilm formation	Iron acquisition, adhesion, biofilm formation	Bielaszewska et al. (2011), Feldmann et al. (2007), Hancock et al. (2008), Léveillé et al. (2006)
Extracellular polysaccharides, cellulose, capsule, LPS	Capsule, cellulose, LPS	Serum resistance, protection against immune response; interaction with eukaryotic cells	Protection against predation, desiccation, intestinal colonization	Diard et al. (2010), Monteiro et al. (2009), Wang et al. (2006), Hafez et al. (2009), Mordhorst et al. (2009)
Toxins	α-Hemolysin	Cell/tissue destruction, release of nutrients	Signaling	Söderblom et al. (2002), Uhlén et al. (2000)
Flagella		Motility/chemotaxis	Motility/chemotaxis	Adler et al. (1973), Lane et al. (2007), (2005), Mesibov and Adler (1972), Schwan (2008)
Metabolic traits	Utilization of D-serine, fructooligosaccharides	Growth advantage, niche colonization	Growth advantage, niche colonization	Fabich et al. (2008), Bernier-Febreau et al. (2004), Le Bouguénec and Schouler (2011), Porcheron et al. (2012), Schouler et al. (2009), Rouquet et al. (2009)



SE11 also identified large horizontally acquired regions in the chromosome or in plasmids, which frequently comprise fimbrial and autotransporter determinants. This finding led to the hypothesis that these cell surface-associated factors may contribute to the adherence of *E. coli* SE11 to host cells or to conjugation. Accordingly, *E. coli* SE11 probably accumulated functions which promote stable colonization of intestinal cells. These data support the idea that adhesion-associated functions are important for the commensality of *E. coli* in the human gut (Oshima et al. 2008). Most likely, these factors may, however, also promote bacterial adhesion in niches outside of the intestine.

The *pks* genomic island present in *E. coli* strains of phylogroup B2 encodes colibactin, a hybrid polyketide/non-ribosomal peptide that causes DNA damage and cell cycle arrest of eukaryotes (Nougayrède et al. 2006). The colibactin-encoding determinant has been detected primarily in extraintestinal pathogenic isolates of *E. coli*, *Klebsiella pneumonia*, *Enterobacter aerogenes* and *Citrobacter koseri*, but also in commensal *E. coli*. The presence of the *pks* island in mainly extraintestinal pathogens may indicate that colibactin contributes to fitness or virulence during extraintestinal infection (Johnson et al. 2008; Krieger et al. 2011; Putze et al. 2009). The frequent detection of the *pks* island and other ExPEC PAIs in *E. coli* isolates from biopsy material of patients suffering from colon cancer again raises the question whether traits encoded on ExPEC PAIs, including colibactin, may contribute to long-term intestinal colonization or pathogenicity of *E. coli* strains, here associated with colon cancer (Bronowski et al. 2008). Screening of the gut microbiota of Swedish infants from birth to 18 months of age revealed that *E. coli* with the capacity to persist in the microbiota carried significantly more often the *pks* island than either intermediate-term colonizers or transient strains. This finding suggests that the *pks* island contributes to the gut-colonizing capacity of group B2 strains (Nowrouzian and Oswald 2012). The recent observation that the probiotic effects of strain EcN to ameliorate colitis severity and modulate cytokine expression cannot be separated from the strain's ability to express functional colibactin (Olier et al. 2012) also demonstrates that, depending on the niche or context, colibactin can be considered a virulence and/or a probiotic factor.

The commensal *E. coli* strain A0 34/86 (O83:K24:H31) has proven for several decades to be clinically safe and efficient in the prophylaxis and treatment of nosocomial infections and diarrhea of preterm and newborn infants. Interestingly, many PAI-associated genes were detected in this strain, including those coding for the important ExPEC toxins α-hemolysin and cytotoxic necrotizing factor 1 (CNF-1). The search for genomic regions specific for *E. coli* A0 34/86 identified some genes to be implicated in the colonization capacity of the strain, enabling it to outcompete pathogens. A genomic fragment coding for gluconate and mannonate metabolism, adhesion (*fim*), invasion (*ibe*), and restriction/modification functions reproducibly enhanced persistence in the intestine of newborn piglets on laboratory strain DH10B (Hejnova et al. 2005). The presence of many ExPEC virulence-associated genes in the genome sequence of this efficient colonizer strain underlines the thin line between ExPEC virulence and bacterial fitness in the intestine. Similar results have been obtained upon comparative genomic and phenotypic



20 A. Leimbach et al.

analysis of different collections of ExPEC, IPEC, and fecal *E. coli* isolates (Dobrindt et al. 2003; Salvador et al. 2012).

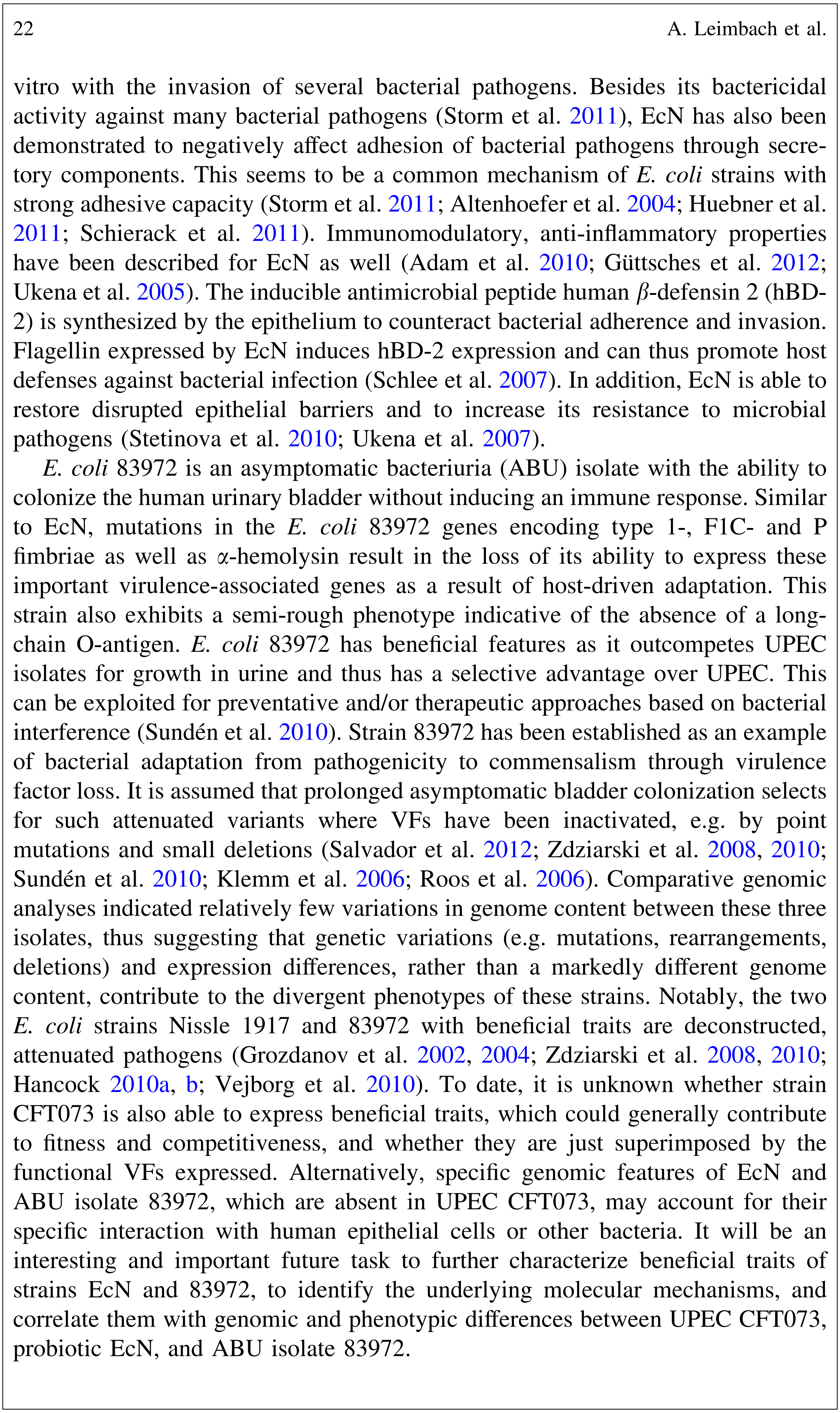
Horizontally acquired determinants which support fitness and competitiveness of *E. coli* pathogens also code for metabolic traits. As an intestinal bacterium, *E. coli* is adapted to utilize energy sources in the mammalian intestine and live and multiply at this site. Analyses of the metabolic versatility of pathogenic and non-pathogenic *E. coli* variants indicate, however, that *E. coli* pathogens can use sugars or other carbon sources that are not used by commensal *E. coli* to colonize the mouse intestine. This strategy enables the pathogen to gain advantage by simultaneously consuming several C-sources that may be available because they are not consumed by the commensal intestinal microbiota (Anfora et al. 2007; Anfora and Welch 2006; Fabich et al. 2008; Roesch et al. 2003). Similarly, studies using various animal models of intestinal colonization showed that the metabolism of short-chain fructooligosaccharides and deoxyribose help avian and human pathogenic *E. coli* to outcompete the normal flora and colonize the intestine. Furthermore, phosphotransferase system (PTS) and non-PTS sugar transporters can expand metabolic capabilities and modulate ExPEC virulence (Bernier-Febreau et al. 2004; Le Bouguénec and Schouler 2011; Porcheron et al. 2012; Schouler et al. 2009). It will, however, be interesting to see, how widespread such metabolic capabilities are among commensals. In conclusion, there is a thin line between the definition of virulence and fitness factors in ExPEC and commensals. In consequence, a clear distinction between ExPEC and commensal *E. coli* strains can be difficult (Tenaillon et al. 2010; Dobrindt 2005; Köhler and Dobrindt 2011; Diard et al. 2010). Nevertheless, commensal fitness determinants required for efficient intestinal colonization and competitiveness serve as a reservoir for virulent strains, in respect to the concept of the pangenome. The virulence genes probably evolved by adaptation to the intestinal growth environment and were selected for a commensal lifestyle. As a consequence, many of these features can be rather considered fitness traits (e.g. iron uptake systems, bacteriocins, toxins, proteases, flagella, adhesins, extracellular polysaccharides), that contribute to the overall ability to colonize the host. They also serve as fitness factors to occupy a niche in a secondary habitat as some ExPEC VFs might even protect against predation by protozoa or nematodes (Alsam et al. 2006; Diard et al. 2007; Steinberg and Levin 2007) (Table 1). This suggests, that ExPEC virulence might just be a by-product of the “main” non-pathogenic or commensal lifestyle (Tenaillon et al. 2010; Diard et al. 2010; Le Gall et al. 2007).

Despite the presence or absence of specific determinants promoting virulence or fitness, the pathogenic potential of *E. coli* can be markedly affected by the individual strain background and different gene regulation patterns. For example, production of the extracellular polysaccharide cellulose in EcN is required for its efficient adhesion to gastrointestinal epithelial cells in vitro as well as to mouse epithelium in vivo, and for enhanced cytokine production by immune cell lines. Accordingly, cellulose expression has been assumed to potentially contribute to the long-term colonization capability of EcN in vivo. However, this contribution of cellulose to bacterial adhesion on epithelial cells seems to depend on the strain background: In contrast to EcN, adherence of commensal strain TOB1 to HT-29 cells was enhanced upon the loss of cellulose production (Monteiro et al. 2009; Wang et al. 2006). *E. coli* K-12 is well established as a harmless laboratory strain. Recent evidences, however, indicate that the typical non-invasive nature of this *E. coli* strain can be reversed under specific circumstances even in the absence of any major genomic flux. Introduction of a mutated histone-like protein HU into *E. coli* K-12 resulted in significant changes in nucleoid organization and global transcription. These changes transformed the mutant *E. coli* into an almost obligate intracellular bacterium. This result demonstrates that even without gross changes in its genome content, changes of the cellular transcription program can lead to widely divergent lifestyles of *E. coli* K-12 in relation to mammalian host cells (Koli et al. 2011).

4.2 Impact of Genome Plasticity on Pathogenicity and Fitness of E. coli B2 Strains: Three Closely Related Strains with Markedly Divergent Phenotypes

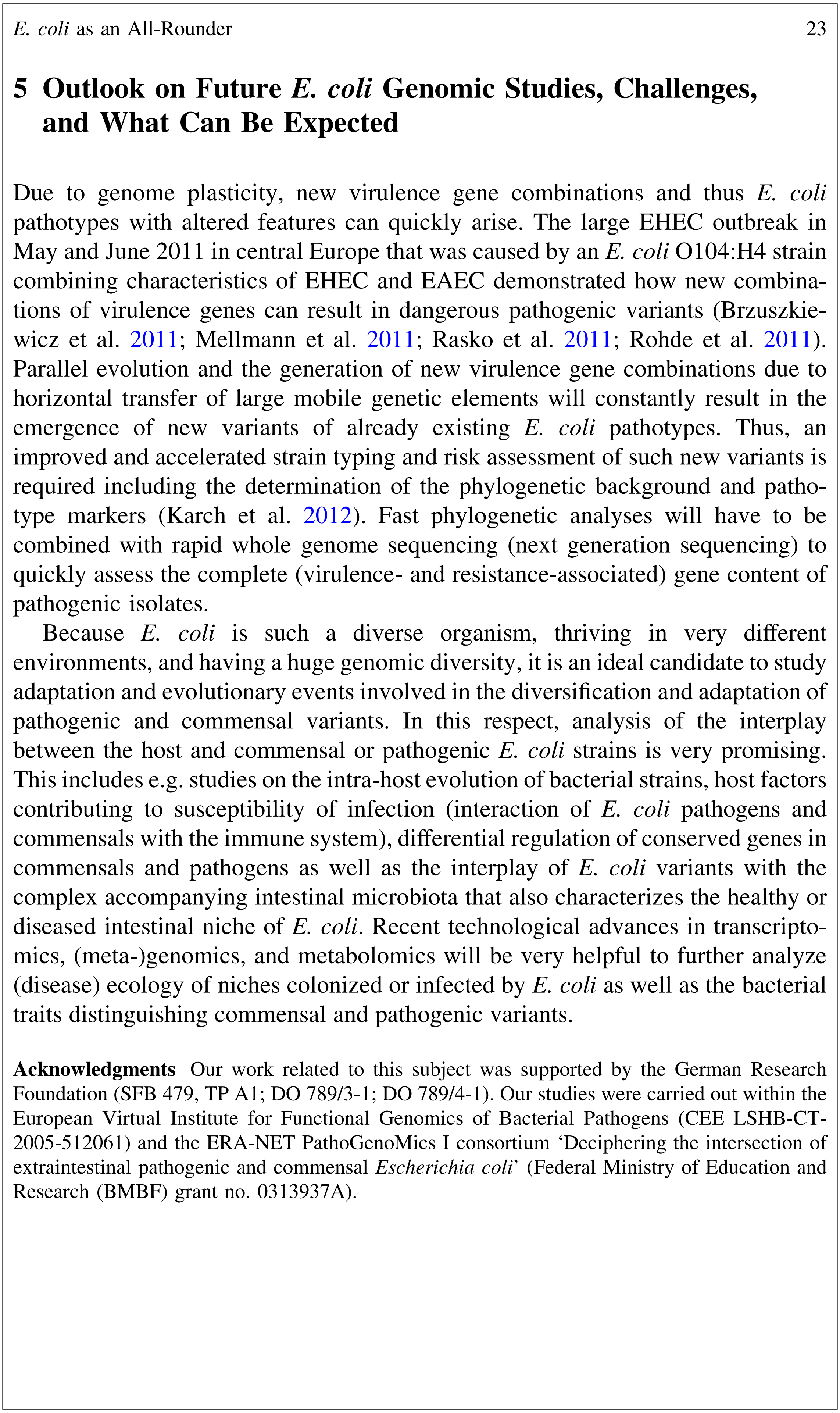
The comparison of three closely related *E. coli* sequence type ST73 isolates CFT073, 83972, and Nissle 1917 (EcN) exemplifies the difficulty to distinguish certain ExPEC and non-pathogenic *E. coli* variants. Strain CFT073 is a highly virulent archetypal uropathogenic isolate, whereas *E. coli* strains 83972 and Nissle 1917 are non-pathogenic strains derived from the urinary tract and the fecal flora, respectively. The three strains belong to the same clonal group (ST73) and are thus phylogenetically very closely related, despite their different environmental origins and disease-causing ability (Grozdanov et al. 2004; Zdziarski et al. 2008).

UPEC strain CFT073 has been isolated from the blood and urine of a woman with acute pyelonephritis and is widely used a model organism to study UPEC pathogenicity. CFT073 expresses a multitude of virulence genes which promote successful colonization and infection of the urinary tract, including several determinants coding for e.g. adhesins, toxins, iron uptake systems, proteases, flagella, and capsule (Welch et al. 2002; Gunther et al. 2002; Lloyd et al. 2009). The probiotic *E. coli* strain Nissle 1917 shows large overall genome content similarity with *E. coli* CFT073. Nevertheless, EcN lacks functional important virulence gene clusters, such as P-fimbrial and α-hemolysin determinants. The absence of a long-chain O-antigen due to a point mutation in the O-antigen polymerase gene *wzy* renders EcN serum-sensitive (Grozdanov et al. 2002). Among others, these traits are held responsible for the non-pathogenic character of this isolate. Beyond that, EcN has been used as a safe and efficient probiotic strain against a variety of intestinal disorders in humans and animals (Kruis et al. 2012; Schultz 2008; von Buenau et al. 2005). EcN was reported to protect gnotobiotic piglets from infection with invasive bacterial pathogens. Furthermore, Nissle 1917 is a good biofilm former and its efficient adhesion to epithelial cells interferes in



vitro with the invasion of several bacterial pathogens. Besides its bactericidal activity against many bacterial pathogens (Storm et al. 2011), EcN has also been demonstrated to negatively affect adhesion of bacterial pathogens through secretory components. This seems to be a common mechanism of *E. coli* strains with strong adhesive capacity (Storm et al. 2011; Altenhoefer et al. 2004; Huebner et al. 2011; Schierack et al. 2011). Immunomodulatory, anti-inflammatory properties have been described for EcN as well (Adam et al. 2010; Güttsches et al. 2012; Ukena et al. 2005). The inducible antimicrobial peptide human β-defensin 2 (hBD-2) is synthesized by the epithelium to counteract bacterial adherence and invasion. Flagellin expressed by EcN induces hBD-2 expression and can thus promote host defenses against bacterial infection (Schlee et al. 2007). In addition, EcN is able to restore disrupted epithelial barriers and to increase its resistance to microbial pathogens (Stetinova et al. 2010; Ukena et al. 2007).

E. coli 83972 is an asymptomatic bacteriuria (ABU) isolate with the ability to colonize the human urinary bladder without inducing an immune response. Similar to EcN, mutations in the *E. coli* 83972 genes encoding type 1-, F1C- and P fimbriae as well as α-hemolysin result in the loss of its ability to express these important virulence-associated genes as a result of host-driven adaptation. This strain also exhibits a semi-rough phenotype indicative of the absence of a long-chain O-antigen. *E. coli* 83972 has beneficial features as it outcompetes UPEC isolates for growth in urine and thus has a selective advantage over UPEC. This can be exploited for preventative and/or therapeutic approaches based on bacterial interference (Sundén et al. 2010). Strain 83972 has been established as an example of bacterial adaptation from pathogenicity to commensalism through virulence factor loss. It is assumed that prolonged asymptomatic bladder colonization selects for such attenuated variants where VFs have been inactivated, e.g. by point mutations and small deletions (Salvador et al. 2012; Zdziarski et al. 2008, 2010; Sundén et al. 2010; Klemm et al. 2006; Roos et al. 2006). Comparative genomic analyses indicated relatively few variations in genome content between these three isolates, thus suggesting that genetic variations (e.g. mutations, rearrangements, deletions) and expression differences, rather than a markedly different genome content, contribute to the divergent phenotypes of these strains. Notably, the two *E. coli* strains Nissle 1917 and 83972 with beneficial traits are deconstructed, attenuated pathogens (Grozdanov et al. 2002, 2004; Zdziarski et al. 2008, 2010; Hancock 2010a, b; Vejborg et al. 2010). To date, it is unknown whether strain CFT073 is also able to express beneficial traits, which could generally contribute to fitness and competitiveness, and whether they are just superimposed by the functional VFs expressed. Alternatively, specific genomic features of EcN and ABU isolate 83972, which are absent in UPEC CFT073, may account for their specific interaction with human epithelial cells or other bacteria. It will be an interesting and important future task to further characterize beneficial traits of strains EcN and 83972, to identify the underlying molecular mechanisms, and correlate them with genomic and phenotypic differences between UPEC CFT073, probiotic EcN, and ABU isolate 83972.

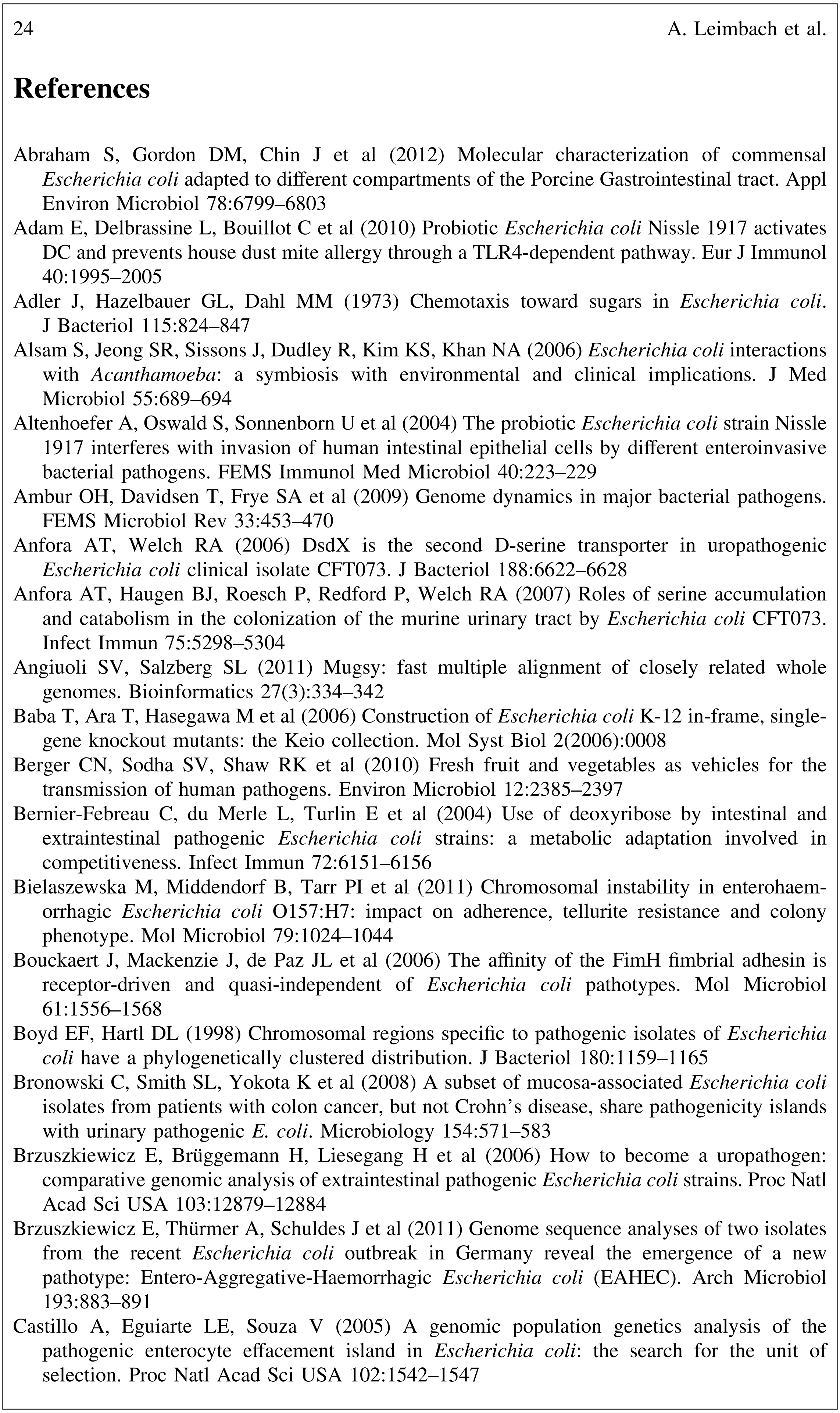


5 Outlook on Future *E. coli* Genomic Studies, Challenges, and What Can Be Expected

Due to genome plasticity, new virulence gene combinations and thus *E. coli* pathotypes with altered features can quickly arise. The large EHEC outbreak in May and June 2011 in central Europe that was caused by an *E. coli* O104:H4 strain combining characteristics of EHEC and EAEC demonstrated how new combinations of virulence genes can result in dangerous pathogenic variants (Brzuszkiewicz et al. 2011; Mellmann et al. 2011; Rasko et al. 2011; Rohde et al. 2011). Parallel evolution and the generation of new virulence gene combinations due to horizontal transfer of large mobile genetic elements will constantly result in the emergence of new variants of already existing *E. coli* pathotypes. Thus, an improved and accelerated strain typing and risk assessment of such new variants is required including the determination of the phylogenetic background and pathotype markers (Karch et al. 2012). Fast phylogenetic analyses will have to be combined with rapid whole genome sequencing (next generation sequencing) to quickly assess the complete (virulence- and resistance-associated) gene content of pathogenic isolates.

Because *E. coli* is such a diverse organism, thriving in very different environments, and having a huge genomic diversity, it is an ideal candidate to study adaptation and evolutionary events involved in the diversification and adaptation of pathogenic and commensal variants. In this respect, analysis of the interplay between the host and commensal or pathogenic *E. coli* strains is very promising. This includes e.g. studies on the intra-host evolution of bacterial strains, host factors contributing to susceptibility of infection (interaction of *E. coli* pathogens and commensals with the immune system), differential regulation of conserved genes in commensals and pathogens as well as the interplay of *E. coli* variants with the complex accompanying intestinal microbiota that also characterizes the healthy or diseased intestinal niche of *E. coli*. Recent technological advances in transcriptomics, (meta-)genomics, and metabolomics will be very helpful to further analyze (disease) ecology of niches colonized or infected by *E. coli* as well as the bacterial traits distinguishing commensal and pathogenic variants.

Acknowledgments Our work related to this subject was supported by the German Research Foundation (SFB 479, TP A1; DO 789/3-1; DO 789/4-1). Our studies were carried out within the European Virtual Institute for Functional Genomics of Bacterial Pathogens (CEE LSHB-CT-2005-512061) and the ERA-NET PathoGenoMics I consortium ‘Deciphering the intersection of extraintestinal pathogenic and commensal *Escherichia coli*’ (Federal Ministry of Education and Research (BMBF) grant no. 0313937A).



References

Abraham S, Gordon DM, Chin J et al (2012) Molecular characterization of commensal *Escherichia coli* adapted to different compartments of the Porcine Gastrointestinal tract. Appl Environ Microbiol 78:6799–6803

Adam E, Delbrassine L, Bouillot C et al (2010) Probiotic *Escherichia coli* Nissle 1917 activates DC and prevents house dust mite allergy through a TLR4-dependent pathway. Eur J Immunol 40:1995–2005

Adler J, Hazelbauer GL, Dahl MM (1973) Chemotaxis toward sugars in *Escherichia coli*. J Bacteriol 115:824–847

Alsam S, Jeong SR, Sissons J, Dudley R, Kim KS, Khan NA (2006) *Escherichia coli* interactions with *Acanthamoeba*: a symbiosis with environmental and clinical implications. J Med Microbiol 55:689–694

Altenhoefer A, Oswald S, Sonnenborn U et al (2004) The probiotic *Escherichia coli* strain Nissle 1917 interferes with invasion of human intestinal epithelial cells by different enteroinvasive bacterial pathogens. FEMS Immunol Med Microbiol 40:223–229

Ambur OH, Davidsen T, Frye SA et al (2009) Genome dynamics in major bacterial pathogens. FEMS Microbiol Rev 33:453–470

Anfora AT, Welch RA (2006) DsdX is the second D-serine transporter in uropathogenic *Escherichia coli* clinical isolate CFT073. J Bacteriol 188:6622–6628

Anfora AT, Haugen BJ, Roesch P, Redford P, Welch RA (2007) Roles of serine accumulation and catabolism in the colonization of the murine urinary tract by *Escherichia coli* CFT073. Infect Immun 75:5298–5304

Angiuoli SV, Salzberg SL (2011) Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics 27(3):334–342

Baba T, Ara T, Hasegawa M et al (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol 2(2006):0008

Berger CN, Sodha SV, Shaw RK et al (2010) Fresh fruit and vegetables as vehicles for the transmission of human pathogens. Environ Microbiol 12:2385–2397

Bernier-Febreau C, du Merle L, Turlin E et al (2004) Use of deoxyribose by intestinal and extraintestinal pathogenic *Escherichia coli* strains: a metabolic adaptation involved in competitiveness. Infect Immun 72:6151–6156

Bielaszewska M, Middendorf B, Tarr PI et al (2011) Chromosomal instability in enterohaemorrhagic *Escherichia coli* O157:H7: impact on adherence, tellurite resistance and colony phenotype. Mol Microbiol 79:1024–1044

Bouckaert J, Mackenzie J, de Paz JL et al (2006) The affinity of the FimH fimbrial adhesin is receptor-driven and quasi-independent of *Escherichia coli* pathotypes. Mol Microbiol 61:1556–1568

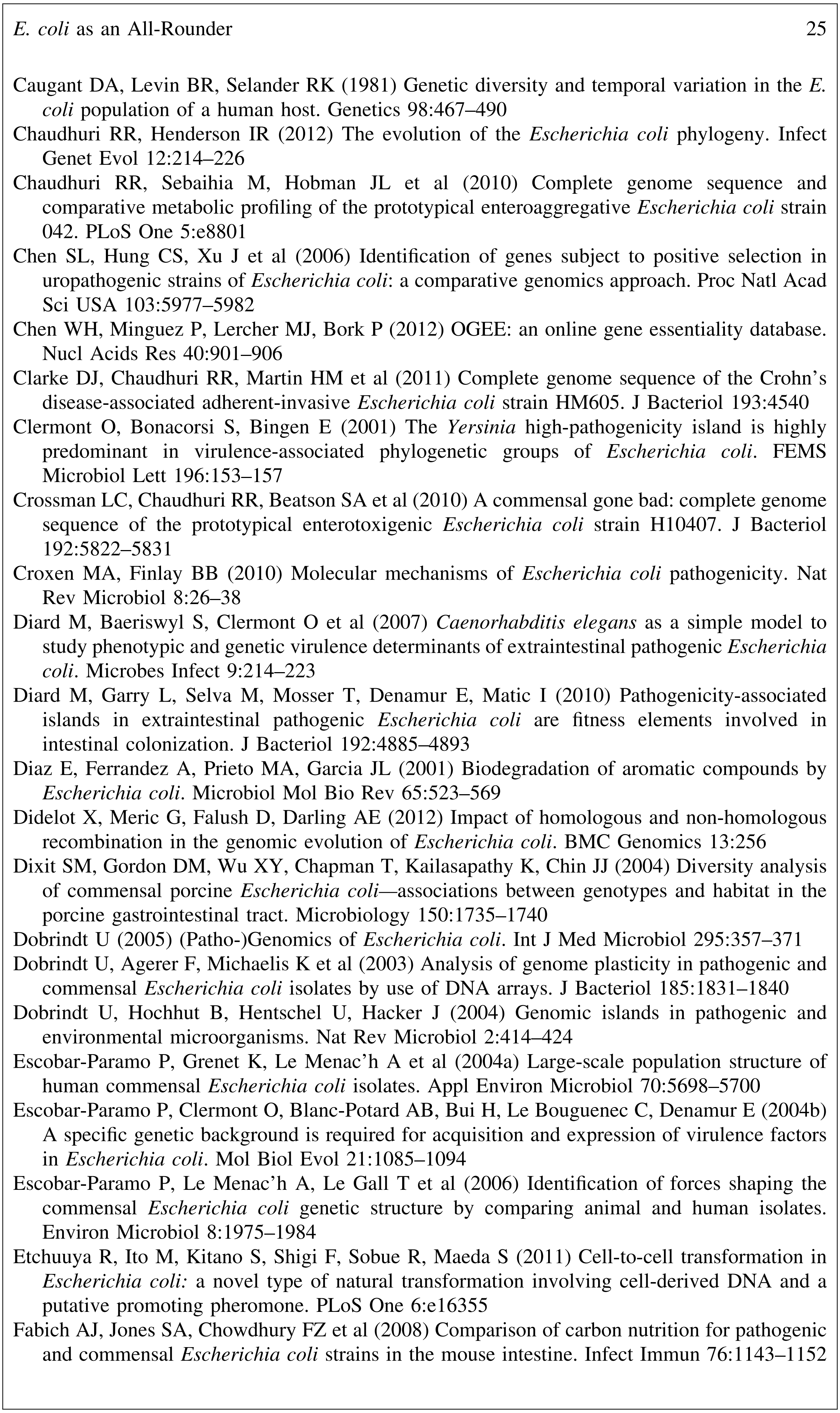
Boyd EF, Hartl DL (1998) Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J Bacteriol 180:1159–1165

Bronowski C, Smith SL, Yokota K et al (2008) A subset of mucosa-associated *Escherichia coli* isolates from patients with colon cancer, but not Crohn's disease, share pathogenicity islands with urinary pathogenic *E. coli*. Microbiology 154:571–583

Brzuszkiewicz E, Brüggemann H, Liesegang H et al (2006) How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. Proc Natl Acad Sci USA 103:12879–12884

Brzuszkiewicz E, Thürmer A, Schuldes J et al (2011) Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). Arch Microbiol 193:883–891

Castillo A, Eguiarte LE, Souza V (2005) A genomic population genetics analysis of the pathogenic enterocyte effacement island in *Escherichia coli*: the search for the unit of selection. Proc Natl Acad Sci USA 102:1542–1547



E. coli as an All-Rounder 25

Caugant DA, Levin BR, Selander RK (1981) Genetic diversity and temporal variation in the *E. coli* population of a human host. Genetics 98:467–490

Chaudhuri RR, Henderson IR (2012) The evolution of the *Escherichia coli* phylogeny. Infect Genet Evol 12:214–226

Chaudhuri RR, Sebaihia M, Hobman JL et al (2010) Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. PLoS One 5:e8801

Chen SL, Hung CS, Xu J et al (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. Proc Natl Acad Sci USA 103:5977–5982

Chen WH, Minguez P, Lercher MJ, Bork P (2012) OGEE: an online gene essentiality database. Nucl Acids Res 40:901–906

Clarke DJ, Chaudhuri RR, Martin HM et al (2011) Complete genome sequence of the Crohn's disease-associated adherent-invasive *Escherichia coli* strain HM605. J Bacteriol 193:4540

Clermont O, Bonacorsi S, Bingen E (2001) The *Yersinia* high-pathogenicity island is highly predominant in virulence-associated phylogenetic groups of *Escherichia coli*. FEMS Microbiol Lett 196:153–157

Crossman LC, Chaudhuri RR, Beatson SA et al (2010) A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407. J Bacteriol 192:5822–5831

Croxen MA, Finlay BB (2010) Molecular mechanisms of *Escherichia coli* pathogenicity. Nat Rev Microbiol 8:26–38

Diard M, Baeriswyl S, Clermont O et al (2007) *Caenorhabditis elegans* as a simple model to study phenotypic and genetic virulence determinants of extraintestinal pathogenic *Escherichia coli*. Microbes Infect 9:214–223

Diard M, Garry L, Selva M, Mosser T, Denamur E, Matic I (2010) Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. J Bacteriol 192:4885–4893

Diaz E, Ferrandez A, Prieto MA, Garcia JL (2001) Biodegradation of aromatic compounds by *Escherichia coli*. Microbiol Mol Bio Rev 65:523–569

Didelot X, Meric G, Falush D, Darling AE (2012) Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13:256

Dixit SM, Gordon DM, Wu XY, Chapman T, Kailasapathy K, Chin JJ (2004) Diversity analysis of commensal porcine *Escherichia coli*—associations between genotypes and habitat in the porcine gastrointestinal tract. Microbiology 150:1735–1740

Dobrindt U (2005) (Patho-)Genomics of *Escherichia coli*. Int J Med Microbiol 295:357–371

Dobrindt U, Agerer F, Michaelis K et al (2003) Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. J Bacteriol 185:1831–1840

Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2:414–424

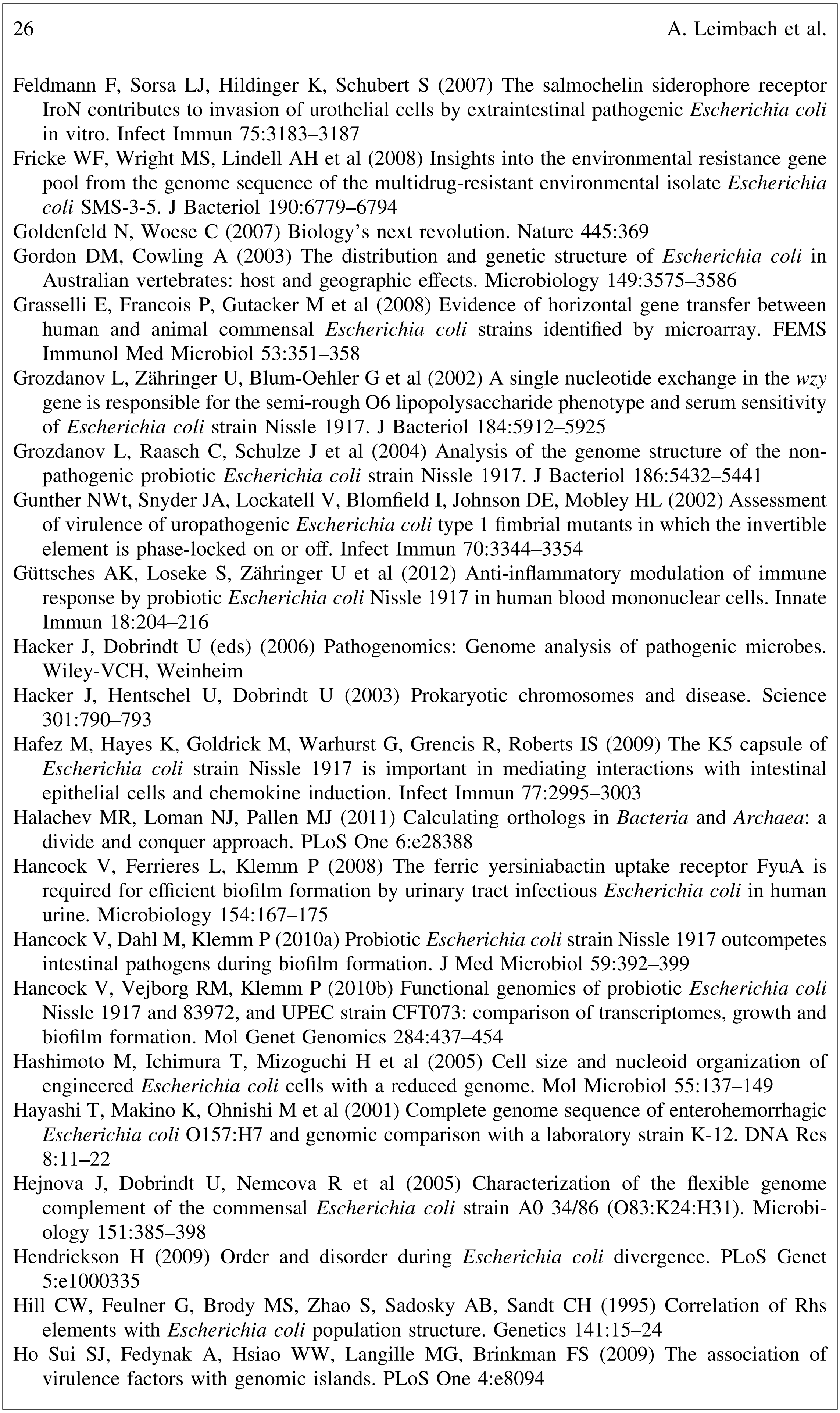
Escobar-Paramo P, Grenet K, Le Menac'h A et al (2004a) Large-scale population structure of human commensal *Escherichia coli* isolates. Appl Environ Microbiol 70:5698–5700

Escobar-Paramo P, Clermont O, Blanc-Potard AB, Bui H, Le Bouguenec C, Denamur E (2004b) A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. Mol Biol Evol 21:1085–1094

Escobar-Paramo P, Le Menac'h A, Le Gall T et al (2006) Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. Environ Microbiol 8:1975–1984

Etchuuya R, Ito M, Kitano S, Shigi F, Sobue R, Maeda S (2011) Cell-to-cell transformation in *Escherichia coli:* a novel type of natural transformation involving cell-derived DNA and a putative promoting pheromone. PLoS One 6:e16355

Fabich AJ, Jones SA, Chowdhury FZ et al (2008) Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. Infect Immun 76:1143–1152



Feldmann F, Sorsa LJ, Hildinger K, Schubert S (2007) The salmochelin siderophore receptor IroN contributes to invasion of urothelial cells by extraintestinal pathogenic *Escherichia coli* in vitro. Infect Immun 75:3183–3187

Fricke WF, Wright MS, Lindell AH et al (2008) Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5. J Bacteriol 190:6779–6794

Goldenfeld N, Woese C (2007) Biology's next revolution. Nature 445:369

Gordon DM, Cowling A (2003) The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. Microbiology 149:3575–3586

Grasselli E, Francois P, Gutacker M et al (2008) Evidence of horizontal gene transfer between human and animal commensal *Escherichia coli* strains identified by microarray. FEMS Immunol Med Microbiol 53:351–358

Grozdanov L, Zähringer U, Blum-Oehler G et al (2002) A single nucleotide exchange in the *wzy* gene is responsible for the semi-rough O6 lipopolysaccharide phenotype and serum sensitivity of *Escherichia coli* strain Nissle 1917. J Bacteriol 184:5912–5925

Grozdanov L, Raasch C, Schulze J et al (2004) Analysis of the genome structure of the non-pathogenic probiotic *Escherichia coli* strain Nissle 1917. J Bacteriol 186:5432–5441

Gunther NWt, Snyder JA, Lockatell V, Blomfield I, Johnson DE, Mobley HL (2002) Assessment of virulence of uropathogenic *Escherichia coli* type 1 fimbrial mutants in which the invertible element is phase-locked on or off. Infect Immun 70:3344–3354

Güttsches AK, Loseke S, Zähringer U et al (2012) Anti-inflammatory modulation of immune response by probiotic *Escherichia coli* Nissle 1917 in human blood mononuclear cells. Innate Immun 18:204–216

Hacker J, Dobrindt U (eds) (2006) Pathogenomics: Genome analysis of pathogenic microbes. Wiley-VCH, Weinheim

Hacker J, Hentschel U, Dobrindt U (2003) Prokaryotic chromosomes and disease. Science 301:790–793

Hafez M, Hayes K, Goldrick M, Warhurst G, Grencis R, Roberts IS (2009) The K5 capsule of *Escherichia coli* strain Nissle 1917 is important in mediating interactions with intestinal epithelial cells and chemokine induction. Infect Immun 77:2995–3003

Halachev MR, Loman NJ, Pallen MJ (2011) Calculating orthologs in *Bacteria* and *Archaea*: a divide and conquer approach. PLoS One 6:e28388

Hancock V, Ferrieres L, Klemm P (2008) The ferric yersiniabactin uptake receptor FyuA is required for efficient biofilm formation by urinary tract infectious *Escherichia coli* in human urine. Microbiology 154:167–175

Hancock V, Dahl M, Klemm P (2010a) Probiotic *Escherichia coli* strain Nissle 1917 outcompetes intestinal pathogens during biofilm formation. J Med Microbiol 59:392–399

Hancock V, Vejborg RM, Klemm P (2010b) Functional genomics of probiotic *Escherichia coli* Nissle 1917 and 83972, and UPEC strain CFT073: comparison of transcriptomes, growth and biofilm formation. Mol Genet Genomics 284:437–454

Hashimoto M, Ichimura T, Mizoguchi H et al (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. Mol Microbiol 55:137–149

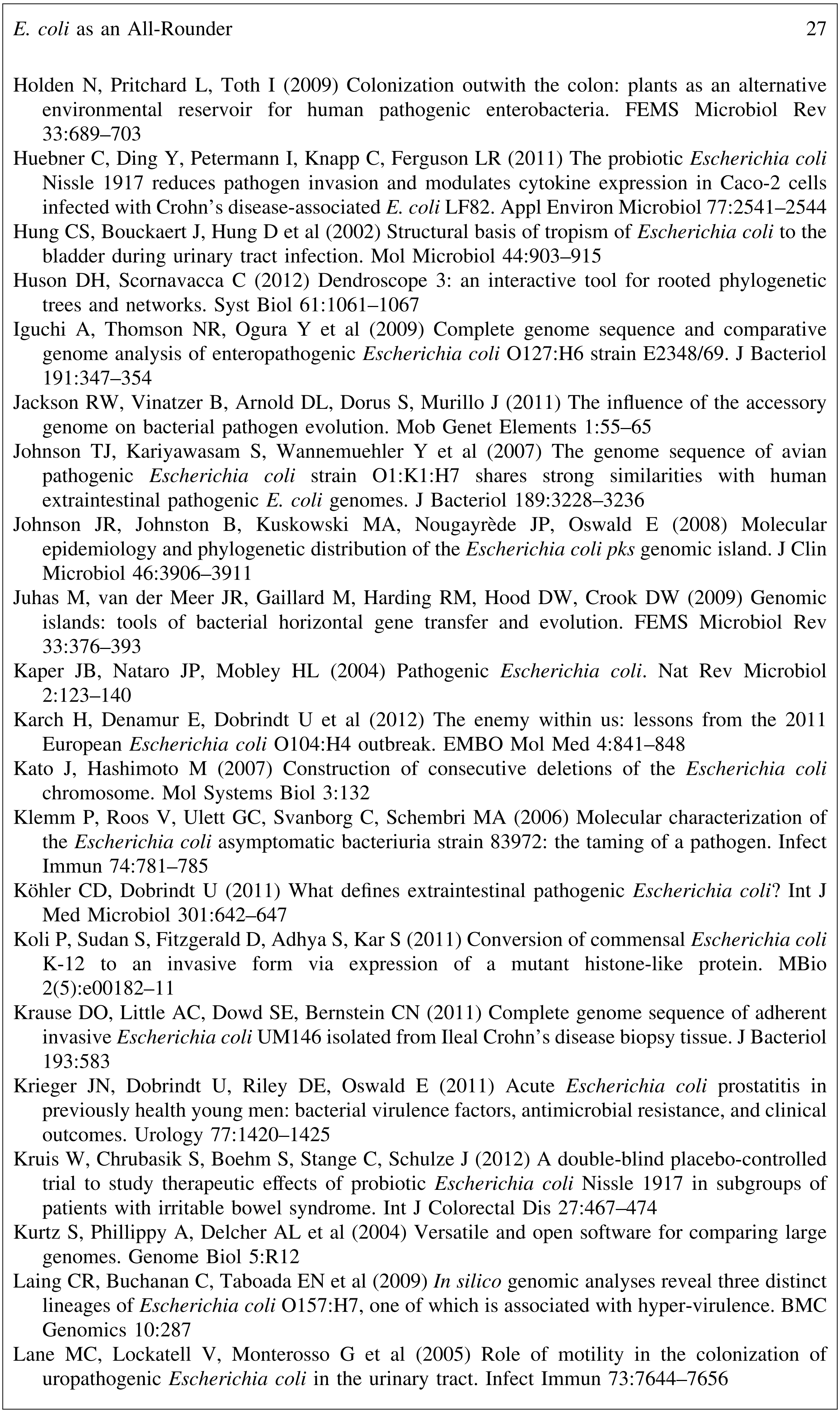
Hayashi T, Makino K, Ohnishi M et al (2001) Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11–22

Hejnova J, Dobrindt U, Nemcova R et al (2005) Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83:K24:H31). Microbiology 151:385–398

Hendrickson H (2009) Order and disorder during *Escherichia coli* divergence. PLoS Genet 5:e1000335

Hill CW, Feulner G, Brody MS, Zhao S, Sadosky AB, Sandt CH (1995) Correlation of Rhs elements with *Escherichia coli* population structure. Genetics 141:15–24

Ho Sui SJ, Fedynak A, Hsiao WW, Langille MG, Brinkman FS (2009) The association of virulence factors with genomic islands. PLoS One 4:e8094



Holden N, Pritchard L, Toth I (2009) Colonization outwith the colon: plants as an alternative environmental reservoir for human pathogenic enterobacteria. FEMS Microbiol Rev 33:689–703

Huebner C, Ding Y, Petermann I, Knapp C, Ferguson LR (2011) The probiotic *Escherichia coli* Nissle 1917 reduces pathogen invasion and modulates cytokine expression in Caco-2 cells infected with Crohn's disease-associated *E. coli* LF82. Appl Environ Microbiol 77:2541–2544

Hung CS, Bouckaert J, Hung D et al (2002) Structural basis of tropism of *Escherichia coli* to the bladder during urinary tract infection. Mol Microbiol 44:903–915

Huson DH, Scornavacca C (2012) Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067

Iguchi A, Thomson NR, Ogura Y et al (2009) Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. J Bacteriol 191:347–354

Jackson RW, Vinatzer B, Arnold DL, Dorus S, Murillo J (2011) The influence of the accessory genome on bacterial pathogen evolution. Mob Genet Elements 1:55–65

Johnson TJ, Kariyawasam S, Wannemuehler Y et al (2007) The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes. J Bacteriol 189:3228–3236

Johnson JR, Johnston B, Kuskowski MA, Nougayrède JP, Oswald E (2008) Molecular epidemiology and phylogenetic distribution of the *Escherichia coli pks* genomic island. J Clin Microbiol 46:3906–3911

Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol Rev 33:376–393

Kaper JB, Nataro JP, Mobley HL (2004) Pathogenic *Escherichia coli*. Nat Rev Microbiol 2:123–140

Karch H, Denamur E, Dobrindt U et al (2012) The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. EMBO Mol Med 4:841–848

Kato J, Hashimoto M (2007) Construction of consecutive deletions of the *Escherichia coli* chromosome. Mol Systems Biol 3:132

Klemm P, Roos V, Ulett GC, Svanborg C, Schembri MA (2006) Molecular characterization of the *Escherichia coli* asymptomatic bacteriuria strain 83972: the taming of a pathogen. Infect Immun 74:781–785

Köhler CD, Dobrindt U (2011) What defines extraintestinal pathogenic *Escherichia coli*? Int J Med Microbiol 301:642–647

Koli P, Sudan S, Fitzgerald D, Adhya S, Kar S (2011) Conversion of commensal *Escherichia coli* K-12 to an invasive form via expression of a mutant histone-like protein. MBio 2(5):e00182–11

Krause DO, Little AC, Dowd SE, Bernstein CN (2011) Complete genome sequence of adherent invasive *Escherichia coli* UM146 isolated from Ileal Crohn's disease biopsy tissue. J Bacteriol 193:583

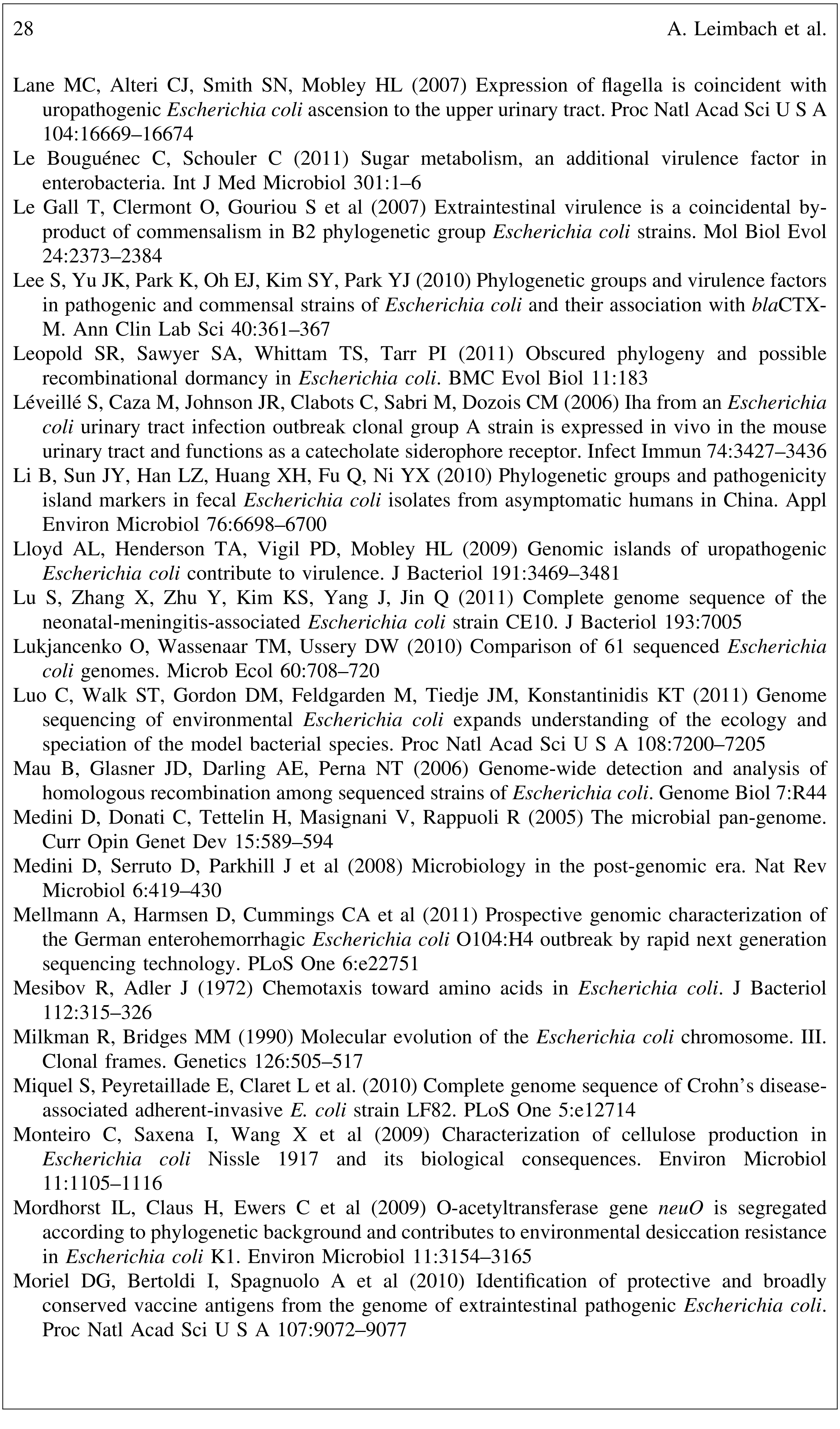
Krieger JN, Dobrindt U, Riley DE, Oswald E (2011) Acute *Escherichia coli* prostatitis in previously health young men: bacterial virulence factors, antimicrobial resistance, and clinical outcomes. Urology 77:1420–1425

Kruis W, Chrubasik S, Boehm S, Stange C, Schulze J (2012) A double-blind placebo-controlled trial to study therapeutic effects of probiotic *Escherichia coli* Nissle 1917 in subgroups of patients with irritable bowel syndrome. Int J Colorectal Dis 27:467–474

Kurtz S, Phillippy A, Delcher AL et al (2004) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Laing CR, Buchanan C, Taboada EN et al (2009) *In silico* genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence. BMC Genomics 10:287

Lane MC, Lockatell V, Monterosso G et al (2005) Role of motility in the colonization of uropathogenic *Escherichia coli* in the urinary tract. Infect Immun 73:7644–7656



Lane MC, Alteri CJ, Smith SN, Mobley HL (2007) Expression of flagella is coincident with uropathogenic *Escherichia coli* ascension to the upper urinary tract. Proc Natl Acad Sci U S A 104:16669–16674

Le Bouguénec C, Schouler C (2011) Sugar metabolism, an additional virulence factor in enterobacteria. Int J Med Microbiol 301:1–6

Le Gall T, Clermont O, Gouriou S et al (2007) Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol Biol Evol 24:2373–2384

Lee S, Yu JK, Park K, Oh EJ, Kim SY, Park YJ (2010) Phylogenetic groups and virulence factors in pathogenic and commensal strains of *Escherichia coli* and their association with *bla*CTX-M. Ann Clin Lab Sci 40:361–367

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

Léveillé S, Caza M, Johnson JR, Clabots C, Sabri M, Dozois CM (2006) Iha from an *Escherichia coli* urinary tract infection outbreak clonal group A strain is expressed in vivo in the mouse urinary tract and functions as a catecholate siderophore receptor. Infect Immun 74:3427–3436

Li B, Sun JY, Han LZ, Huang XH, Fu Q, Ni YX (2010) Phylogenetic groups and pathogenicity island markers in fecal *Escherichia coli* isolates from asymptomatic humans in China. Appl Environ Microbiol 76:6698–6700

Lloyd AL, Henderson TA, Vigil PD, Mobley HL (2009) Genomic islands of uropathogenic *Escherichia coli* contribute to virulence. J Bacteriol 191:3469–3481

Lu S, Zhang X, Zhu Y, Kim KS, Yang J, Jin Q (2011) Complete genome sequence of the neonatal-meningitis-associated *Escherichia coli* strain CE10. J Bacteriol 193:7005

Lukjancenko O, Wassenaar TM, Ussery DW (2010) Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 60:708–720

Luo C, Walk ST, Gordon DM, Feldgarden M, Tiedje JM, Konstantinidis KT (2011) Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. Proc Natl Acad Sci U S A 108:7200–7205

Mau B, Glasner JD, Darling AE, Perna NT (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. Genome Biol 7:R44

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. Curr Opin Genet Dev 15:589–594

Medini D, Serruto D, Parkhill J et al (2008) Microbiology in the post-genomic era. Nat Rev Microbiol 6:419–430

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

Mesibov R, Adler J (1972) Chemotaxis toward amino acids in *Escherichia coli*. J Bacteriol 112:315–326

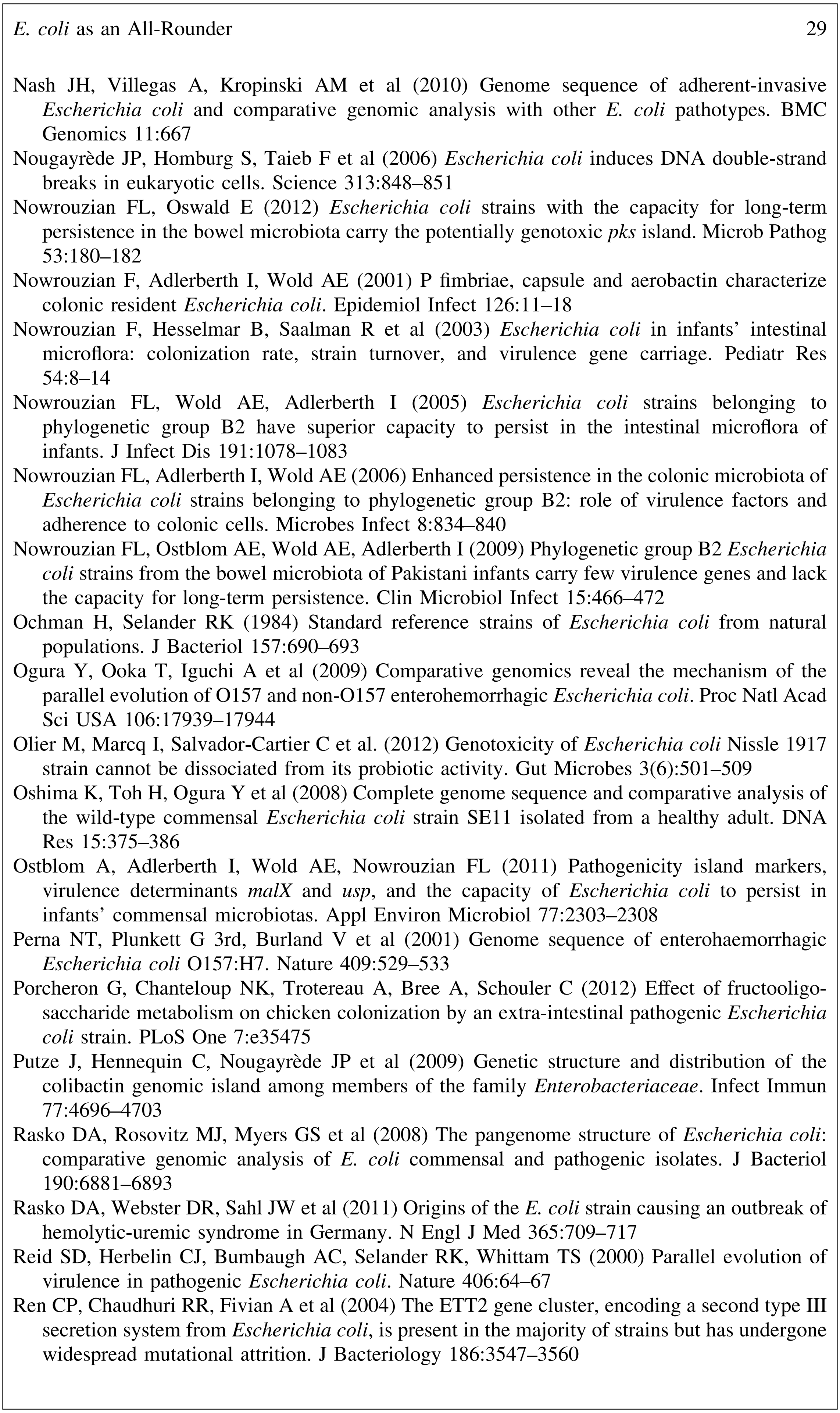
Milkman R, Bridges MM (1990) Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. Genetics 126:505–517

Miquel S, Peyretaillade E, Claret L et al. (2010) Complete genome sequence of Crohn's disease-associated adherent-invasive *E. coli* strain LF82. PLoS One 5:e12714

Monteiro C, Saxena I, Wang X et al (2009) Characterization of cellulose production in *Escherichia coli* Nissle 1917 and its biological consequences. Environ Microbiol 11:1105–1116

Mordhorst IL, Claus H, Ewers C et al (2009) O-acetyltransferase gene *neuO* is segregated according to phylogenetic background and contributes to environmental desiccation resistance in *Escherichia coli* K1. Environ Microbiol 11:3154–3165

Moriel DG, Bertoldi I, Spagnuolo A et al (2010) Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*. Proc Natl Acad Sci U S A 107:9072–9077



E. coli as an All-Rounder 29

Nash JH, Villegas A, Kropinski AM et al (2010) Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes. BMC Genomics 11:667

Nougayrède JP, Homburg S, Taieb F et al (2006) *Escherichia coli* induces DNA double-strand breaks in eukaryotic cells. Science 313:848–851

Nowrouzian FL, Oswald E (2012) *Escherichia coli* strains with the capacity for long-term persistence in the bowel microbiota carry the potentially genotoxic *pks* island. Microb Pathog 53:180–182

Nowrouzian F, Adlerberth I, Wold AE (2001) P fimbriae, capsule and aerobactin characterize colonic resident *Escherichia coli*. Epidemiol Infect 126:11–18

Nowrouzian F, Hesselmar B, Saalman R et al (2003) *Escherichia coli* in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage. Pediatr Res 54:8–14

Nowrouzian FL, Wold AE, Adlerberth I (2005) *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. J Infect Dis 191:1078–1083

Nowrouzian FL, Adlerberth I, Wold AE (2006) Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. Microbes Infect 8:834–840

Nowrouzian FL, Ostblom AE, Wold AE, Adlerberth I (2009) Phylogenetic group B2 *Escherichia coli* strains from the bowel microbiota of Pakistani infants carry few virulence genes and lack the capacity for long-term persistence. Clin Microbiol Infect 15:466–472

Ochman H, Selander RK (1984) Standard reference strains of *Escherichia coli* from natural populations. J Bacteriol 157:690–693

Ogura Y, Ooka T, Iguchi A et al (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. Proc Natl Acad Sci USA 106:17939–17944

Olier M, Marcq I, Salvador-Cartier C et al. (2012) Genotoxicity of *Escherichia coli* Nissle 1917 strain cannot be dissociated from its probiotic activity. Gut Microbes 3(6):501–509

Oshima K, Toh H, Ogura Y et al (2008) Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. DNA Res 15:375–386

Ostblom A, Adlerberth I, Wold AE, Nowrouzian FL (2011) Pathogenicity island markers, virulence determinants *malX* and *usp*, and the capacity of *Escherichia coli* to persist in infants' commensal microbiotas. Appl Environ Microbiol 77:2303–2308

Perna NT, Plunkett G 3rd, Burland V et al (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–533

Porcheron G, Chanteloup NK, Trotereau A, Bree A, Schouler C (2012) Effect of fructooligosaccharide metabolism on chicken colonization by an extra-intestinal pathogenic *Escherichia coli* strain. PLoS One 7:e35475

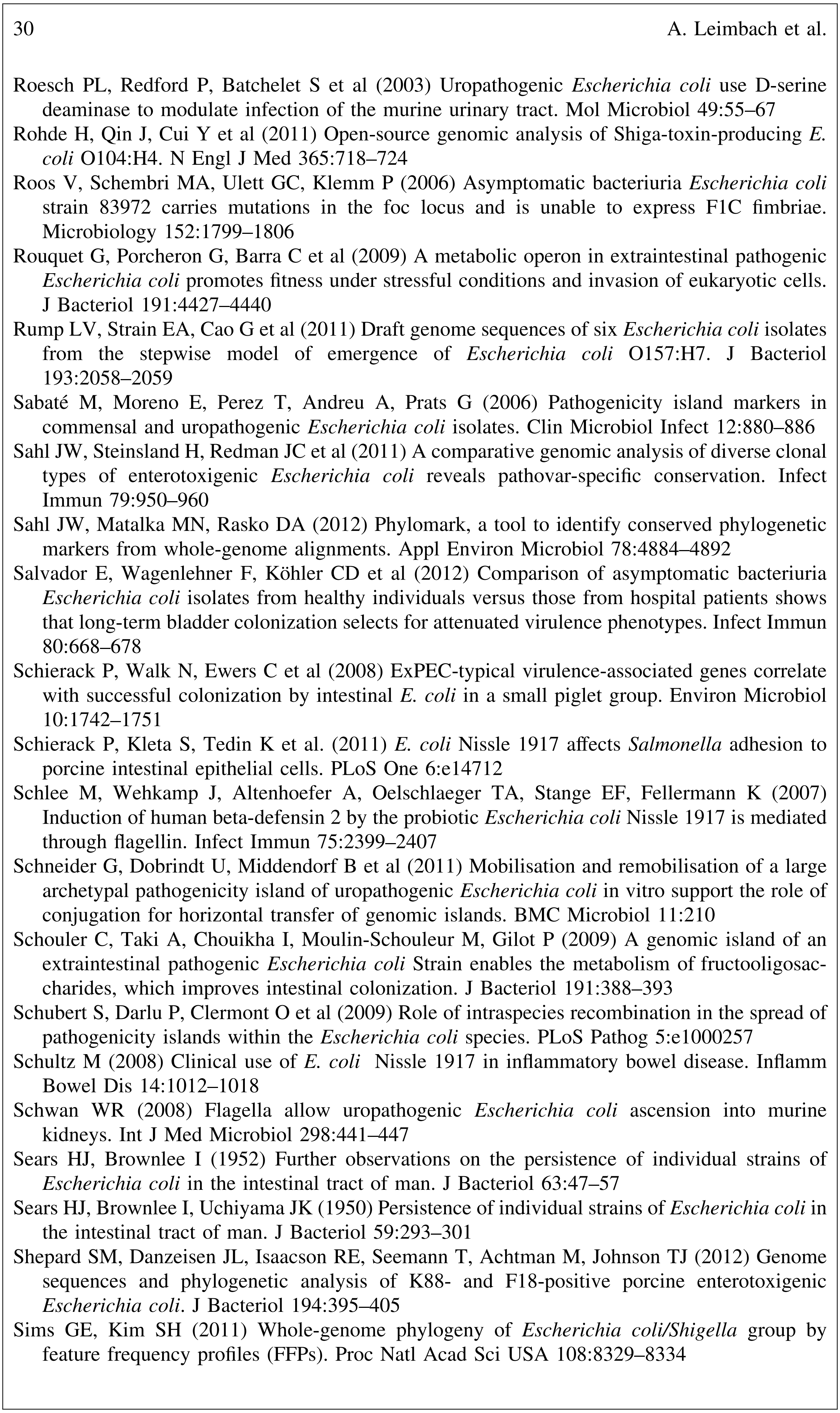
Putze J, Hennequin C, Nougayrède JP et al (2009) Genetic structure and distribution of the colibactin genomic island among members of the family *Enterobacteriaceae*. Infect Immun 77:4696–4703

Rasko DA, Rosovitz MJ, Myers GS et al (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol 190:6881–6893

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

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406:64–67

Ren CP, Chaudhuri RR, Fivian A et al (2004) The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition. J Bacteriology 186:3547–3560



Roesch PL, Redford P, Batchelet S et al (2003) Uropathogenic *Escherichia coli* use D-serine deaminase to modulate infection of the murine urinary tract. Mol Microbiol 49:55–67

Rohde H, Qin J, Cui Y et al (2011) Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4. N Engl J Med 365:718–724

Roos V, Schembri MA, Ulett GC, Klemm P (2006) Asymptomatic bacteriuria *Escherichia coli* strain 83972 carries mutations in the foc locus and is unable to express F1C fimbriae. Microbiology 152:1799–1806

Rouquet G, Porcheron G, Barra C et al (2009) A metabolic operon in extraintestinal pathogenic *Escherichia coli* promotes fitness under stressful conditions and invasion of eukaryotic cells. J Bacteriol 191:4427–4440

Rump LV, Strain EA, Cao G et al (2011) Draft genome sequences of six *Escherichia coli* isolates from the stepwise model of emergence of *Escherichia coli* O157:H7. J Bacteriol 193:2058–2059

Sabaté M, Moreno E, Perez T, Andreu A, Prats G (2006) Pathogenicity island markers in commensal and uropathogenic *Escherichia coli* isolates. Clin Microbiol Infect 12:880–886

Sahl JW, Steinsland H, Redman JC et al (2011) A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation. Infect Immun 79:950–960

Sahl JW, Matalka MN, Rasko DA (2012) Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl Environ Microbiol 78:4884–4892

Salvador E, Wagenlehner F, Köhler CD et al (2012) Comparison of asymptomatic bacteriuria *Escherichia coli* isolates from healthy individuals versus those from hospital patients shows that long-term bladder colonization selects for attenuated virulence phenotypes. Infect Immun 80:668–678

Schierack P, Walk N, Ewers C et al (2008) ExPEC-typical virulence-associated genes correlate with successful colonization by intestinal *E. coli* in a small piglet group. Environ Microbiol 10:1742–1751

Schierack P, Kleta S, Tedin K et al. (2011) *E. coli* Nissle 1917 affects *Salmonella* adhesion to porcine intestinal epithelial cells. PLoS One 6:e14712

Schlee M, Wehkamp J, Altenhoefer A, Oelschlaeger TA, Stange EF, Fellermann K (2007) Induction of human beta-defensin 2 by the probiotic *Escherichia coli* Nissle 1917 is mediated through flagellin. Infect Immun 75:2399–2407

Schneider G, Dobrindt U, Middendorf B et al (2011) Mobilisation and remobilisation of a large archetypal pathogenicity island of uropathogenic *Escherichia coli* in vitro support the role of conjugation for horizontal transfer of genomic islands. BMC Microbiol 11:210

Schouler C, Taki A, Chouikha I, Moulin-Schouleur M, Gilot P (2009) A genomic island of an extraintestinal pathogenic *Escherichia coli* Strain enables the metabolism of fructooligosaccharides, which improves intestinal colonization. J Bacteriol 191:388–393

Schubert S, Darlu P, Clermont O et al (2009) Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. PLoS Pathog 5:e1000257

Schultz M (2008) Clinical use of *E. coli* Nissle 1917 in inflammatory bowel disease. Inflamm Bowel Dis 14:1012–1018

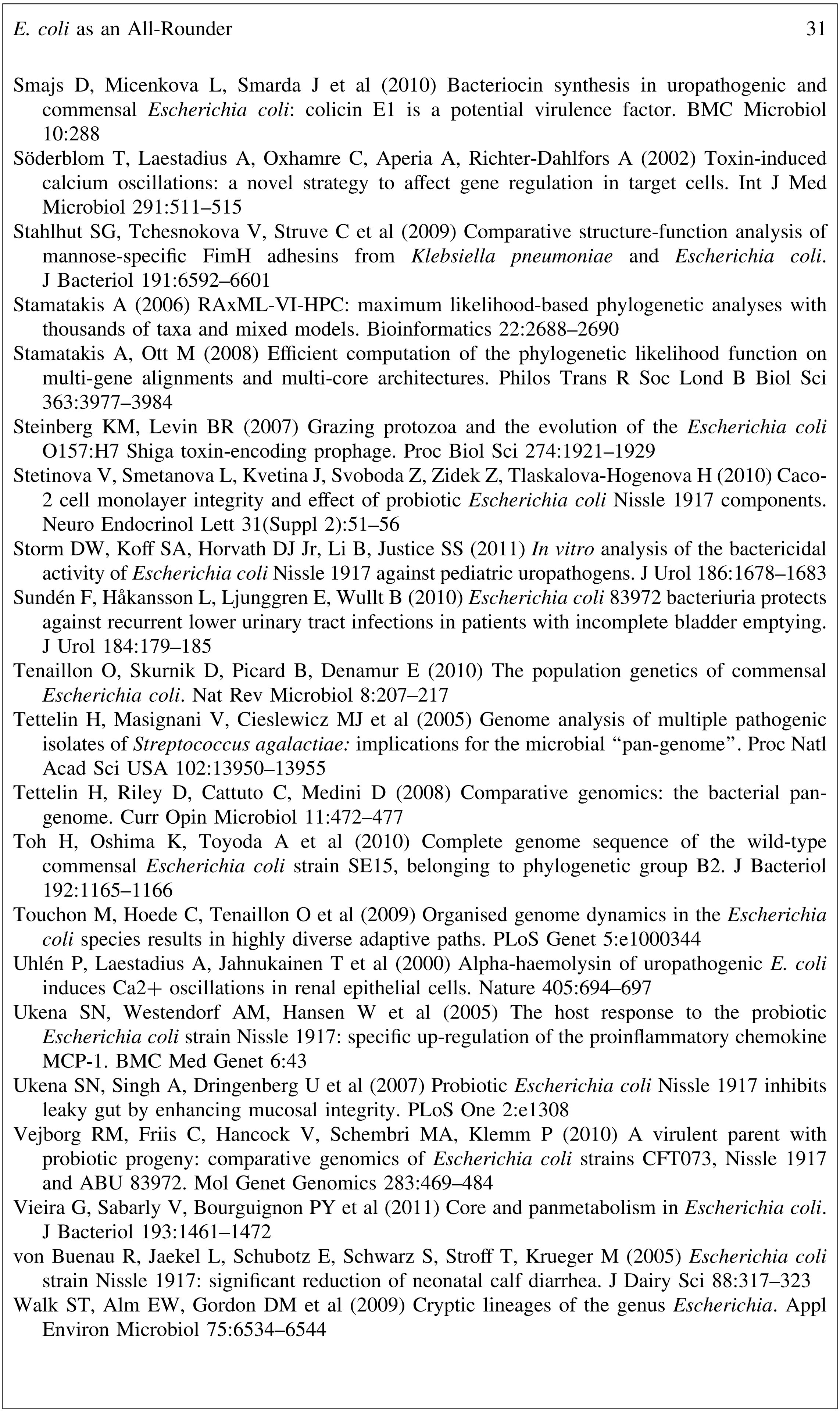
Schwan WR (2008) Flagella allow uropathogenic *Escherichia coli* ascension into murine kidneys. Int J Med Microbiol 298:441–447

Sears HJ, Brownlee I (1952) Further observations on the persistence of individual strains of *Escherichia coli* in the intestinal tract of man. J Bacteriol 63:47–57

Sears HJ, Brownlee I, Uchiyama JK (1950) Persistence of individual strains of *Escherichia coli* in the intestinal tract of man. J Bacteriol 59:293–301

Shepard SM, Danzeisen JL, Isaacson RE, Seemann T, Achtman M, Johnson TJ (2012) Genome sequences and phylogenetic analysis of K88- and F18-positive porcine enterotoxigenic *Escherichia coli*. J Bacteriol 194:395–405

Sims GE, Kim SH (2011) Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). Proc Natl Acad Sci USA 108:8329–8334



E. coli as an All-Rounder 31

Smajs D, Micenkova L, Smarda J et al (2010) Bacteriocin synthesis in uropathogenic and commensal *Escherichia coli*: colicin E1 is a potential virulence factor. BMC Microbiol 10:288

Söderblom T, Laestadius A, Oxhamre C, Aperia A, Richter-Dahlfors A (2002) Toxin-induced calcium oscillations: a novel strategy to affect gene regulation in target cells. Int J Med Microbiol 291:511–515

Stahlhut SG, Tchesnokova V, Struve C et al (2009) Comparative structure-function analysis of mannose-specific FimH adhesins from *Klebsiella pneumoniae* and *Escherichia coli*. J Bacteriol 191:6592–6601

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Stamatakis A, Ott M (2008) Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos Trans R Soc Lond B Biol Sci 363:3977–3984

Steinberg KM, Levin BR (2007) Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. Proc Biol Sci 274:1921–1929

Stetinova V, Smetanova L, Kvetina J, Svoboda Z, Zidek Z, Tlaskalova-Hogenova H (2010) Caco-2 cell monolayer integrity and effect of probiotic *Escherichia coli* Nissle 1917 components. Neuro Endocrinol Lett 31(Suppl 2):51–56

Storm DW, Koff SA, Horvath DJ Jr, Li B, Justice SS (2011) *In vitro* analysis of the bactericidal activity of *Escherichia coli* Nissle 1917 against pediatric uropathogens. J Urol 186:1678–1683

Sundén F, Håkansson L, Ljunggren E, Wullt B (2010) *Escherichia coli* 83972 bacteriuria protects against recurrent lower urinary tract infections in patients with incomplete bladder emptying. J Urol 184:179–185

Tenaillon O, Skurnik D, Picard B, Denamur E (2010) The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol 8:207–217

Tettelin H, Masignani V, Cieslewicz MJ et al (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae:* implications for the microbial "pan-genome". Proc Natl Acad Sci USA 102:13950–13955

Tettelin H, Riley D, Cattuto C, Medini D (2008) Comparative genomics: the bacterial pan-genome. Curr Opin Microbiol 11:472–477

Toh H, Oshima K, Toyoda A et al (2010) Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2. J Bacteriol 192:1165–1166

Touchon M, Hoede C, Tenaillon O et al (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344

Uhlén P, Laestadius A, Jahnukainen T et al (2000) Alpha-haemolysin of uropathogenic *E. coli* induces Ca2+ oscillations in renal epithelial cells. Nature 405:694–697

Ukena SN, Westendorf AM, Hansen W et al (2005) The host response to the probiotic *Escherichia coli* strain Nissle 1917: specific up-regulation of the proinflammatory chemokine MCP-1. BMC Med Genet 6:43

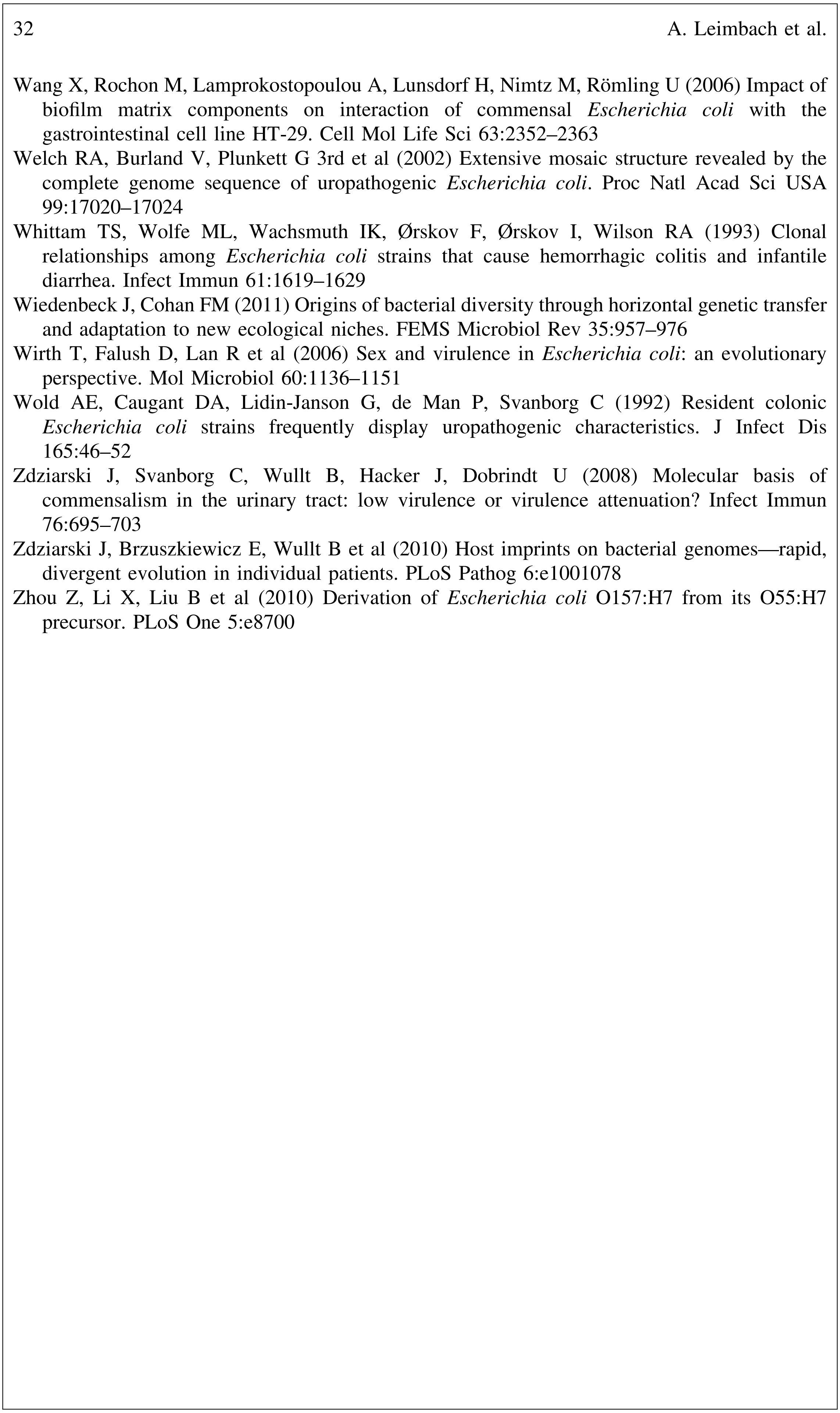
Ukena SN, Singh A, Dringenberg U et al (2007) Probiotic *Escherichia coli* Nissle 1917 inhibits leaky gut by enhancing mucosal integrity. PLoS One 2:e1308

Vejborg RM, Friis C, Hancock V, Schembri MA, Klemm P (2010) A virulent parent with probiotic progeny: comparative genomics of *Escherichia coli* strains CFT073, Nissle 1917 and ABU 83972. Mol Genet Genomics 283:469–484

Vieira G, Sabarly V, Bourguignon PY et al (2011) Core and panmetabolism in *Escherichia coli*. J Bacteriol 193:1461–1472

von Buenau R, Jaekel L, Schubotz E, Schwarz S, Stroff T, Krueger M (2005) *Escherichia coli* strain Nissle 1917: significant reduction of neonatal calf diarrhea. J Dairy Sci 88:317–323

Walk ST, Alm EW, Gordon DM et al (2009) Cryptic lineages of the genus *Escherichia*. Appl Environ Microbiol 75:6534–6544



Wang X, Rochon M, Lamprokostopoulou A, Lunsdorf H, Nimtz M, Römling U (2006) Impact of biofilm matrix components on interaction of commensal *Escherichia coli* with the gastrointestinal cell line HT-29. Cell Mol Life Sci 63:2352–2363

Welch RA, Burland V, Plunkett G 3rd et al (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. Proc Natl Acad Sci USA 99:17020–17024

Whittam TS, Wolfe ML, Wachsmuth IK, Ørskov F, Ørskov I, Wilson RA (1993) Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. Infect Immun 61:1619–1629

Wiedenbeck J, Cohan FM (2011) Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches. FEMS Microbiol Rev 35:957–976

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

Wold AE, Caugant DA, Lidin-Janson G, de Man P, Svanborg C (1992) Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics. J Infect Dis 165:46–52

Zdziarski J, Svanborg C, Wullt B, Hacker J, Dobrindt U (2008) Molecular basis of commensalism in the urinary tract: low virulence or virulence attenuation? Infect Immun 76:695–703

Zdziarski J, Brzuszkiewicz E, Wullt B et al (2010) Host imprints on bacterial genomes—rapid, divergent evolution in individual patients. PLoS Pathog 6:e1001078

Zhou Z, Li X, Liu B et al (2010) Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor. PLoS One 5:e8700

1.3 AUTOTRANSPORTER PROTEINS

Eight protein secretion systems (type I–VIII) have been identified in Gram-negative bacteria. A protein needs to pass three barriers, the inner membrane (IM) (cytoplasmic), the peptidoglycan mesh, and the outer membrane (OM), to be secreted out of a Gram-negative cell. Several secretion nanomachines act in a two-step mechanism and are dependent on general translocation mechanisms to traverse the IM (Desvaux et al., 2009; Gawarzewski et al., 2013). Classical monomeric single polypeptide autotransporter (AT) proteins belong to the type Va secretion pathway, which was seen as the simplest of the secretion mechanisms to transport proteins to the surface of Gram-negative bacteria (Bernstein, 2015; Drobnak et al., 2015).

AT protein domains

ATs are unique as they were earlier regarded to be able to catalyze their own translocation across the OM of Gram-negative bacteria. The quite diverse AT proteins retained a common domain structure to facilitate this OM translocation (Figure 7 on the facing page). Transport of ATs over the IM is dependent on the general Sec translocon, which is directed by the amino-terminal (N-terminal) *signal peptide (SP)* of ATs. Further in the direction to the carboxy (C) terminus are the *passenger* domain, which is the secreted and functional part of the AT, a *linker*, and finally the C-terminal 12-stranded transmembrane β-*barrel* domain. The last two domains function primarily in the translocation of the passenger domain over the OM. The linker is quite diverse in sequence and protein structure between different AT proteins, but many linkers have an α-helical domain (that spans the β-barrel pore during secretion) and a structurally disordered protein region. The secreted passenger either remains covalently linked to the β-barrel or is cleaved in the linker (autocatalytically or by an extracellular protease) to remain non-covalently attached to the OM or be secreted (Drobnak et al., 2015).

AT secretion mechanism

Recently, more detailed investigations into the precise secretion mechanism revealed a more complex process for OM translocation, which is less autonomous and includes several essential chaperones in the periplasm (Figure 7). Nevertheless, the AT multi-domain structure, especially the passenger domain, is essential for efficient translocation and folding of the N-terminal functional region. After IM transport via the Sec general secretion pathway the SP is cleaved from the AT precursor. The AT, especially the C-terminal β-barrel, are stabilized by chaperones in an intermediate structure during crossing of the periplasm. Afterwards, the β-barrel domain inserts into the OM to form a hydrophilic pore that spans the OM, a process that is catalyzed by the β-barrel assembly machinery (BAM). The passenger domain can then pass through the pore (which is subsequently occupied by the linker α-helix), which might be additionally assisted by the BAM complex, and reach the extracellular milieu (Bernstein, 2015; Drobnak et al., 2015; Gawarzewski

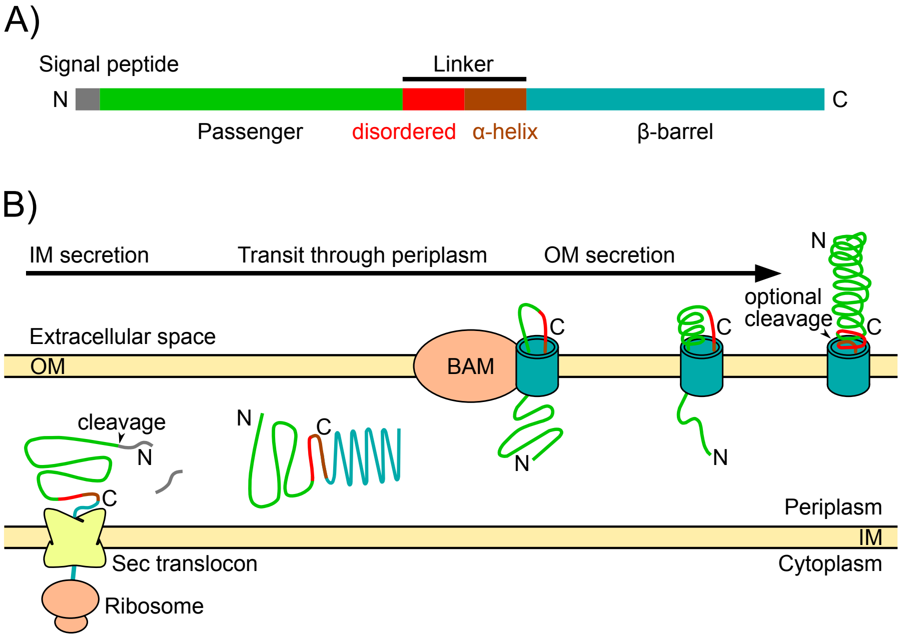


Figure 7: Autotransporter (AT) conserved multi-domain structure (A) and simplified type Va secretion mechanism diagram (B). Translocation over the IM requires the N-terminal signal peptide which targets the AT to the Sec complex. The signal peptide is cleaved of via a periplasmic signal peptidase after IM transit. Subsequently, the β-barrel domain forms a pore in the outer membrane, catalyzed by the BAM complex, for translocation of the passenger domain and display on the cell surface. Finally, the passenger domain might be released via proteolytic cleavage. Figure redrawn based on Drobnak et al. (2015) with adaptations, using `Inkscape`.

et al., 2013). Finally, the passenger domain can fold into its functional tertiary protein structure at the cell surface. The other four currently known type V secretion systems (T5SSs) (two-partner Vb, trimeric Vc, fused two-partner Vd, and inversed AT intimin/invasin Ve pathways) follow a very similar secretion mechanism but either consist of several polypeptides or have a different domain order. However, their secretion mechanisms are not as extensively studied (Gawarzewski et al., 2013). For example, the type Vc secretion pathway consists of three identical polypeptide chains that form a trimeric polypeptide passenger domain and a single 12-stranded β-barrel (Bernstein, 2015; Gawarzewski et al., 2013).

Despite the common domain structures of AT proteins, *E. coli* ATs are quite diverse in their sequence especially in the functional passenger domain (Drobnak et al., 2015). This sequence diversity of the passenger domain reflects the wide variety of virulence-associated functions ATs can perform, like adhesion, biofilm formation, serum resistance, enzymatic (esterases and proteases, e. g. serine protease autotransporters of *Enterobacteriaceae* (SPATE)), and cytotoxic activity (Celik et al., 2012; Gawarzewski et al., 2013; Henderson and Nataro, 2001; Wells et al., 2010). Therefore, AT proteins traditionally were regarded as virulence factors (VFs) of pathogenic *E. coli* isolates. SPATEs, e. g. , are implicated in mucosal damage and colonization in the EAEC pathotype (Rasko et al., 2011). Many specific functions were associated with individual *E. coli* pathotypes to support their particular pathogenesis mechanisms (Easton et al., 2011; Restieri et al., 2007; van der Woude and Henderson, 2008; Wells et al., 2010). However, a more detailed analysis, especially incorporating the phylogenetic background of the strains, has not been done. Particularly, since many of the AT functions do not necessarily relate to virulence but can also be described as fitness factors (FFs) for commensal *E. coli*.

E. coli *ATs as VFs*

1.4 GERMAN 2011 O104:H4 STEC OUTBREAK

In 2011, from May to July, the to date largest and most deadly Shiga toxin-producing *E. coli* (STEC) epidemic swept over Germany and neighboring countries. This epidemic resulted in about 4,000 infections, many requiring hospitalization, and more than 50 fatalities (Croxen et al., 2013; Frank et al., 2011; Monecke et al., 2011).

Initially, it was assumed the infectious agent is a classical EHEC – a notorious intestinal pathogenic *E. coli* (IPEC) pathotype which causes food-borne outbreaks of severe diarrhea and hemorrhagic colitis. Most of these outbreaks are caused by several major EHEC serotypes[18], most

[18] A serotype is an antigenically distinct variety in a bacterial species used for typing. In *E. coli* this is based on the O-antigen of lipopolysaccharide (LPS) (O), flagellin of flagella (H), and capsular (K) antigens, which are numbered consecutively (Kaper et al., 2004).

prominently O157:H7 (Croxen et al., 2013; Kaper et al., 2004). However, this epidemic was unusual in that it affected mainly healthy adults (especially young women) instead of children (Frank et al., 2011). Urgent public health measures were set in place and the Robert Koch Institute (RKI) in cooperation with the national consulting laboratory for haemolytic uraemic syndrome (HUS) at the Münster University Hospital (UKM) were able to identify the strain (Table 6 on page 198).

Classical microbiological diagnostic methods showed that it was a STEC of a rare O104:H4 serotype, with its main VF being the phage-encoded Shiga toxin (Stx) type 2 (*stx2*)[19]. However, the outbreak strain was missing the typical EHEC locus of enterocyte effacement (LEE) island with its encoded type III secretion system and its intimin adhesin (*eae*) and Tir translocated receptor effectors important for the formation of attaching/effacing (A/E) lesions. *E. coli* encoding for Stx but missing the LEE GI, are generally summarized under the abbreviation STEC (Croxen et al., 2013). The unusual serotype and VF repertoire was the reason diagnostic laboratories had difficulties in identifying the culprit with classical diagnostic measures like selective media enrichment and immunoassay serotyping (Chattaway et al., 2011).

We were one of the teams capable of fast genomic analyses of these food-borne isolates during the *ongoing* outbreak to elucidate their pathogenesis.

1.5 BOVINE MASTITIS

Bovine mastitis is an inflammatory immune reaction of the udder mammary tissue, primarily in response to invading microorganisms. The inflammation has the purpose of neutralizing the infectious agent and returning the mammary tissue to a normal function. Fungi, yeast, and most importantly bacteria are the major microbial causes of bovine mastitis. The inflammation leads to the deterioration of milk secretion tissue[20] and as a consequence adversely effects milk quantity and quality (Bradley, 2002; Long et al., 2001; Rainard et al., 2016; Younis et al., 2016). Mastitis is a worldwide problem and is the most costly infectious disease in cows, causing multi-billion dollar economic losses through decreased milk quality/production, treatment costs[21], early dry-off, or in extreme cases culling or death of the animal (Bradley, 2002; Hogeveen et al., 2011). The most important bacterial mastitis pathogens in the modern dairy industry are *Staphylococcus aureus*, *E.*

[19] Endocytosis of Stx into epithelial cells disrupts protein synthesis at the ribosome and ultimately leads to cell death. On a global scale, absorption of Stx into the bloodstream leads to kidney damage and finally kidney failure, thus patients need maintenance of fluid/electrolyte levels and dialysis to filter the toxin out of the blood stream (Karch et al., 2012).

[20] Apoptosis and necrosis are induced by proinflammatory mediators or bacterial effectors.

[21] Antibiotics are used pervasively in treating mastitis (Section 6.4.2.2 on page 209).

coli, and *Streptococcus uberis* (Rainard and Riollet, 2006; Schukken et al., 2011).

Bovine udder infections occur according to the following steps (Figure 8 on the next page): Pathogenic bacteria contaminate the teat skin and subsequently invade the teat canal. From there they can disseminate in the milk duct systems and adhere to the lining epithelium, e. g. in the alveoli. Bacteria that can successfully cause intramammary infections (IMIs) replicate in the milk and can withstand the washing out during milking either through sufficient replication, motility, and/or adhesion. Several pathogens might form biofilms on alveolar epithelial cells, others invade epithelial cells and evade detection of the immune system (Gomes et al., 2016; Shpigel et al., 2008).

The udder has several intrinsic mechanisms to repel microbial invasion and prevent growth of infectious agents. The first line of defense are physical barriers, like the teat canal that seals the teat through the sphincter muscle after each milking and the keratin plug that is formed in the teat canal during the dry period[22]. Additional mechanisms include iron sequestration via lactoferrin (especially during the dry period), antimicrobial peptides, lysozyme, complement, antibodies, and most importantly the presence of different leukocyte cells (like polymorphonuclear neutrophils (PMNs), macrophages, and to a lesser extent lymphocytes) (Section 1.5.2 on page 53) (Isobe et al., 2009; Rainard and Riollet, 2006; Sordillo and Streicher, 2002).

Given that bovine milk exhibits a complex and dynamic microbial diversity, the mammary gland harbors a natural microbiota that inhibits the propagation of pathogenic bacteria. Many mastitis pathogens grow exponentially in the mammary gland and therefore cause a dysbiosis in this microbiota. However, in cured cows microbiota complexity recovers quickly after acute clinical mastitis at the same time as pathogen incidence declines. These effects depend strongly on the bacterial mastitis pathogen involved (Ganda et al., 2016).

1.5.1 *Bovine mastitis pathogens and disease pathogenesis*

Mastitis bacterial pathogens are classically divided into two subtypes depending on the primary reservoir and mode of transmission (Table 1 on page 52): “contagious” or “environmental” (Blowey and Edmondson, 2010; Bradley, 2002). *Contagious* pathogens can persist in the cow udder and spread the disease via direct propagation, because they are adapted to the bovine udder habitat. *Staphylococcus aureus*, *Streptococcus dysgalactiae*, and *Streptococcus agalactiae* are the major representatives of this class. *E. coli* and *Streptococcus uberis* make up the majority of *environmental* pathogens, which can survive outside the host, are not adapted to survival within the udder, and rather cause disease opportunistically by infecting each animal separately (Bradley, 2002; Passey

[22] Keratin also has a microbicidal activity.

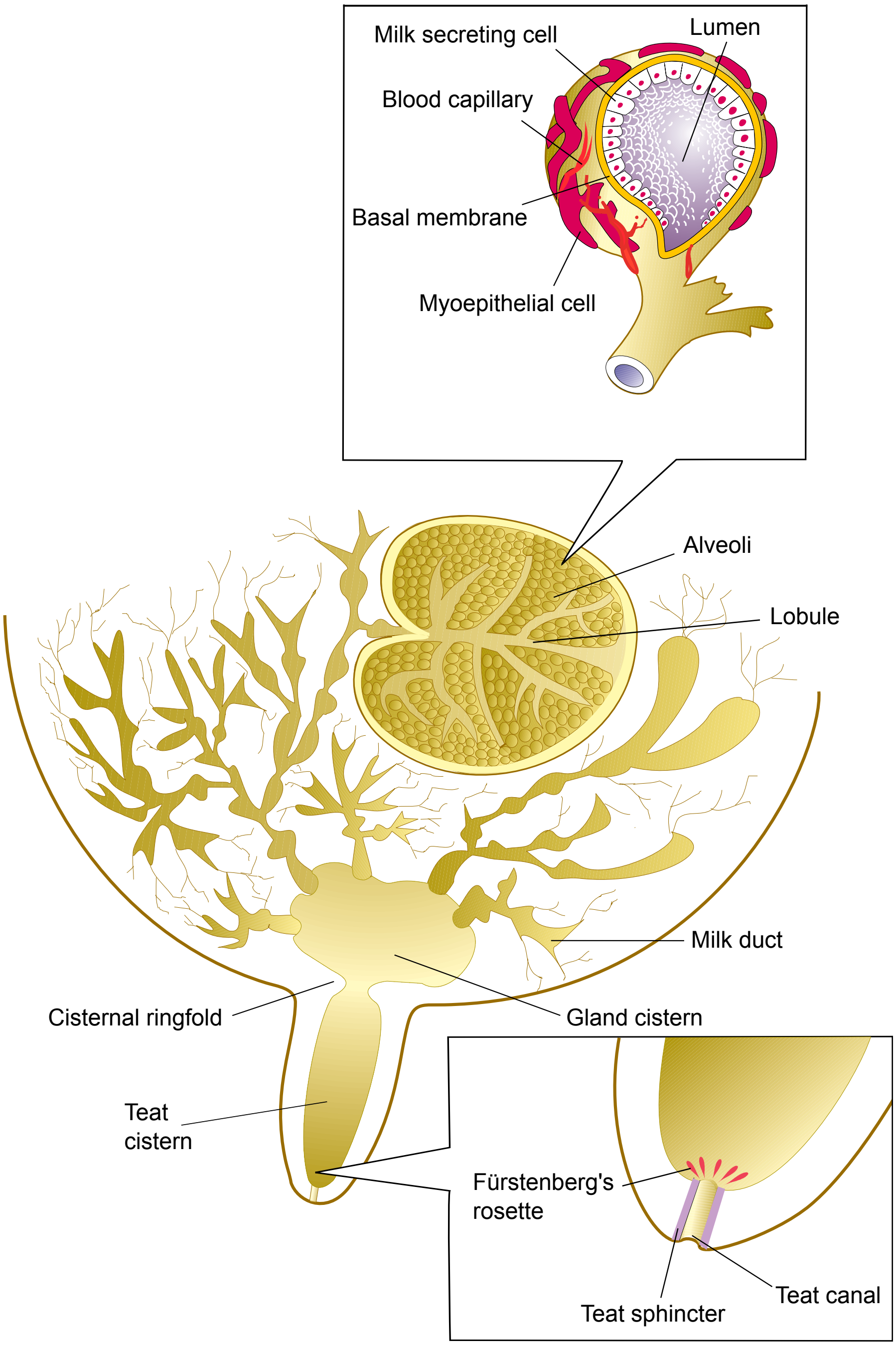


Figure 8: Illustration of the bovine udder anatomy with enlargements of the teat tip and a milk secretory alveolus. Figure adapted from the brochure "Efficient Milking" (© DeLaval; `http://www.delaval.ch/ImageVaultFiles/id_27591/cf_5/Efficient-milking.PDF`) with `Inkscape` and reproduced with permission of the company.

et al., 2008). Their infection sources are bedding materials, soil, and manure (Hogan and Larry Smith, 2003). *E. coli* is a prime example of an environmental mastitis pathogen, especially since it is a commensal of the bovine gastrointestinal tract and thus ubiquitous present on cow farms (Burvenich et al., 2003).

Interestingly, at the same time as the prevalence of contagious mastitis was successfully reduced[23], environmental mastitis caused by *E. coli* and *Streptococcus uberis* stepped into the breach and their incidence increased in many countries. These environmental pathogens are more difficult to control, as isolation of infected animals and milking hygiene will not completely protect healthy animals (Blowey and Edmondson, 2010; Bradley, 2002; Hogan and Larry Smith, 2003).

Table 1: Major bacterial bovine mastitis pathogens.

PATHOGEN	GRAM	INFECTION MODE	PATHOGENESIS
Streptococcus agalactiae	+	contagious	chronic
Streptococcus dysgalactiae	+	contagious	chronic
Staphylococcus aureus	+	contagious	chronic
Coliforms			
Escherichia coli	-	environmental	acute
Klebsiella pneumoniae/oxytoca	-	environmental	acute
Streptococcus uberis	+	environmental	acute/chronic

The disease outcome of an udder infection depends on its etiology:

CHRONIC MASTITIS *S. aureus* particularly causes a chronic mastitis that is less severe, but with sometimes lifelong *persistence* of the bacteria. Because these strains can persist in the udder for a prolonged time, often without obvious external indications of symptoms (presenting themselves as subclinical), chronic infections usually have recurrent episodes of mastitis involving the same strain (Bannerman, 2009; Petzl et al., 2008; Schukken et al., 2011). Invasion of udder epithelial cells has been proposed for some bacterial pathogens as a virulence mechanism for persistence. In this case, the intracellular bacteria might serve as a point of origin for recurrent mastitis episodes and avoid detection by the immune system (Dogan et al., 2006; Schukken et al., 2011).

subclinical mastitis *– no visible signs of the disease*

ACUTE MASTITIS Acute mastitis infections, show a contrary disease development. They have a very fast onset with often severe systemic

[23] The most important control measures to reduce their incidence were improved milking hygiene, dry period antibiotic therapy, and culling of chronically infected animals.

clinical symptoms, like a swollen and painful udder, elevated rectal temperature, dullness, loss of appetite, and diarrhea (Dufour et al., 2011; Hogan and Larry Smith, 2003). The infection is usually *transient*, the invaders rapidly eliminated, and animals recover within days. Nevertheless, the inflammation can become systemic and lead to sepsis concurrent with fever, where cows either die or have to be culled. Although infections are generally cleared rapidly, long-term detrimental effects on milk quality can remain. *E. coli* is the leading cause of acute mastitis in dairy animals, but infections can actually result in large variations of clinical symptoms (Blum et al., 2014; Burvenich et al., 2003; Hogan and Larry Smith, 2003; Zadoks et al., 2011).

clinical mastitis – *visible signs of mastitis*

Interestingly, the representative pathogens causing chronic (*S. aureus*) or acute (*E. coli*) mastitis induce a different immune response in the host during an infection. This might be the underlying mechanisms that lead to the different disease outcomes.

1.5.2 *The cow immune system in the udder*

The inflammation process in mastitis decidingly depends on the *innate immune* response. Alveolar macrophages and epithelial cells with the Toll-like receptor (TLR) class of receptors detect general *pathogen-associated molecular patterns (PAMPs)* and respond with the secretion of inflammatory mediators and modulators, like cytokines and acute phase proteins. These in turn attract leukocytes, especially blood neutrophils[24], to the inflammation site into the cistern and ducts of the alveolar milk space, that either clear the infection and/or elicit additionally an adaptive immune response (Paape et al., 2003; Sordillo and Streicher, 2002). The most effective protection against mastitis is an integrated response of the innate and adaptive arms of the immune system. Macrophages are essential in connecting the innate and adaptive immune response by serving as antigen-presenting cells (Rainard and Riollet, 2006).

The most important PAMPs for Gram-positive bacteria are lipoteichoic acid (LTA) and peptidoglycan, recognized by TLR2. Gram-negative bacteria are characterized by the OM lipopolysaccharide (LPS) endotoxin component, detected by TLR4 (Rainard and Riollet, 2006; Rainard et al., 2016; Schukken et al., 2011).

The difference in disease progression of chronic and acute infections is considered to be a consequence of an impaired proinflammatory cytokine response during a *S. aureus* infection through a lack of nuclear factor κB (NF-κB) activation in mammary epithelial cells. As a result, an udder infected with *S. aureus* shows a delayed chemokine interleukin (IL)-8 (CXCL8) and proinflammatory cytokines TNF-α (Tumor necrosis factor) and IL-1 upregulation in contrast to the cytokine storm in an inflamed udder infected by *E. coli*. Thus, the overall increase

[24] PMNs make up to 70–80% of all migrated leukocytes in IMIs.

in milk somatic cell count (SCC)[25] is delayed in *S. aureus* infections compared to *E. coli*, accompanied by a delayed neutrophil recruitment. This results in an initial immune response that the invading bacteria can survive and contributes to the persistent and chronic course of *S. aureus* infections (Bannerman, 2009; Rainard and Riollet, 2006; Schukken et al., 2011; Wellnitz et al., 2006; Yang et al., 2008b; Younis et al., 2016).

PMNs essential countermeasure to microbial mastitis

The most important countermeasure to invading mammary bacteria is the recruitment and diapedesis of blood PMNs. PMNs clear the infectious agent by phagocytosis and respiratory burst or with neutrophil extracellular traps (NETs), and strengthen the inflammatory response through a release of cytokines. A prompt and efficient recruitment of these effector cells is the main factor influencing the severity of the disease, especially during *E. coli* infections (Bradley, 2002; Burvenich et al., 2003; Paape et al., 2003; Rainard and Riollet, 2006). Otherwise, exponential and unrestricted pathogen growth in the udder (putatively combined with an inhibition of neutrophil migration) might overwhelm the acute phase immunity reaction. An increased presence of LPS is then too high for endotoxin detoxification and clearance through bovine PMNs, triggering an overshooting immune response and inflammation through TNFα release, and ultimately leading to endotoxic shock and severe systemic clinical symptoms (Burvenich et al., 2003; Rainard and Riollet, 2006; Sordillo and Streicher, 2002).

Bacterial FFs which ensure the survival of the pathogen in the presence of the host response or at least elongate the immune system response time are considered to be essential in mastitis. Especially factors for the evasion of neutrophils, like resistance to phagocytosis, NETs, respiratory burst of hydroxyl or oxygen radicals, or survival within neutrophils have been implicated. Opsonization of invading bacteria with complement or antibodies strongly stimulate PMNs to initiate phagocytosis and bactericidal activities (Paape et al., 2003).

cow factors define course of mastitis

Nevertheless, the course of mastitis depends not only on the type of pathogen but also strongly on the cow's genetic predisposition[26], its age[27] (Burvenich et al., 2003; Rainard and Riollet, 2006), the immune status, the lactational stage, as well as environmental factors[28] (Wellnitz et al., 2006). These factors can be summarized as "cow" and "environmental" factors, and in fact have been proposed as the main defining parameters in *E. coli* bovine mastitis (Burvenich et al., 2003). Because of these different *cow factors*, cattle vary strongly in their ability to overcome mastitis infections and play an active role in the development of disease (Younis et al., 2016).

[25] The SCC is the number of cells (leukocytes and epithelial cells) per milliliter in milk, used as a measure of the immune response.

[26] E. g. different milk concentrations of humoral defenses, immune cells . . . in the healthy udder.

[27] Primiparous cows have a stronger immune response than older animals (Burvenich et al., 2003).

[28] The environment of a cow like nutritional status, barn environment etc.

1.5.3 E. coli *bovine mastitis*

E. coli is a member of the environmental mastitis pathogens (Section 1.5.1 on page 50) and invasion of the udder is considered to happen when the teat orifice is open, after milking or teat damage. As a consequence, the pathogen is regarded to be an opportunistic pathogen (Bean et al., 2004). About 80% of coliform mastitis cases are caused by *E. coli* (Suojala et al., 2013).

E. coli *opportunistic environmental bovine mastitis pathogen*

On the herd level, especially high producing cows and well-managed herds with a low bulk SCC are prone to *E. coli* mastitis infections (Blowey and Edmondson, 2010; Burvenich et al., 2003). Individually, cows are more susceptible to *E. coli* IMI during the early dry and periparturient period in comparison to mid-lactation, because of a compromised immune system (Burvenich et al., 2007; Burvenich et al., 2003). Additionally, infections acquired during the dry period can persist subclinically and subsequently lead to a clinical mastitis in early lactation (Bradley, 2002; Schukken et al., 2011; Zadoks et al., 2011). Generally, *E. coli* IMIs cause an acute onset of bovine mastitis (Table 1 on page 52), but infection occasionally leads to a persistent state usually with a mild or subclinical pathogenesis (Döpfer et al., 1999; Fairbrother et al., 2015; Schukken et al., 2011; Zadoks et al., 2011). Invasion of and persistence in epithelial mammary cells is considered to be a distinctive feature of persistent/chronic *E. coli* bovine udder infections with subclinical pathology (Almeida et al., 2011; Dogan et al., 2006; Döpfer et al., 2000). However, results of a study with a larger strain panel were inconsistent (Dogan et al., 2012). Nevertheless, in a review article Shpigel et al. (2008) reported on the existence of *E. coli* strains that supposedly trigger acute/transient or chronic/persistent infections in a consistent way.

Because *E. coli* are adapted to the bovine gastrointestinal tract, a rapid metabolic adjustment to mammary secretion is important to promote growth. The acute onset of *E. coli* mastitis is a direct consequence of exponential intramammary growth, that also expresses itself in dramatic changes in the udder microbiota during an infection (Ganda et al., 2016). Bacterial numbers in milk are in direct correlation to the severity of the disease, triggering a stronger inflammation response through mammary epithelial cells and macrophages (Hogan and Larry Smith, 2003; Kornalijnslijper et al., 2004; Schukken et al., 2011; Vangroenweghe et al., 2004).

Three important necessary adaptations to the udder milieu have been proposed for mastitis pathogenic *E. coli*: Utilization of milk lactose as carbon and energy source, growth under microaerobic conditions, and the scavenging of iron from chelators like citrate and lactoferrin[29] (Lippolis et al., 2009). These adaptations are inherent capabilities of

[29] Lactoferrin concentration is especially high during the dry period and its concentration increases dramatically during acute mastitis.

members of the species *E. coli*, albeit in different phenotypical manifestations. Additional requirements have been proposed to elicit IMIs: evading cellular host defenses (especially phagocytosis by neutrophils e. g. via capsules that block opsonization[30]), serum resistance, biofilm formation, adhesion to and invasion of epithelial cells, motility to invade the teat canal and milk duct system, and toxins[31] (Blum et al., 2008; Gomes et al., 2016; Hogan and Larry Smith, 2003; Rainard and Riollet, 2006; Shpigel et al., 2008; Wenz et al., 2006). *Serum resistance* was traditionally considered the most important virulence-associated trait of *E. coli* capable of eliciting IMI. Considering that different antibodies, complement, and defensins (e. g. lingual antimicrobial peptide (LAP)) are present abundantly in milk following bacterial udder invasion, this is a sensible expectation (Isobe et al., 2009; Petzl et al., 2008; Wenz et al., 2006). However, the serum resistance phenotype and associated genes[32] were detected with varying frequencies and are similarly also present in fecal commensal bovine *E. coli* isolates from the cow shed (Blum et al., 2008; Blum and Leitner, 2013; Kaipainen et al., 2002; Nemeth et al., 1991, 1994).

serum resistance traditionally considered most important E. coli *virulence property*

Overall, the bovine gastrointestinal tract is a natural reservoir for commensal and pathogenic *E. coli* of high phylogenetic and genotypic diversity (Houser et al., 2008). Consequently, also isolates from bovine mastitis show a large heterogeneity in serotypes and genotypes regardless of disease severity – and importantly comparable to fecal isolates. This supports the classification of this environmental pathogen as an opportunistic one (Blum et al., 2008; Dogan et al., 2012; Nemeth et al., 1994; Suojala et al., 2011; Wenz et al., 2006; Zadoks et al., 2011). Countless studies have tried to associate specific putative VFs with *E. coli* IMI isolates with varying degrees of success (Table 2 on page 58). Although several of them show significant association of some VFs, all studies combined do not share a common set of VFs for mastitis *E. coli* in contrast to typical IPEC pathotypes, and no significant differences have been found between *E. coli* strains from transient/persistent infections or severity of disease. The *E. coli* isolates rather lack most of the characterized VFs present in other pathogenic *E. coli* or pathotypes (including ExPEC VFs) (Bean et al., 2004; Blum and Leitner, 2013; Burvenich et al., 2003; Fernandes et al., 2011; Ghanbarpour and Oswald, 2010; Kaipainen et al., 2002; Kempf et al., 2016; Nemeth et al., 1994; Suojala et al., 2011; Wenz et al., 2006; Zadoks et al., 2011). Keep in mind, that most of the publications in Table 2 applied PCR for VF detection in the *E. coli* isolates, which of course depends strongly on the primer panel chosen[33]. Several of these studies used overlapping gene panels

no common set of VFs for mastitis E. coli *isolates*

[30] Other surface exposed structures of *E. coli*, like O-antigen moieties of LPS, have also been implicated in affecting phagocytosis susceptibility (Hogan and Larry Smith, 2003).

[31] Especially toxins that damage the mammary tissue, like cytotoxins and hemolysins.

[32] Genes associated with serum resistance encode e. g. for OM proteases or capsules.

[33] In these experimental set ups you can only find what you are looking for.

for detection. Additionally, overall VF presence in the respective strain panels was low (mostly below 30% of mastitis-associated *E. coli*).

LPS fundamental property of E. coli *to cause udder inflammation*

Consequently, virulence of *E. coli* mastitis isolates is solely attributed to the general pyogenic properties of *E. coli* PAMPs, foremost LPS (Section 1.5.2 on page 53). Different severities of *E. coli* mastitis (mild to severe) and/or outcomes (acute or subclinical) must then be a consequence of variation in the environmental and cow factors (Burvenich et al., 2003; Zadoks et al., 2011) (Section 1.5.2 on page 53). An intramammary injection with LPS induces the same local signs as observed during *E. coli* infections (Burvenich et al., 2003; Long et al., 2001; Rainard and Riollet, 2006; Shpigel et al., 2008). However, the effects of LPS in an *in vivo* infection are augmented by other *E. coli* PAMPs, which leads to the acute inflammatory systemic reaction (Shpigel et al., 2008).

mammary pathogenic E. coli *(MPEC) pathotype*

In spite of the presented evidence, several studies contrary reported on an *E. coli* genetic subset of strains from the overall bovine *E. coli* population predisposed to elicit bovine mastitis. This subset has a restricted geno- and phenotype in comparison to commensal fecal bovine *E. coli* (Blum et al., 2008; Blum and Leitner, 2013; Bradley, 2002). The strains e. g. showed a faster lactose utilization, correlated with a faster growth in milk, as well as lower phagocytosis susceptibility by PMNs (Blum et al., 2008; Blum and Leitner, 2013). Because of the adaptation requirement of *E. coli* to the udder and reports on strains having properties to cause predominantly subclinical mastitis (Almeida et al., 2011; Dogan et al., 2006; Döpfer et al., 1999, 2000), a *mammary pathogenic* E. coli *(MPEC) pathotype* has been proposed (Bradley and Green, 2001; Shpigel et al., 2008). Since mastitis is an extraintestinal disease, analogy was drawn to ExPEC pathotypes where *E. coli* strains harboring different combinations of VFs or FFs can lead to the same disease, in this case IMI (Section 1.2.1 on page 13). Thus, mastitis *E. coli* isolates could include a large genotypical diversity. Furthermore, these variable VF combinations might enable *E. coli* to elicit the different mastitis disease outcomes associated with the species (Kempf et al., 2016; Shpigel et al., 2008). Detailed comparative genomics studies proposed concordantly that different lineages of *E. coli* might be more capable in causing mastitis, i. e. more than one pathogenic subset might exist. Therefore, a selection of *E. coli* strains within the intramammary environment must take place (Goldstone et al., 2016; Kempf et al., 2016).

Table 2: A selection of publications that examined the presence of VF genes in *E. coli* isolated from milk of dairy cattle with mastitis.

STUDY	ISOLATES FROM	LOOKED FOR	NO. OF ISOLATES	PREDOMINANTLY FOUND
Kaipainen et al. (2002)	Finland & Israel	*E. coli* VFs	273	*traT*, *cnf1/2*, *aer*, *f17*, *sfa*
Lira et al. (2004)	Brazil	STEC VFs	182	few isolates with *stx1/2*, *eaeA*, *hly*
Bean et al. (2004)	New Zealand	O157:H7 STEC VFs	80	several isolates with Stx
Wenz et al. (2006)	USA	*eae*, *cs31a*, *cnf1/2*	123	only low presence of VFs
Dogan et al. (2006)	USA	*E. coli* VFs	6	no VFs found
Ghanbarpour and Oswald (2010)	Iran	*E. coli* VFs	127	*f17A*, *iucD*, *cnf2*
Suojala et al. (2011)	Finland	ExPEC and EHEC VFs	154	*irp2*, *iucD*, *papC*, *iss*
Fernandes et al. (2011)	Brazil	*E. coli* VFs	27	*stb*, *cs31a*, Stx2
Dogan et al. (2012)	USA	*E. coli* VFs	28	type II secretion system (T2SS), type IV secretion system (T4SS), type VI secretion system (T6SS), *lpfA*, *fyuA*
Blum and Leitner (2013)	Israel	*E. coli* VFs	63	*lpfA*, *astA* (EAST-1), *iss*
Liu et al. (2014)	China	ExPEC and EHEC VFs	70	*f17A*, *irp2*, *astA*, *iucD*, *colV*

Continued on next page.

Table 2: A selection of publications that examined the presence of VF genes in *E. coli* isolated from milk of dairy cattle with mastitis (continued).

STUDY	ISOLATES FROM	LOOKED FOR	NO. OF ISOLATES	PREDOMINANTLY FOUND
Fairbrother et al. (2015)	Canada	ExPEC VFs	97	*hra1*, *hlyA*, yersiniabactin, *iss*
Richards et al. (2015)	USA	Comparative genomics	4	T6SS
Blum et al. (2015)	Israel	Comparative genomics	3 + 1 bovine commensal	genes involved in LPS synthesis, sugar metabolism, *fecA*
Kempf et al. (2016)	France & Israel	Comparative genomics & *E. coli* VFs	5 + 1 bovine commensal	ferric iron(III)-dicitrate uptake system (Fec), AraC family regulator, *dosP*, Clp-like protein
Goldstone et al. (2016)	Europe & Israel	Comparative genomics	66 (phylo-group A)	*ycdU-ymdE*, phenylacetic acid degradation (*feaRB*, *paaFGHIJKXY*), Fec

Studies colored according to main method used: PCR, DNA colony or microarray hybridization, comparative genomics.

putative mastitis-associated E. coli *(MAEC) VFs actually commensal FFs*

Nevertheless, the genetic abilities of the putative MPEC pathotype might just be facultative, a by-product of their function in commensalism – with the primary ability to colonize and persist in the bovine gastrointestinal tract, as has been proposed for ExPEC (Diard et al., 2010; Le Gall et al., 2007; Leimbach et al., 2013; Nowrouzian et al., 2005; Schierack et al., 2008; Tenaillon et al., 2010; Tourret and Denamur, 2016). A commensal *E. coli* that turns into a pathogenic strain, requires not only the acquisition of fitness factors, but also genetic information that directly contributes to pathogenesis, which incidentally is the definition of a pathotype.

2 THESIS AIMS

The overall theme of this thesis is the genomic analysis of pathogenic *E. coli* isolates and their pathotype definition with regard to their phylogenetic history. Because traditionally *E. coli* pathotypes were regarded as clonal and very similar in their VF repertoire (Section 1.2.1 on page 13) many studies disregarded the phylogenetic background of the analyzed strains. However, putative functional relatedness through HGT or parallel evolution can be overshadowed by vertical phylogenetic ancestry.

In the end, the aims of this thesis were manifold and evolved during its execution:

1. Establishment of a bioinformatical database, tools, and workflows for the comparative genomics needs of the thesis
2. Determination of the distribution of AT proteins in *E. coli* pathotypes and phylogroups in order to analyze if ATs are associated with pathotypes
3. Genomic evaluation of STEC isolates from the ongoing 2011 German outbreak by characterizing the phylogenetic relationship to other *E. coli*, as well as *E. coli* VF and antibiotic resistance gene carriage
4. Comparative genomics of *E. coli* isolates from bovine mastitis. The main goal was to shed light on the contradictory hypotheses if an MPEC pathotype with a distinct set of virulence-associated genes exists.
   a) Characterization of a novel LPS O-antigen moiety of *E. coli* strain 1303
   b) Genome finishing of two *E. coli* mastitis isolates and high quality annotation
   c) WGS of six *E. coli* mastitis and six fecal commensal *E. coli* isolates of different phylogenetic groups and their annotation
   d) Detailed genomic analysis of phylogenetically diverse *E. coli* mastitis isolates in comparison to bovine commensal isolates to potentially identify virulence-associated fitness traits that contribute to the establishment of bovine mastitis

Part II

E. COLI VIRULENCE FACTOR COLLECTION AND BACTERIAL GENOMICS SCRIPTS

3 *E. COLI* VIRULENCE FACTOR COLLECTION

3.1 INTRODUCTION

The advancements of genomic sciences and high-throughput techniques in microbiology (Section 1.1.1.1 on page 6 and Section 1.1.3 on page 11) has sparked the creation of numerous data resources and increased their importance tremendously (Chandras et al., 2009; Helmy et al., 2016). This is also the case for specific *E. coli* databases. Although there is a wealth of publications characterizing the VF content of *E. coli* genomes, surprisingly there is no database for VF protein sequences that is open to all, easy to use for comparative genomics, well annotated, up-to-date, and comprehensively includes[1] *ExPEC VFs* (Figure 9).

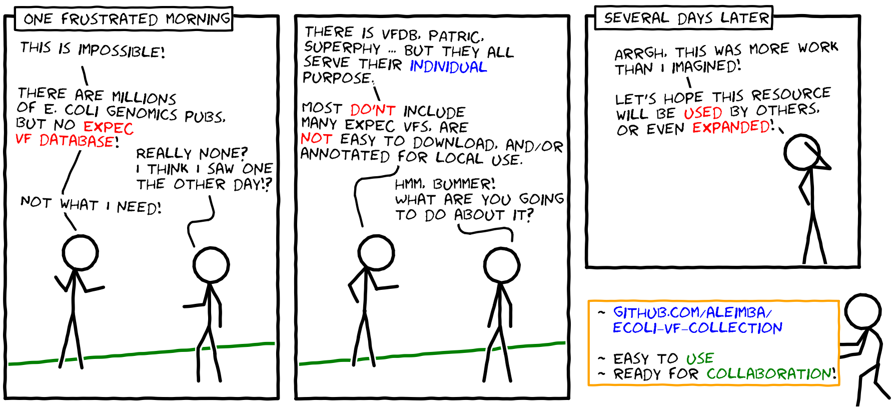

ONE FRUSTRATED MORNING
THIS IS IMPOSSIBLE!
THERE ARE MILLIONS OF E. COLI GENOMICS PUBS. BUT NO EXPEC VF DATABASE!
REALLY NONE? I THINK I SAW ONE THE OTHER DAY!?
NOT WHAT I NEED!
THERE IS VFDB, PATRIC, SUPERPHY ... BUT THEY ALL SERVE THEIR INDIVIDUAL PURPOSE.
MOST DON'T INCLUDE MANY EXPEC VFS, ARE NOT EASY TO DOWNLOAD, AND/OR ANNOTATED FOR LOCAL USE.
HMM, BUMMER! WHAT ARE YOU GOING TO DO ABOUT IT?
SEVERAL DAYS LATER
ARRGH, THIS WAS MORE WORK THAN I IMAGINED!
LET'S HOPE THIS RESOURCE WILL BE USED BY OTHERS, OR EVEN EXPANDED!
~ GITHUB.COM/ALEIMBA/ ECOLI-VF-COLLECTION
~ EASY TO USE
~ READY FOR COLLABORATION!

Figure 9: A comic on why the `ecoli_VF_collection` was devised. Figure created in Randall Munroe's `xkcd` style (`https://xkcd.com/`) with Comix I/O (`http://cmx.io/`) and edited in `Inkscape`.

In older *E. coli* isolate characterization studies, that used PCR for VF detection, usually primer nucleotide sequences are given, but often corresponding accession numbers for the genes are omitted (see e. g. Johnson et al. (2006b) and Müller et al. (2007), and many similar studies). This restricts findability and easy access to the corresponding gene sequences. Instead one has to resort to searching gene names, which can be difficult as sequence databases are littered with small fragments of genes, ambiguous annotation, different alleles . . . Other authors provide the VF nucleotide and/or protein sequences in the supplemental material of their publications[2]. However, this is often in a format that

[1] Most important for my purpose of characterizing bovine mastitis *E. coli* isolates (Section 3.2 on page 67).

[2] Providing data as supplemental material can be a problem of its own regarding its lack of reviewing during peer-review and burried citations (Pop and Salzberg, 2015).

does not provide convenient computable access, like spreadsheets (e. g. Table S1 in Kempf et al. (2016)) or PDF files (e. g. Dataset S9 in Salipante et al. (2015)).

virulence factor database (VFDB)

There are several web services that specialize in VFs of bacterial pathogens. Four examples are the *virulence factor database (VFDB)*[3], the *Pathosystems Resource Integration Center (PATRIC)* (Mao et al., 2015; Wattam et al., 2017), the `VirulenceFinder` of the Center for Genomic Epidemiology (CGE) (Joensen et al., 2014), and *SuperPhy* (Whiteside et al., 2016). All of them are tailored for the "traditional wet-lab" microbiologist as target audience that is interested in manual queries via the web, and not computational biologists interested in batch data download or advanced programming interfaces (API) (Helmy et al., 2016). Thus, access to the complete underlying VF sequences is not always provided or sometimes the sequences can only be accessed manually via a graphical user interface. The most comprehensive and up-to-date resource is the VFDB, especially for classical IPEC VFs and corresponding alleles. Therefore, many other databases are derivatives of VFDB's dataset. Last but not least, all of these databases are closed, i. e. collaborative working on the datasets is not possible.

missing education and support for open dissemination of datasets

Because the above-mentioned databases lack several ExPEC VFs, many authors went repeatedly to the trouble of collecting sequences needed for their studies (some listed in Table 4 on page 69). Sadly, the current lack of knowledge and training in biological sciences inhibits an open, replicable, and easy dissemination of these small-scale datasets (Section 1.1.2 on page 9). Thus, researchers have to reinvent the wheel again and again. The current scientific system has no incentive to go the extra mile and create shareable and accessible data, as only publication citations and not data citations count to a scientists' reputation and career.

FAIR Data Principles

Nevertheless, sustained access to these resources is essential for scientific advancement like reproducibility (e. g. re-analysis), testing of new hypothesis the original database authors might not have foreseen, and developing new technologies (Chandras et al., 2009). There is a need for an easy and fast system to publish small-scale datasets, that fosters *findability*, *accessibility*, *interoperability*, and *reusability*/adaptability (the *FAIR Data Principles* (Wilkinson et al., 2016)). Ideally, resources should be designed for collaboration in order to reduce duplication of work (Helmy et al., 2016).

[3] The four VFDB releases include different datasets: R1 with experimentally validated VFs (Chen et al., 2005), R2 with VFs derived from comparative genomics including homologs from complete genomes (intra-genera) (Yang et al., 2008a), and R3 with a curated dataset categorized in VF classes (VF centric, inter-genera) (Chen et al., 2012a). The most recent release includes a new graphical interface and a reorganization of the datasets into an experimentally verified core VF dataset, setA, and a comparative genomics dataset, setB, to remove redundancies and enhance data quality (Chen et al., 2016).

The most popular site that conforms with many of these premises is the collaborative software development and hosting platform GitHub[4], launched in 2008. This platform was successfully used to publicly make data and analysis available during the 2011 German STEC outbreak (Section 6.2.2 on page 197). GitHub is based upon Git, a version-control system, build for tracking changes in computer files and coordinating this work between different contributors. The GitHub platform has enjoyed tremendous success in the life sciences for sharing, maintaining, and updating datasets and code, because of its transparency, speed, and ease of use. Therefore, GitHub repositories are increasingly being cited in the literature (Perkel, 2016).

GitHub for sharing code and data

3.2 PURPOSE, DESIGN, AND COMPOSITION OF THE *E. COLI* VIRULENCE FACTOR COLLECTION

For the characterization of bovine MAEC and commensal isolates in Leimbach et al. (2017) (Figure 4 and Figure S6 in Section 5.3.4 on page 140) I needed a comprehensive collection of *E. coli* VF protein sequences for my `prot_finder` pipeline (that runs on `BLASTP` (Altschul et al., 1990; Camacho et al., 2009); Table 5 on page 74). As MAEC are considered ExPEC (Section 1.5.3 on page 55) this collection required the inclusion of not only IPEC but especially *ExPEC VFs*. After extensive research into the MAEC and ExPEC typing/characterization literature and the aforementioned VF databases, I regrettably experienced several shortcomings (Figure 9 on page 65) and thus decided to start a separate VF collection with a focus on ExPEC VFs, the `ecoli_VF_collection` (`https://github.com/aleimba/ecoli_VF_collection`).

`ecoli_VF_collection`

In order to avoid reinventing the wheel, I based my VF collection on all VFDB releases and collected VFs from there and from the primary literature that fit to my purpose. Furthermore, the Petty et al. (2014) study on uropathogenic *E. coli* (UPEC) sequence type (ST)131 served as an exemplary publication which includes ExPEC VFs in a GitHub repository[5]. However, the repository only includes single genes from VF operons/gene cluster.

I collected 1,069 VF protein sequences and categorized them into twelve VF classes (Table 3 on the next page). The protein sequences corresponding to a class are stored in separate `FASTA` files and include the locus tag, gene name (if available), accession number (if available), product description, origin *E. coli* strain with serotype and pathotype (and replicon if applicable), and finally the VF class in their header lines (`https://github.com/aleimba/ecoli_VF_collection/tree/master/data`).

In order to make this collection of VFs sustainable, it is hosted on GitHub and includes a detailed `README` manual with file format de-

[4] `https://github.com/`
[5] `https://github.com/BeatsonLab-MicrobialGenomics/VFDB/`

Table 3: VF classes, number of operons/gene clusters, and VF genes included in the `ecoli_VF_collection`.

VF CLASS	OPERON/GENE CLUSTER	GENES
Adhesion & invasion	13	26
Autotransporter (AT, T5SS)	39	40
Chaperone-usher (CU) fimbriae	59	304
Flagella	3	93
Iron uptake	18	87
Serum resistance	12	41
T2SS	3	41
T3SS	2	163
T6SS	5	127
Toxins	42	100
Type 4 pilus	3	44
Other virulence genes	1	3
Total	200	1,069

scriptions, instructions on how to download and use the files, as well as how to contribute (`https://github.com/aleimba/ecoli_VF_collection`) (Leimbach, 2016b). Additionally, there is a tab-delimited description list[6] of each VF with the *E. coli* reference strain and its pathotype, source (VFDB or manually collected), VF class, and respective locus tag or protein_id accession number (which serve as unique SeqIDs). Last but not least, a tab-delimited file is included specifying the source publications from the literature search[7] (Table 4 on the facing page). The `ecoli_VF_collection` is licensed with an open Creative Commons Attribution 4.0 International License (CC BY 4.0) (cc) BY to maximize reuse. For longtime archival the current snapshot of the repository is stored publicly in Zenodo[8] with a citable Digital Object Identifier (DOI) (Leimbach, 2016b):

> LEIMBACH A. 2016. `ecoli_VF_collection`: v0.1 `https://github.com/aleimba/ecoli_VF_collection`. *Zenodo*. DOI: 10.5281/zenodo.56686

[6] `https://github.com/aleimba/ecoli_VF_collection/blob/master/source/ecoli_VF_collection_description.tsv`

[7] `https://github.com/aleimba/ecoli_VF_collection/blob/master/source/source_publications.tsv`

[8] `https://www.zenodo.org/`

Table 4: Source publications for *E. coli* VFs included in the `ecoli_VF_collection`, ordered alphabetically by first author.

PUBLICATION	INCLUDED VFS	COMMENT
Archer et al. (2011)	CU fimbriae; flagella; T2SS; type III secretion system (T3SS); T6SS	Overview of "large structural components"
Bekal et al. (2003)	IPEC VFs	*E. coli* VF pathoarray
Blum and Leitner (2013)	Bovine-associated *E. coli*	
Blum et al. (2015)	Bovine-associated *E. coli*	
Burgos and Beutin (2010)	Toxins	Differences in EHEC and UPEC α-hemolysins
Chaudhuri et al. (2010)	AT; CU fimbriae; iron uptake	
Clermont et al. (2011)	ExPEC VFs	
Crossman et al. (2010)	CU fimbriae; enterotoxigenic *E. coli* (ETEC) VFs	
Croxen and Finlay (2010)	Review on pathogenic *E. coli* (mostly IPEC)	
Croxen et al. (2013)	Review on pathogenic *E. coli* (mostly IPEC)	
Dogan et al. (2006)	Bovine-associated *E. coli*	
Dogan et al. (2012)	Bovine-associated *E. coli*	VF microarray and PCR
Fairbrother et al. (2015)	Bovine-associated *E. coli*	

Continued on next page.

Table 4: Source publications for *E. coli* VFs included in the `ecoli_VF_collection`, ordered alphabetically by first author (continued).

PUBLICATION	INCLUDED VFS	COMMENT
Garcia et al. (2011)	Iron uptake	Locus tags of iron receptors
Ghanbarpour and Oswald (2010)	Bovine-associated *E. coli*	
Huja et al. (2015)	ExPEC VFs; iron uptake	
Hwang et al. (2007)	Serum resistance	Omptins *ompP* and *ompT*
Ideses et al. (2005)	T3SS	ETT2
Joensen et al. (2014)	IPEC VFs	`VirulenceFinder`
Johnson and Stell (2000)	ExPEC VFs	Basis for many UPEC VF PCRs
Johnson et al. (2006a)	Toxins	Plasmid pAPEC-O1-ColBM with ColB/M colicins
Johnson et al. (2006b)	Toxins	Plasmid pAPEC-O2-ColV with a ColV colicin
Johnson et al. (2008b)	ExPEC VFs	
Johnson et al. (2008a)	ExPEC VFs	
Kaipainen et al. (2002)	Bovine-associated *E. coli*	
Kaper et al. (2004)	Review on pathogenic *E. coli* (mostly IPEC)	
Kempf et al. (2016)	Bovine-associated *E. coli*	

Continued on next page.

Table 4: Source publications for *E. coli* VFs included in the `ecoli_VF_collection`, ordered alphabetically by first author (continued).

PUBLICATION	INCLUDED VFS	COMMENT
Korea et al. (2010)	CU fimbriae	
Köhler and Dobrindt (2011)	ExPEC VFs	
Li et al. (2015)	T6SS	`SecReT6` T6SS database
Ma et al. (2013)	T6SS	
Moulin-Schouleur et al. (2007)	ExPEC VFs	
Müller et al. (2007)	IPEC VFs	Multiplex PCR schema to discern IPEC pathotypes
Nyholm et al. (2015)	IPEC VFs	
Olesen et al. (2012)	ExPEC VFs	
Petty et al. (2014)	ExPEC VFs	Includes tool `SeqFindr`
Ren et al. (2004)	T3SS	ETT2
Ren et al. (2005)	Flagella	Flag-2
Rijavec et al. (2008)	ExPEC VFs	
Rodriguez-Siek et al. (2005)	ExPEC VFs	
Salipante et al. (2015)	ExPEC VFs	Large VF panel from VFDB and literature

Continued on next page.

Table 4: Source publications for *E. coli* VFs included in the `ecoli_VF_collection`, ordered alphabetically by first author (continued).

PUBLICATION	INCLUDED VFS	COMMENT
Schneider et al. (2004)	Adhesion & invasion; serum resistance	Describes pathogenic GI V of UPEC 536
Suojala et al. (2011)	Bovine-associated *E. coli*	
Sváb et al. (2013b)	Adhesion & invasion; bovine-associated *E. coli*	Long polar fimbriae variants
Torres et al. (2009)	Adhesion & invasion	Long polar fimbriae variants
Whitfield and Roberts (1999)	Serum resistance	Definition of the *E. coli* capsule groups
Wurpel et al. (2013)	CU fimbriae	
Zude et al. (2014)	ATs	Section 5.1 on page 79

4 BACTERIAL GENOMICS SCRIPTS

For the bioinformatical analyses of the publications I co-authored during this thesis (see page xi), several workflows and scripts had to be established and written. Thus, I created a collection of `Perl`[1] scripts for bacterial genomics (Table 5 on the following page) and hosted the source code on GitHub (`https://github.com/aleimba/bac-genomics-scripts`). Some of these tools include `bash` wrappers to easily run the whole pipeline and/or call upon the statistical computing language `R`[2] (R Core Team, 2017). Also, each has an individual and detailed `README` file that describes the purpose and usage, lists all options, dependencies (mostly the `BioPerl` module collection (Stajich et al., 2002)), outputs, gives examples, and furthermore includes a changelog for each version. The scripts follow the recommendations for usable bioinformatics command line software by Seemann (2013), such as including a help text with option `-h` and a version switch `-v`.

`bac-genomics-scripts` *collection*

The main `README` gives a summary for each tool/pipeline, installation recommendations, run tips, and email contact (besides the GitHub issue system). All scripts are licensed under the open source copyleft GNU General Public License v3.0 (GPLv3) for a great variety of permissions and reuse/distribution possibilities, and to ensure the work remains freely available. The current version of the scripts is publicly archived in Zenodo with a citable DOI (Leimbach, 2016a):

> LEIMBACH A. 2016. `bac-genomics-scripts`: Bovine *E. coli* mastitis comparative genomics edition `https://github.com/aleimba/bac-genomics-scripts`. *Zenodo*.
> DOI: 10.5281/zenodo.215824

[1] `https://www.perl.org/`
[2] `http://www.r-project.org/`

Table 5: Scripts and pipelines currently in the `bac-genomics-scripts` collection.

SCRIPT/PIPELINE	VERSION	DESCRIPTION
`calc_fastq-stats`	0.1	Calculate basic statistics for bases and reads in `FASTQ` files
`cat_seq`	0.1	Concatenate a multi-sequence file (`EMBL`, `GENBANK` . . . format) to a single artificial file
`cdd2cog`	0.2	Assign Clusters of Orthologous Groups (COG) categories to proteins with `RPS-BLAST+` and NCBI's Conserved Domain Database (CDD) (Marchler-Bauer et al., 2015; Tatusov et al., 2003)
`cds_extractor`	0.7.1	Extract CDS protein/nucleotide sequences from `EMBL`/`GENBANK` files
`ecoli_mlst`	0.3	Determine STs and extract alleles according to Achtman's *E. coli* multi-locus sequence typing (MLST) scheme with `NUCmer` (Kurtz et al., 2004; Wirth et al., 2006)
`genomes_feature_table`	0.5	Create a feature table for `EMBL` or `GENBANK` files
`ncbi_ftp_download` pipeline[3]	0.2.1	Batch downloading of bacterial genomes for a genus/species from NCBI's FTP server; **deprecated**
`order_fastx`	0.1	Order sequence entries in `FASTA`/`FASTQ` files according to an ID list
`po2anno`	0.2.2	Create an annotation comparison matrix based on `Proteinortho5` orthologous/paralogous protein detection (Lechner et al., 2011, 2014)

Continued on next page.

[3] This pipeline is deprecated because NCBI reorganized its FTP server.

Table 5: Scripts and pipelines currently in the bac-genomics-scripts collection (continued).

SCRIPT/PIPELINE	VERSION	DESCRIPTION
po2group_stats	0.1.3	Calculate pan-genome-wide association statistics from Proteinortho5 ortholog/paralog analysis and plot Venn diagrams with R package gplots (Warnes et al., 2016)
prot_finder pipeline	0.7.1	BLASTP presence/absence matrix and pan-genome-wide association (plus Venn diagrams) of query proteins in genomes
rename_fasta_id	0.1	Rename and enumerate FASTA ID lines
revcom_seq	0.2	Reverse complement (multi-)sequence files
rod_finder pipeline	0.4	Regions of difference (RODs) detection between a query genome and reference genome(s) with BLASTN
sam_insert-size	0.2	Paired-end library insert size estimation and read length statistics from BAM/SAM files
sample_fastx-txt	0.1	Randomly subsample FASTA, FASTQ, or TEXT files with *reservoir sampling*
seq_format-converter	0.2	Convert a sequence file to another format with BioPerl
tbl2tab	0.2	Convert NCBI's TBL format[4] to tab-delimited and back, e. g. for manual annotation curation
trunc_seq	0.2	Truncate sequence files according to given coordinates, while retaining annotations

[4] https://www.ncbi.nlm.nih.gov/genbank/genomesubmit_annotation/

Part III

PUBLICATIONS

5 PUBLICATIONS

This chapter includes all publications in which I participated as co-author and that relate to the PhD thesis topic. Original publisher PDF pages or supplementary material are indicated by frames enclosing the respective pages. For a full list of publications co-authored in the course of this thesis please refer to page xi.

5.1 AUTOTRANSPORTER PREVALENCE IN *E. COLI*

Many autotransporter (AT) proteins have been characterized for *E. coli* and their functional domains associated with *virulence* for the majority of these (Section 1.3 on page 46). Because of this association, several ATs were regarded as *markers* for *E. coli pathotypes* that support their individual pathogenesis mechanism. In order to examine this putative pathotype association we set out to analyze the distribution of AT proteins in *E. coli* genomes in regard to *pathotype* and *phylogroup* classification.

5.1.1 *Prevalence of autotransporters in* Escherichia coli*: what is the impact of phylogeny and pathotype?*

Zude I*, LEIMBACH A*, Dobrindt U. 2014. Prevalence of autotransporters in *Escherichia coli*: what is the impact of phylogeny and pathotype? *Int. J. Med. Microbiol.* 304:243–256.
DOI: 10.1016/j.ijmm.2013.10.006

* *Authors contributed equally*

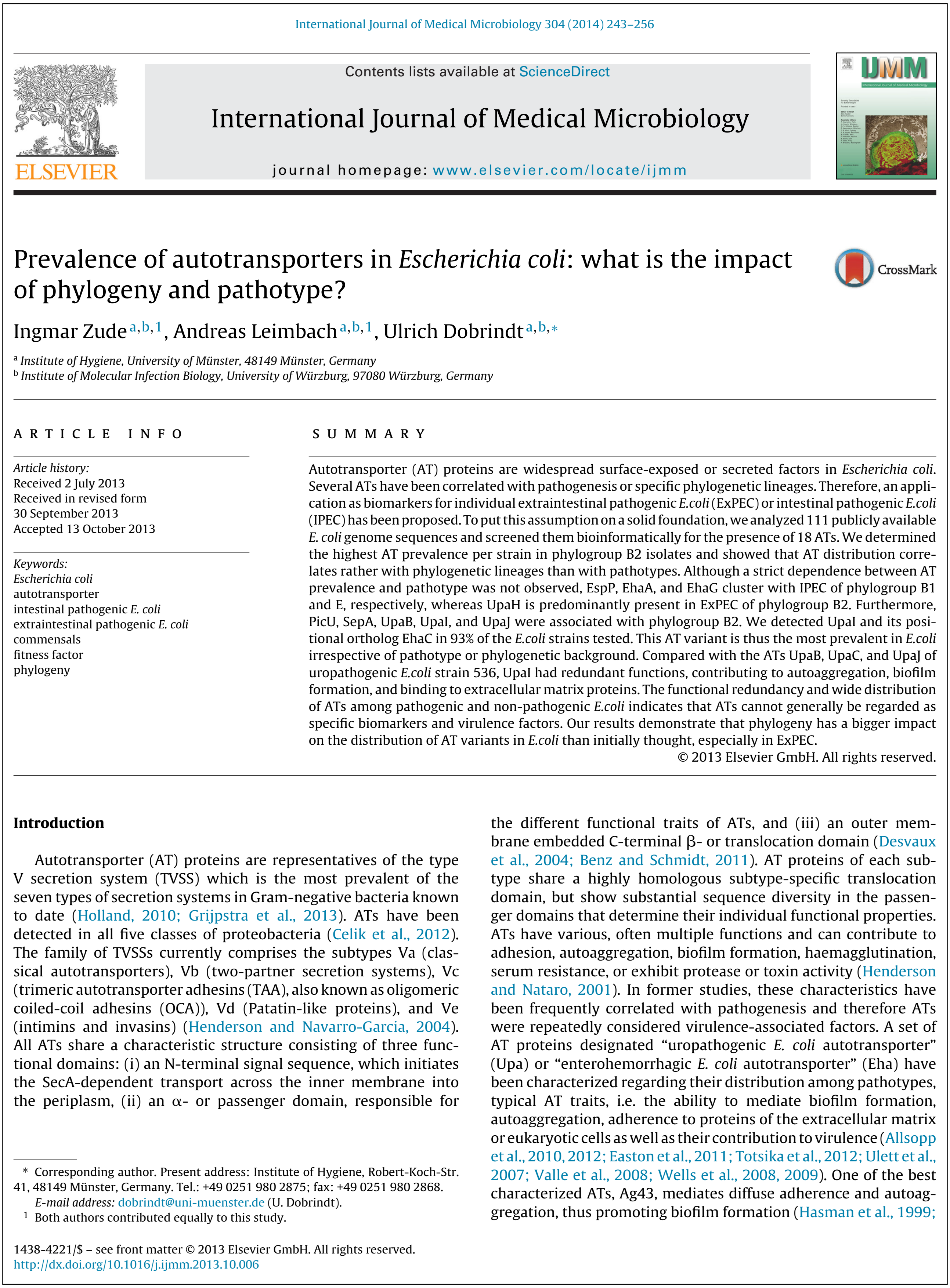
5.1.1.1 *Contributions*

Zude et al. (2014) describes presence/absence distribution of *E. coli* AT proteins (Section 1.3 on page 46) in a large *E. coli* strain panel in relation to phylogeny and pathotype. Additionally, several newly identified ExPEC ATs were phenotypically characterized.

My contribution constitutes all the bioinformatical research and study design of the publication, the respective figures/tables, as well as the accompanying statistical analyses. The corresponding scripts/pipelines I programmed and used for this work are outlined in Chapter 4 on page 73. I wrote all parts of the manuscript that touch upon this analysis. The wet-lab part of the publication was executed and written by Ingmar Zude. Detailed individual author contributions for each part of the paper and each figure/table can be found in Table 10 and Table 11 on page 230, respectively.

5.1.1.2 *Main paper*

Reprinted from Zude et al. (2014) with permission from Elsevier. The publication can be found on pages 81–94 or at:
http://www.sciencedirect.com/science/article/pii/S1438422113001562



International Journal of Medical Microbiology 304 (2014) 243–256

Contents lists available at ScienceDirect

International Journal of Medical Microbiology

journal homepage: www.elsevier.com/locate/ijmm

Prevalence of autotransporters in *Escherichia coli*: what is the impact of phylogeny and pathotype?

Ingmar Zude[a,b,1], Andreas Leimbach[a,b,1], Ulrich Dobrindt[a,b,*]

[a] *Institute of Hygiene, University of Münster, 48149 Münster, Germany*
[b] *Institute of Molecular Infection Biology, University of Würzburg, 97080 Würzburg, Germany*

ARTICLE INFO

Article history:
Received 2 July 2013
Received in revised form
30 September 2013
Accepted 13 October 2013

Keywords:
Escherichia coli
autotransporter
intestinal pathogenic *E. coli*
extraintestinal pathogenic *E. coli*
commensals
fitness factor
phylogeny

SUMMARY

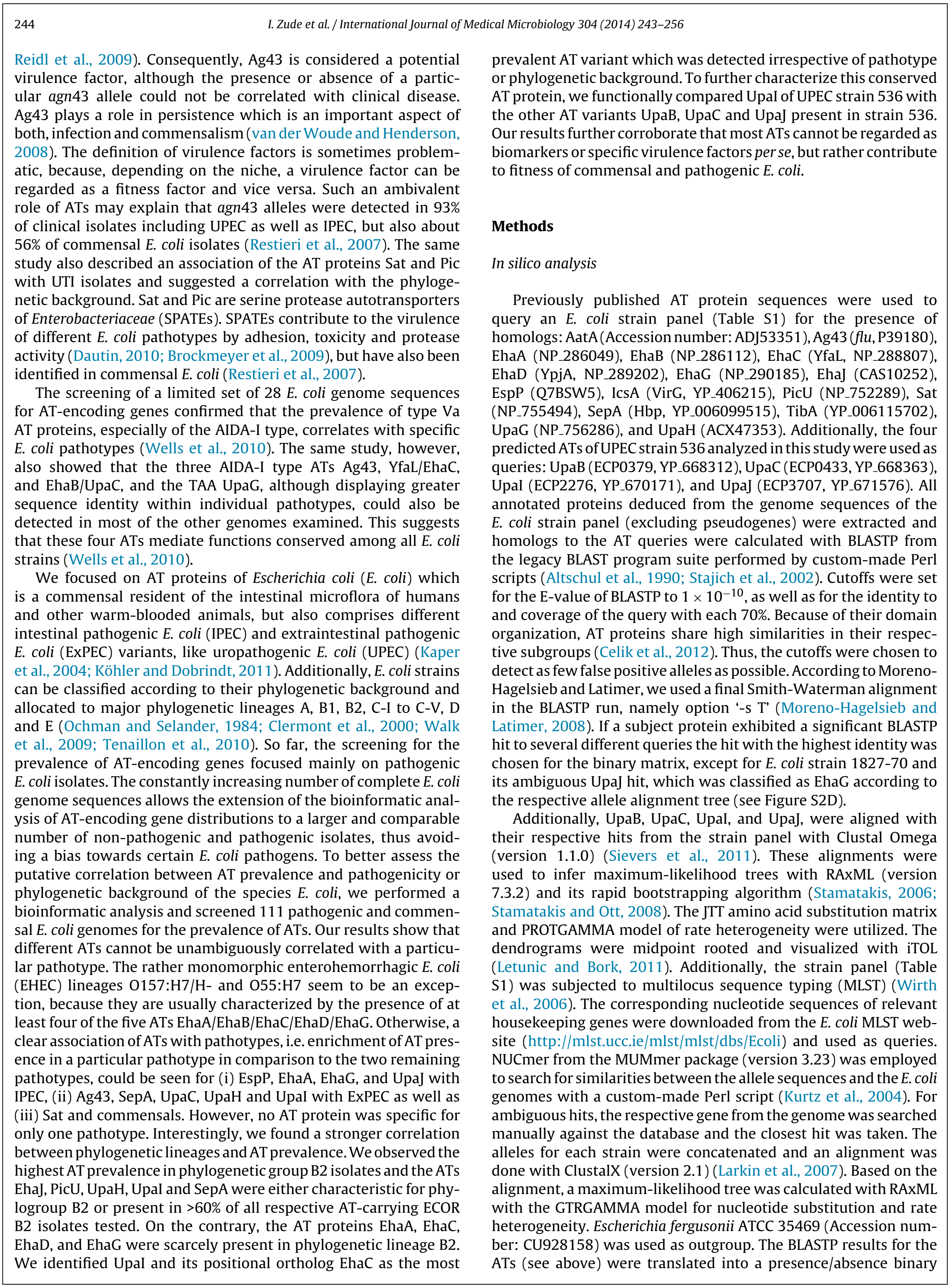
Autotransporter (AT) proteins are widespread surface-exposed or secreted factors in *Escherichia coli*. Several ATs have been correlated with pathogenesis or specific phylogenetic lineages. Therefore, an application as biomarkers for individual extraintestinal pathogenic *E.coli* (ExPEC) or intestinal pathogenic *E.coli* (IPEC) has been proposed. To put this assumption on a solid foundation, we analyzed 111 publicly available *E. coli* genome sequences and screened them bioinformatically for the presence of 18 ATs. We determined the highest AT prevalence per strain in phylogroup B2 isolates and showed that AT distribution correlates rather with phylogenetic lineages than with pathotypes. Although a strict dependence between AT prevalence and pathotype was not observed, EspP, EhaA, and EhaG cluster with IPEC of phylogroup B1 and E, respectively, whereas UpaH is predominantly present in ExPEC of phylogroup B2. Furthermore, PicU, SepA, UpaB, UpaI, and UpaJ were associated with phylogroup B2. We detected UpaI and its positional ortholog EhaC in 93% of the *E.coli* strains tested. This AT variant is thus the most prevalent in *E.coli* irrespective of pathotype or phylogenetic background. Compared with the ATs UpaB, UpaC, and UpaJ of uropathogenic *E.coli* strain 536, UpaI had redundant functions, contributing to autoaggregation, biofilm formation, and binding to extracellular matrix proteins. The functional redundancy and wide distribution of ATs among pathogenic and non-pathogenic *E.coli* indicates that ATs cannot generally be regarded as specific biomarkers and virulence factors. Our results demonstrate that phylogeny has a bigger impact on the distribution of AT variants in *E.coli* than initially thought, especially in ExPEC.

Introduction

Autotransporter (AT) proteins are representatives of the type V secretion system (TVSS) which is the most prevalent of the seven types of secretion systems in Gram-negative bacteria known to date (Holland, 2010; Grijpstra et al., 2013). ATs have been detected in all five classes of proteobacteria (Celik et al., 2012). The family of TVSSs currently comprises the subtypes Va (classical autotransporters), Vb (two-partner secretion systems), Vc (trimeric autotransporter adhesins (TAA), also known as oligomeric coiled-coil adhesins (OCA)), Vd (Patatin-like proteins), and Ve (intimins and invasins) (Henderson and Navarro-Garcia, 2004). All ATs share a characteristic structure consisting of three functional domains: (i) an N-terminal signal sequence, which initiates the SecA-dependent transport across the inner membrane into the periplasm, (ii) an α- or passenger domain, responsible for the different functional traits of ATs, and (iii) an outer membrane embedded C-terminal β- or translocation domain (Desvaux et al., 2004; Benz and Schmidt, 2011). AT proteins of each subtype share a highly homologous subtype-specific translocation domain, but show substantial sequence diversity in the passenger domains that determine their individual functional properties. ATs have various, often multiple functions and can contribute to adhesion, autoaggregation, biofilm formation, haemagglutination, serum resistance, or exhibit protease or toxin activity (Henderson and Nataro, 2001). In former studies, these characteristics have been frequently correlated with pathogenesis and therefore ATs were repeatedly considered virulence-associated factors. A set of AT proteins designated “uropathogenic *E. coli* autotransporter” (Upa) or “enterohemorrhagic *E. coli* autotransporter” (Eha) have been characterized regarding their distribution among pathotypes, typical AT traits, i.e. the ability to mediate biofilm formation, autoaggregation, adherence to proteins of the extracellular matrix or eukaryotic cells as well as their contribution to virulence (Allsopp et al., 2010, 2012; Easton et al., 2011; Totsika et al., 2012; Ulett et al., 2007; Valle et al., 2008; Wells et al., 2008, 2009). One of the best characterized ATs, Ag43, mediates diffuse adherence and autoaggregation, thus promoting biofilm formation (Hasman et al., 1999;

[*] Corresponding author. Present address: Institute of Hygiene, Robert-Koch-Str. 41, 48149 Münster, Germany. Tel.: +49 0251 980 2875; fax: +49 0251 980 2868.
E-mail address: dobrindt@uni-muenster.de (U. Dobrindt).
[1] Both authors contributed equally to this study.

1438-4221/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.ijmm.2013.10.006



Reidl et al., 2009). Consequently, Ag43 is considered a potential virulence factor, although the presence or absence of a particular *agn*43 allele could not be correlated with clinical disease. Ag43 plays a role in persistence which is an important aspect of both, infection and commensalism (van der Woude and Henderson, 2008). The definition of virulence factors is sometimes problematic, because, depending on the niche, a virulence factor can be regarded as a fitness factor and vice versa. Such an ambivalent role of ATs may explain that *agn*43 alleles were detected in 93% of clinical isolates including UPEC as well as IPEC, but also about 56% of commensal *E. coli* isolates (Restieri et al., 2007). The same study also described an association of the AT proteins Sat and Pic with UTI isolates and suggested a correlation with the phylogenetic background. Sat and Pic are serine protease autotransporters of *Enterobacteriaceae* (SPATEs). SPATEs contribute to the virulence of different *E. coli* pathotypes by adhesion, toxicity and protease activity (Dautin, 2010; Brockmeyer et al., 2009), but have also been identified in commensal *E. coli* (Restieri et al., 2007).

The screening of a limited set of 28 *E. coli* genome sequences for AT-encoding genes confirmed that the prevalence of type Va AT proteins, especially of the AIDA-I type, correlates with specific *E. coli* pathotypes (Wells et al., 2010). The same study, however, also showed that the three AIDA-I type ATs Ag43, YfaL/EhaC, and EhaB/UpaC, and the TAA UpaG, although displaying greater sequence identity within individual pathotypes, could also be detected in most of the other genomes examined. This suggests that these four ATs mediate functions conserved among all *E. coli* strains (Wells et al., 2010).

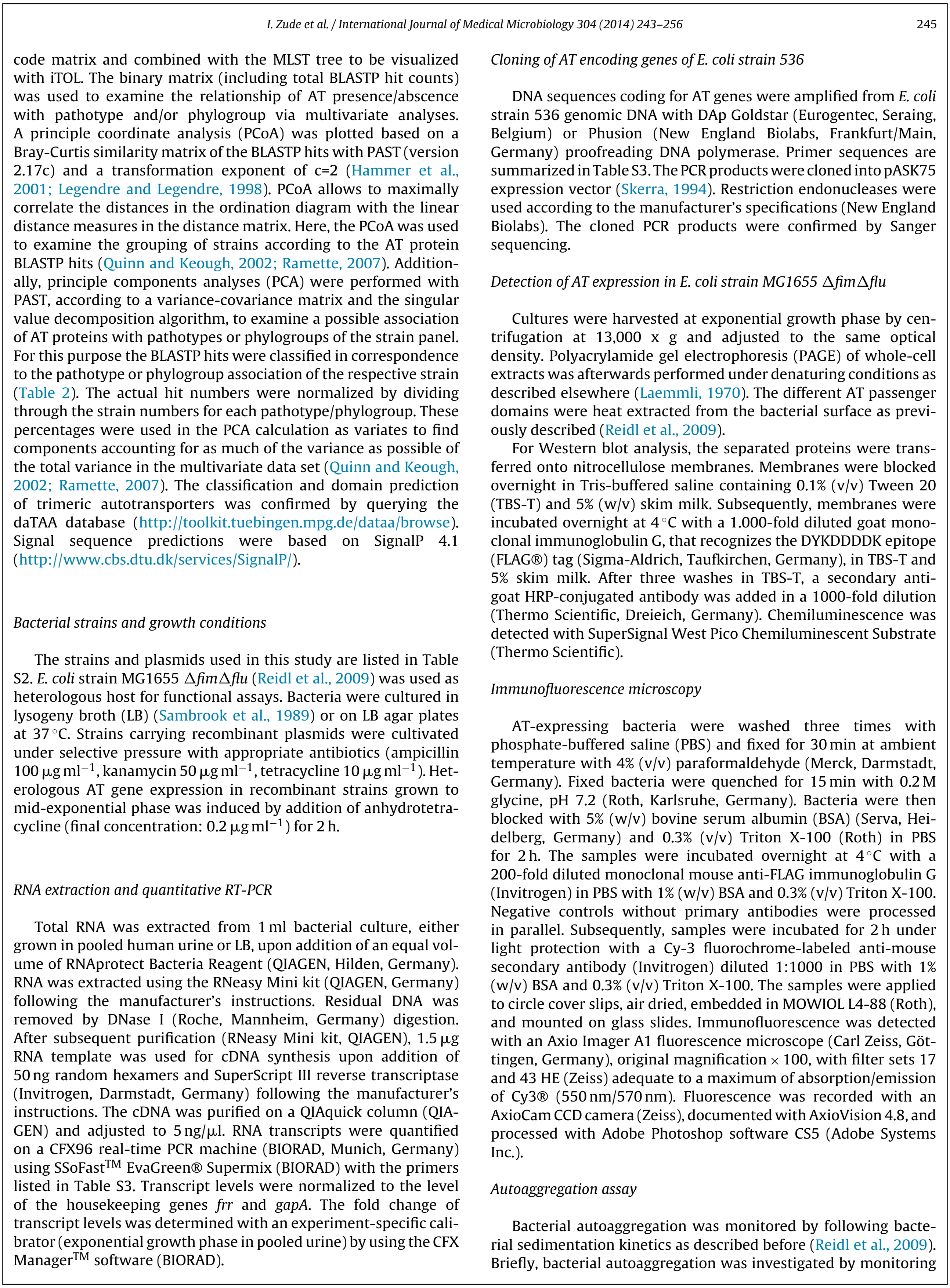
We focused on AT proteins of *Escherichia coli* (*E. coli*) which is a commensal resident of the intestinal microflora of humans and other warm-blooded animals, but also comprises different intestinal pathogenic *E. coli* (IPEC) and extraintestinal pathogenic *E. coli* (ExPEC) variants, like uropathogenic *E. coli* (UPEC) (Kaper et al., 2004; Köhler and Dobrindt, 2011). Additionally, *E. coli* strains can be classified according to their phylogenetic background and allocated to major phylogenetic lineages A, B1, B2, C-I to C-V, D and E (Ochman and Selander, 1984; Clermont et al., 2000; Walk et al., 2009; Tenaillon et al., 2010). So far, the screening for the prevalence of AT-encoding genes focused mainly on pathogenic *E. coli* isolates. The constantly increasing number of complete *E. coli* genome sequences allows the extension of the bioinformatic analysis of AT-encoding gene distributions to a larger and comparable number of non-pathogenic and pathogenic isolates, thus avoiding a bias towards certain *E. coli* pathogens. To better assess the putative correlation between AT prevalence and pathogenicity or phylogenetic background of the species *E. coli*, we performed a bioinformatic analysis and screened 111 pathogenic and commensal *E. coli* genomes for the prevalence of ATs. Our results show that different ATs cannot be unambiguously correlated with a particular pathotype. The rather monomorphic enterohemorrhagic *E. coli* (EHEC) lineages O157:H7/H- and O55:H7 seem to be an exception, because they are usually characterized by the presence of at least four of the five ATs EhaA/EhaB/EhaC/EhaD/EhaG. Otherwise, a clear association of ATs with pathotypes, i.e. enrichment of AT presence in a particular pathotype in comparison to the two remaining pathotypes, could be seen for (i) EspP, EhaA, EhaG, and UpaJ with IPEC, (ii) Ag43, SepA, UpaC, UpaH and UpaI with ExPEC as well as (iii) Sat and commensals. However, no AT protein was specific for only one pathotype. Interestingly, we found a stronger correlation between phylogenetic lineages and AT prevalence. We observed the highest AT prevalence in phylogenetic group B2 isolates and the ATs EhaJ, PicU, UpaH, UpaI and SepA were either characteristic for phylogroup B2 or present in >60% of all respective AT-carrying ECOR B2 isolates tested. On the contrary, the AT proteins EhaA, EhaC, EhaD, and EhaG were scarcely present in phylogenetic lineage B2. We identified UpaI and its positional ortholog EhaC as the most prevalent AT variant which was detected irrespective of pathotype or phylogenetic background. To further characterize this conserved AT protein, we functionally compared UpaI of UPEC strain 536 with the other AT variants UpaB, UpaC and UpaJ present in strain 536. Our results further corroborate that most ATs cannot be regarded as biomarkers or specific virulence factors *per se*, but rather contribute to fitness of commensal and pathogenic *E. coli*.

Methods

In silico analysis

Previously published AT protein sequences were used to query an *E. coli* strain panel (Table S1) for the presence of homologs: AatA (Accession number: ADJ53351), Ag43 (*flu*, P39180), EhaA (NP_286049), EhaB (NP_286112), EhaC (YfaL, NP_288807), EhaD (YpjA, NP_289202), EhaG (NP_290185), EhaJ (CAS10252), EspP (Q7BSW5), IcsA (VirG, YP_406215), PicU (NP_752289), Sat (NP_755494), SepA (Hbp, YP_006099515), TibA (YP_006115702), UpaG (NP_756286), and UpaH (ACX47353). Additionally, the four predicted ATs of UPEC strain 536 analyzed in this study were used as queries: UpaB (ECP0379, YP_668312), UpaC (ECP0433, YP_668363), UpaI (ECP2276, YP_670171), and UpaJ (ECP3707, YP_671576). All annotated proteins deduced from the genome sequences of the *E. coli* strain panel (excluding pseudogenes) were extracted and homologs to the AT queries were calculated with BLASTP from the legacy BLAST program suite performed by custom-made Perl scripts (Altschul et al., 1990; Stajich et al., 2002). Cutoffs were set for the E-value of BLASTP to 1×10^{-10}, as well as for the identity to and coverage of the query with each 70%. Because of their domain organization, AT proteins share high similarities in their respective subgroups (Celik et al., 2012). Thus, the cutoffs were chosen to detect as few false positive alleles as possible. According to Moreno-Hagelsieb and Latimer, we used a final Smith-Waterman alignment in the BLASTP run, namely option '-s T' (Moreno-Hagelsieb and Latimer, 2008). If a subject protein exhibited a significant BLASTP hit to several different queries the hit with the highest identity was chosen for the binary matrix, except for *E. coli* strain 1827-70 and its ambiguous UpaJ hit, which was classified as EhaG according to the respective allele alignment tree (see Figure S2D).

Additionally, UpaB, UpaC, UpaI, and UpaJ, were aligned with their respective hits from the strain panel with Clustal Omega (version 1.1.0) (Sievers et al., 2011). These alignments were used to infer maximum-likelihood trees with RAxML (version 7.3.2) and its rapid bootstrapping algorithm (Stamatakis, 2006; Stamatakis and Ott, 2008). The JTT amino acid substitution matrix and PROTGAMMA model of rate heterogeneity were utilized. The dendrograms were midpoint rooted and visualized with iTOL (Letunic and Bork, 2011). Additionally, the strain panel (Table S1) was subjected to multilocus sequence typing (MLST) (Wirth et al., 2006). The corresponding nucleotide sequences of relevant housekeeping genes were downloaded from the *E. coli* MLST website (http://mlst.ucc.ie/mlst/mlst/dbs/Ecoli) and used as queries. NUCmer from the MUMmer package (version 3.23) was employed to search for similarities between the allele sequences and the *E. coli* genomes with a custom-made Perl script (Kurtz et al., 2004). For ambiguous hits, the respective gene from the genome was searched manually against the database and the closest hit was taken. The alleles for each strain were concatenated and an alignment was done with ClustalX (version 2.1) (Larkin et al., 2007). Based on the alignment, a maximum-likelihood tree was calculated with RAxML with the GTRGAMMA model for nucleotide substitution and rate heterogeneity. *Escherichia fergusonii* ATCC 35469 (Accession number: CU928158) was used as outgroup. The BLASTP results for the ATs (see above) were translated into a presence/absence binary



code matrix and combined with the MLST tree to be visualized with iTOL. The binary matrix (including total BLASTP hit counts) was used to examine the relationship of AT presence/abscence with pathotype and/or phylogroup via multivariate analyses. A principle coordinate analysis (PCoA) was plotted based on a Bray-Curtis similarity matrix of the BLASTP hits with PAST (version 2.17c) and a transformation exponent of c=2 (Hammer et al., 2001; Legendre and Legendre, 1998). PCoA allows to maximally correlate the distances in the ordination diagram with the linear distance measures in the distance matrix. Here, the PCoA was used to examine the grouping of strains according to the AT protein BLASTP hits (Quinn and Keough, 2002; Ramette, 2007). Additionally, principle components analyses (PCA) were performed with PAST, according to a variance-covariance matrix and the singular value decomposition algorithm, to examine a possible association of AT proteins with pathotypes or phylogroups of the strain panel. For this purpose the BLASTP hits were classified in correspondence to the pathotype or phylogroup association of the respective strain (Table 2). The actual hit numbers were normalized by dividing through the strain numbers for each pathotype/phylogroup. These percentages were used in the PCA calculation as variates to find components accounting for as much of the variance as possible of the total variance in the multivariate data set (Quinn and Keough, 2002; Ramette, 2007). The classification and domain prediction of trimeric autotransporters was confirmed by querying the daTAA database (http://toolkit.tuebingen.mpg.de/dataa/browse). Signal sequence predictions were based on SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/).

Bacterial strains and growth conditions

The strains and plasmids used in this study are listed in Table S2. *E. coli* strain MG1655 Δ*fim*Δ*flu* (Reidl et al., 2009) was used as heterologous host for functional assays. Bacteria were cultured in lysogeny broth (LB) (Sambrook et al., 1989) or on LB agar plates at 37 °C. Strains carrying recombinant plasmids were cultivated under selective pressure with appropriate antibiotics (ampicillin 100 μg ml^{-1}, kanamycin 50 μg ml^{-1}, tetracycline 10 μg ml^{-1}). Heterologous AT gene expression in recombinant strains grown to mid-exponential phase was induced by addition of anhydrotetracycline (final concentration: 0.2 μg ml^{-1}) for 2 h.

RNA extraction and quantitative RT-PCR

Total RNA was extracted from 1 ml bacterial culture, either grown in pooled human urine or LB, upon addition of an equal volume of RNAprotect Bacteria Reagent (QIAGEN, Hilden, Germany). RNA was extracted using the RNeasy Mini kit (QIAGEN, Germany) following the manufacturer's instructions. Residual DNA was removed by DNase I (Roche, Mannheim, Germany) digestion. After subsequent purification (RNeasy Mini kit, QIAGEN), 1.5 μg RNA template was used for cDNA synthesis upon addition of 50 ng random hexamers and SuperScript III reverse transcriptase (Invitrogen, Darmstadt, Germany) following the manufacturer's instructions. The cDNA was purified on a QIAquick column (QIAGEN) and adjusted to 5 ng/μl. RNA transcripts were quantified on a CFX96 real-time PCR machine (BIORAD, Munich, Germany) using SSoFast^{TM} EvaGreen® Supermix (BIORAD) with the primers listed in Table S3. Transcript levels were normalized to the level of the housekeeping genes *frr* and *gapA*. The fold change of transcript levels was determined with an experiment-specific calibrator (exponential growth phase in pooled urine) by using the CFX Manager^{TM} software (BIORAD).

Cloning of AT encoding genes of E. coli strain 536

DNA sequences coding for AT genes were amplified from *E. coli* strain 536 genomic DNA with DAp Goldstar (Eurogentec, Seraing, Belgium) or Phusion (New England Biolabs, Frankfurt/Main, Germany) proofreading DNA polymerase. Primer sequences are summarized in Table S3. The PCR products were cloned into pASK75 expression vector (Skerra, 1994). Restriction endonucleases were used according to the manufacturer's specifications (New England Biolabs). The cloned PCR products were confirmed by Sanger sequencing.

Detection of AT expression in E. coli strain MG1655 ΔfimΔflu

Cultures were harvested at exponential growth phase by centrifugation at 13,000 x g and adjusted to the same optical density. Polyacrylamide gel electrophoresis (PAGE) of whole-cell extracts was afterwards performed under denaturing conditions as described elsewhere (Laemmli, 1970). The different AT passenger domains were heat extracted from the bacterial surface as previously described (Reidl et al., 2009).

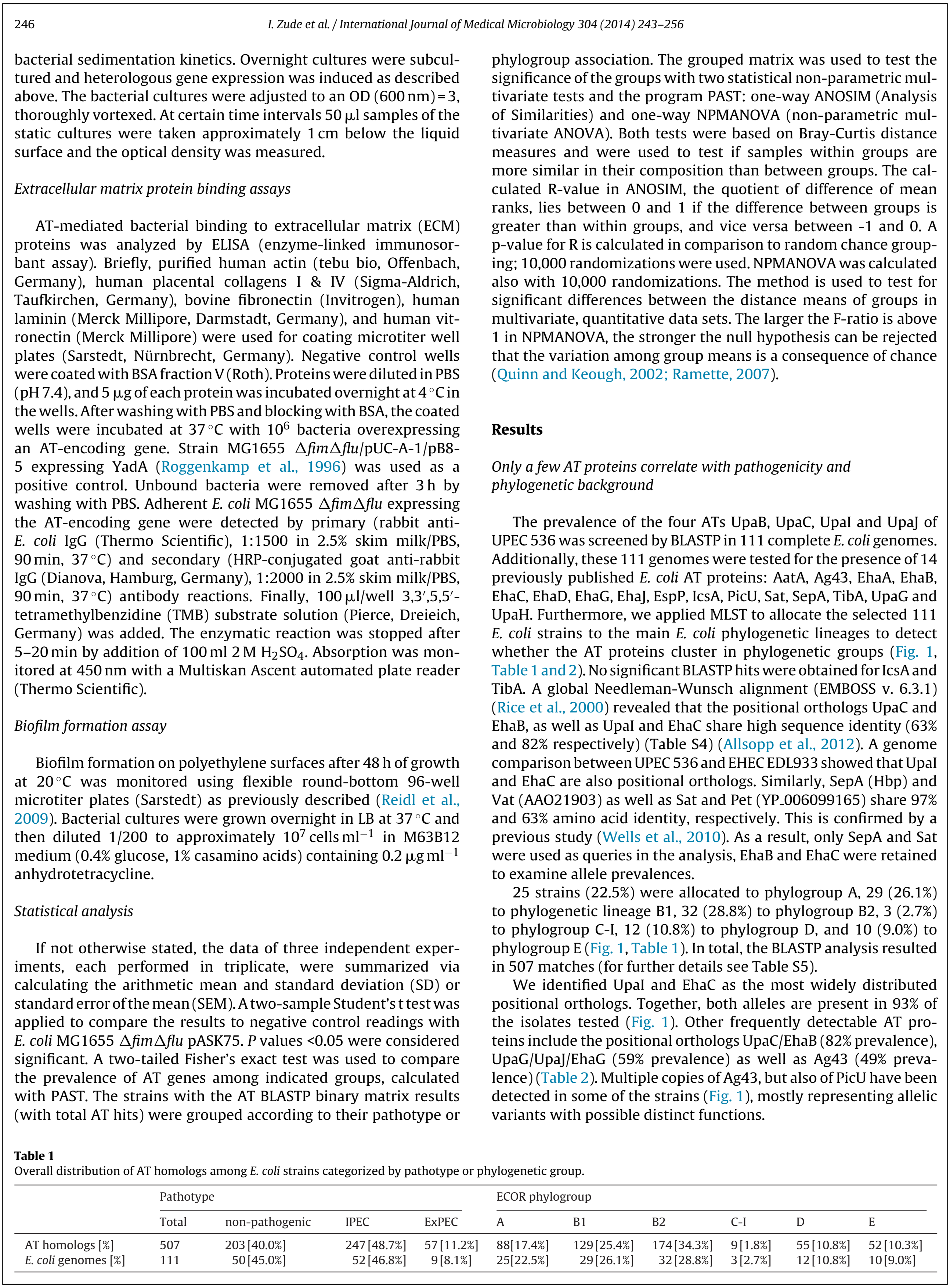
For Western blot analysis, the separated proteins were transferred onto nitrocellulose membranes. Membranes were blocked overnight in Tris-buffered saline containing 0.1% (v/v) Tween 20 (TBS-T) and 5% (w/v) skim milk. Subsequently, membranes were incubated overnight at 4 °C with a 1.000-fold diluted goat monoclonal immunoglobulin G, that recognizes the DYKDDDDK epitope (FLAG®) tag (Sigma-Aldrich, Taufkirchen, Germany), in TBS-T and 5% skim milk. After three washes in TBS-T, a secondary anti-goat HRP-conjugated antibody was added in a 1000-fold dilution (Thermo Scientific, Dreieich, Germany). Chemiluminescence was detected with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific).

Immunofluorescence microscopy

AT-expressing bacteria were washed three times with phosphate-buffered saline (PBS) and fixed for 30 min at ambient temperature with 4% (v/v) paraformaldehyde (Merck, Darmstadt, Germany). Fixed bacteria were quenched for 15 min with 0.2 M glycine, pH 7.2 (Roth, Karlsruhe, Germany). Bacteria were then blocked with 5% (w/v) bovine serum albumin (BSA) (Serva, Heidelberg, Germany) and 0.3% (v/v) Triton X-100 (Roth) in PBS for 2 h. The samples were incubated overnight at 4 °C with a 200-fold diluted monoclonal mouse anti-FLAG immunoglobulin G (Invitrogen) in PBS with 1% (w/v) BSA and 0.3% (v/v) Triton X-100. Negative controls without primary antibodies were processed in parallel. Subsequently, samples were incubated for 2 h under light protection with a Cy-3 fluorochrome-labeled anti-mouse secondary antibody (Invitrogen) diluted 1:1000 in PBS with 1% (w/v) BSA and 0.3% (v/v) Triton X-100. The samples were applied to circle cover slips, air dried, embedded in MOWIOL L4-88 (Roth), and mounted on glass slides. Immunofluorescence was detected with an Axio Imager A1 fluorescence microscope (Carl Zeiss, Göttingen, Germany), original magnification × 100, with filter sets 17 and 43 HE (Zeiss) adequate to a maximum of absorption/emission of Cy3® (550 nm/570 nm). Fluorescence was recorded with an AxioCam CCD camera (Zeiss), documented with AxioVision 4.8, and processed with Adobe Photoshop software CS5 (Adobe Systems Inc.).

Autoaggregation assay

Bacterial autoaggregation was monitored by following bacterial sedimentation kinetics as described before (Reidl et al., 2009). Briefly, bacterial autoaggregation was investigated by monitoring



bacterial sedimentation kinetics. Overnight cultures were subcultured and heterologous gene expression was induced as described above. The bacterial cultures were adjusted to an OD (600 nm) = 3, thoroughly vortexed. At certain time intervals 50 μl samples of the static cultures were taken approximately 1 cm below the liquid surface and the optical density was measured.

Extracellular matrix protein binding assays

AT-mediated bacterial binding to extracellular matrix (ECM) proteins was analyzed by ELISA (enzyme-linked immunosorbant assay). Briefly, purified human actin (tebu bio, Offenbach, Germany), human placental collagens I & IV (Sigma-Aldrich, Taufkirchen, Germany), bovine fibronectin (Invitrogen), human laminin (Merck Millipore, Darmstadt, Germany), and human vitronectin (Merck Millipore) were used for coating microtiter well plates (Sarstedt, Nürnbrecht, Germany). Negative control wells were coated with BSA fraction V (Roth). Proteins were diluted in PBS (pH 7.4), and 5 μg of each protein was incubated overnight at 4 °C in the wells. After washing with PBS and blocking with BSA, the coated wells were incubated at 37 °C with 10^6 bacteria overexpressing an AT-encoding gene. Strain MG1655 Δ*fim*Δ*flu*/pUC-A-1/pB8-5 expressing YadA (Roggenkamp et al., 1996) was used as a positive control. Unbound bacteria were removed after 3 h by washing with PBS. Adherent *E. coli* MG1655 Δ*fim*Δ*flu* expressing the AT-encoding gene were detected by primary (rabbit anti-*E. coli* IgG (Thermo Scientific), 1:1500 in 2.5% skim milk/PBS, 90 min, 37 °C) and secondary (HRP-conjugated goat anti-rabbit IgG (Dianova, Hamburg, Germany), 1:2000 in 2.5% skim milk/PBS, 90 min, 37 °C) antibody reactions. Finally, 100 μl/well 3,3′,5,5′-tetramethylbenzidine (TMB) substrate solution (Pierce, Dreieich, Germany) was added. The enzymatic reaction was stopped after 5–20 min by addition of 100 ml 2 M H_2SO_4. Absorption was monitored at 450 nm with a Multiskan Ascent automated plate reader (Thermo Scientific).

Biofilm formation assay

Biofilm formation on polyethylene surfaces after 48 h of growth at 20 °C was monitored using flexible round-bottom 96-well microtiter plates (Sarstedt) as previously described (Reidl et al., 2009). Bacterial cultures were grown overnight in LB at 37 °C and then diluted 1/200 to approximately 10^7 cells ml^{-1} in M63B12 medium (0.4% glucose, 1% casamino acids) containing 0.2 μg ml^{-1} anhydrotetracycline.

Statistical analysis

If not otherwise stated, the data of three independent experiments, each performed in triplicate, were summarized via calculating the arithmetic mean and standard deviation (SD) or standard error of the mean (SEM). A two-sample Student's t test was applied to compare the results to negative control readings with *E. coli* MG1655 Δ*fim*Δ*flu* pASK75. *P* values <0.05 were considered significant. A two-tailed Fisher's exact test was used to compare the prevalence of AT genes among indicated groups, calculated with PAST. The strains with the AT BLASTP binary matrix results (with total AT hits) were grouped according to their pathotype or phylogroup association. The grouped matrix was used to test the significance of the groups with two statistical non-parametric multivariate tests and the program PAST: one-way ANOSIM (Analysis of Similarities) and one-way NPMANOVA (non-parametric multivariate ANOVA). Both tests were based on Bray-Curtis distance measures and were used to test if samples within groups are more similar in their composition than between groups. The calculated R-value in ANOSIM, the quotient of difference of mean ranks, lies between 0 and 1 if the difference between groups is greater than within groups, and vice versa between -1 and 0. A p-value for R is calculated in comparison to random chance grouping; 10,000 randomizations were used. NPMANOVA was calculated also with 10,000 randomizations. The method is used to test for significant differences between the distance means of groups in multivariate, quantitative data sets. The larger the F-ratio is above 1 in NPMANOVA, the stronger the null hypothesis can be rejected that the variation among group means is a consequence of chance (Quinn and Keough, 2002; Ramette, 2007).

Results

Only a few AT proteins correlate with pathogenicity and phylogenetic background

The prevalence of the four ATs UpaB, UpaC, UpaI and UpaJ of UPEC 536 was screened by BLASTP in 111 complete *E. coli* genomes. Additionally, these 111 genomes were tested for the presence of 14 previously published *E. coli* AT proteins: AatA, Ag43, EhaA, EhaB, EhaC, EhaD, EhaG, EhaJ, EspP, IcsA, PicU, Sat, SepA, TibA, UpaG and UpaH. Furthermore, we applied MLST to allocate the selected 111 *E. coli* strains to the main *E. coli* phylogenetic lineages to detect whether the AT proteins cluster in phylogenetic groups (Fig. 1, Table 1 and 2). No significant BLASTP hits were obtained for IcsA and TibA. A global Needleman-Wunsch alignment (EMBOSS v. 6.3.1) (Rice et al., 2000) revealed that the positional orthologs UpaC and EhaB, as well as UpaI and EhaC share high sequence identity (63% and 82% respectively) (Table S4) (Allsopp et al., 2012). A genome comparison between UPEC 536 and EHEC EDL933 showed that UpaI and EhaC are also positional orthologs. Similarly, SepA (Hbp) and Vat (AAO21903) as well as Sat and Pet (YP_006099165) share 97% and 63% amino acid identity, respectively. This is confirmed by a previous study (Wells et al., 2010). As a result, only SepA and Sat were used as queries in the analysis, EhaB and EhaC were retained to examine allele prevalences.

25 strains (22.5%) were allocated to phylogroup A, 29 (26.1%) to phylogenetic lineage B1, 32 (28.8%) to phylogroup B2, 3 (2.7%) to phylogroup C-I, 12 (10.8%) to phylogroup D, and 10 (9.0%) to phylogroup E (Fig. 1, Table 1). In total, the BLASTP analysis resulted in 507 matches (for further details see Table S5).

We identified UpaI and EhaC as the most widely distributed positional orthologs. Together, both alleles are present in 93% of the isolates tested (Fig. 1). Other frequently detectable AT proteins include the positional orthologs UpaC/EhaB (82% prevalence), UpaG/UpaJ/EhaG (59% prevalence) as well as Ag43 (49% prevalence) (Table 2). Multiple copies of Ag43, but also of PicU have been detected in some of the strains (Fig. 1), mostly representing allelic variants with possible distinct functions.

Table 1
Overall distribution of AT homologs among *E. coli* strains categorized by pathotype or phylogenetic group.

	Pathotype				ECOR phylogroup					
	Total	non-pathogenic	IPEC	ExPEC	A	B1	B2	C-I	D	E
AT homologs [%]	507	203 [40.0%]	247 [48.7%]	57 [11.2%]	88 [17.4%]	129 [25.4%]	174 [34.3%]	9 [1.8%]	55 [10.8%]	52 [10.3%]
E. coli genomes [%]	111	50 [45.0%]	52 [46.8%]	9 [8.1%]	25 [22.5%]	29 [26.1%]	32 [28.8%]	3 [2.7%]	12 [10.8%]	10 [9.0%]

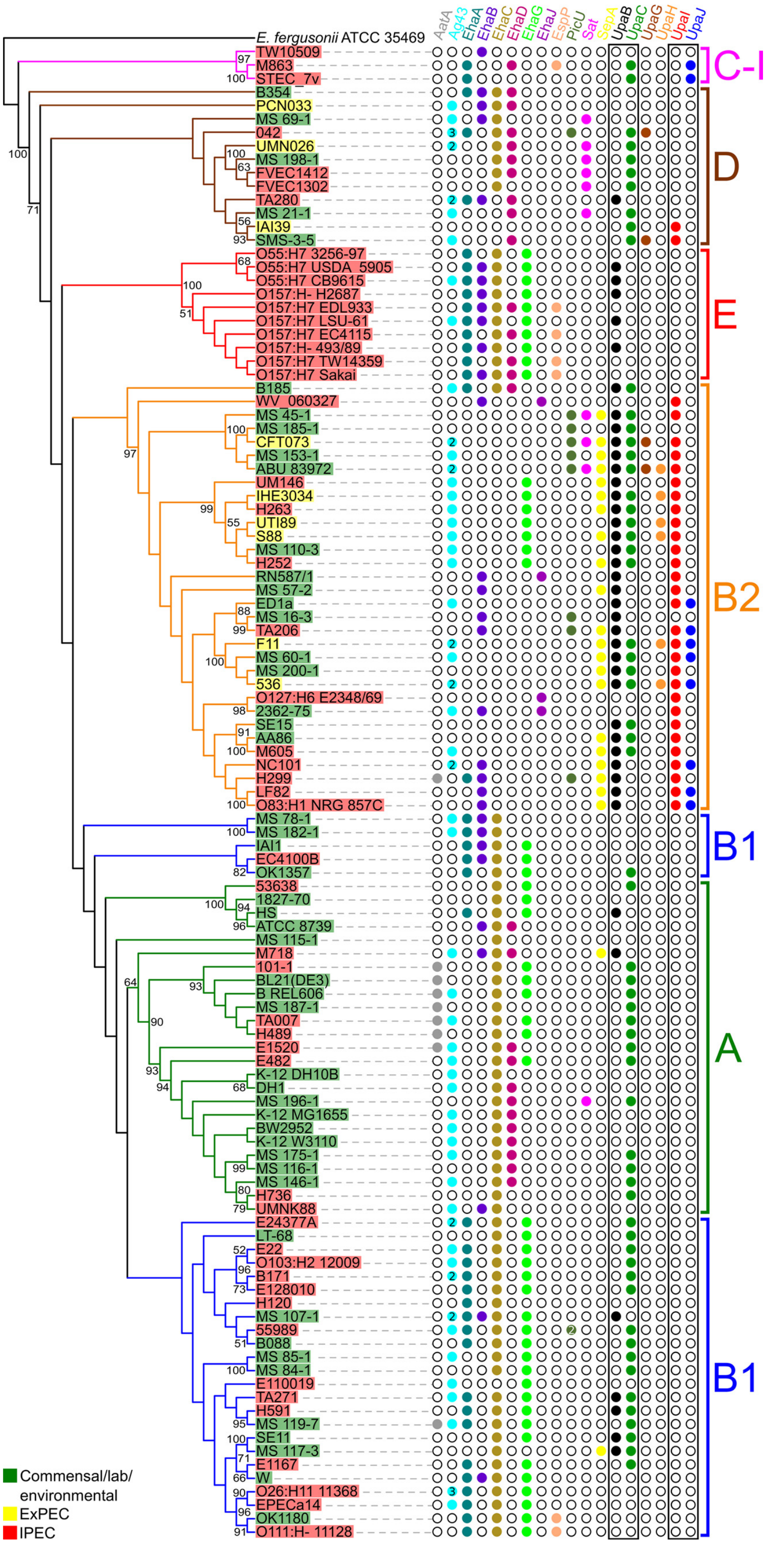


Fig. 1. MLST-based phylogeny of 111 *E. coli* genomes and prevalence of *E. coli* AT proteins assessed by BLASTP. ECOR phylogenetic groups A (green), B1 (blue), B2 (ocher), C-I (violet), D (brown) and E (red) are indicated. Bootstrap values >50 from 1000 resamplings are shown at respective nodes. *Escherichia fergusonii* was used as an outgroup to root the tree. The binary code shows the presence/absence of the respective AT proteins AatA, Ag43, EhaA, EhaB, EhaC, EhaD, EhaG, EhaJ, EspP, PicU, Sat, SepA, UpaB, UpaC, UpaG, UpaH, UpaI, and UpaJ. Digits in the binary code indicate the number of corresponding AT alleles in the respective genome. Further details on the BLASTP results are shown in supplementary (Table S4).

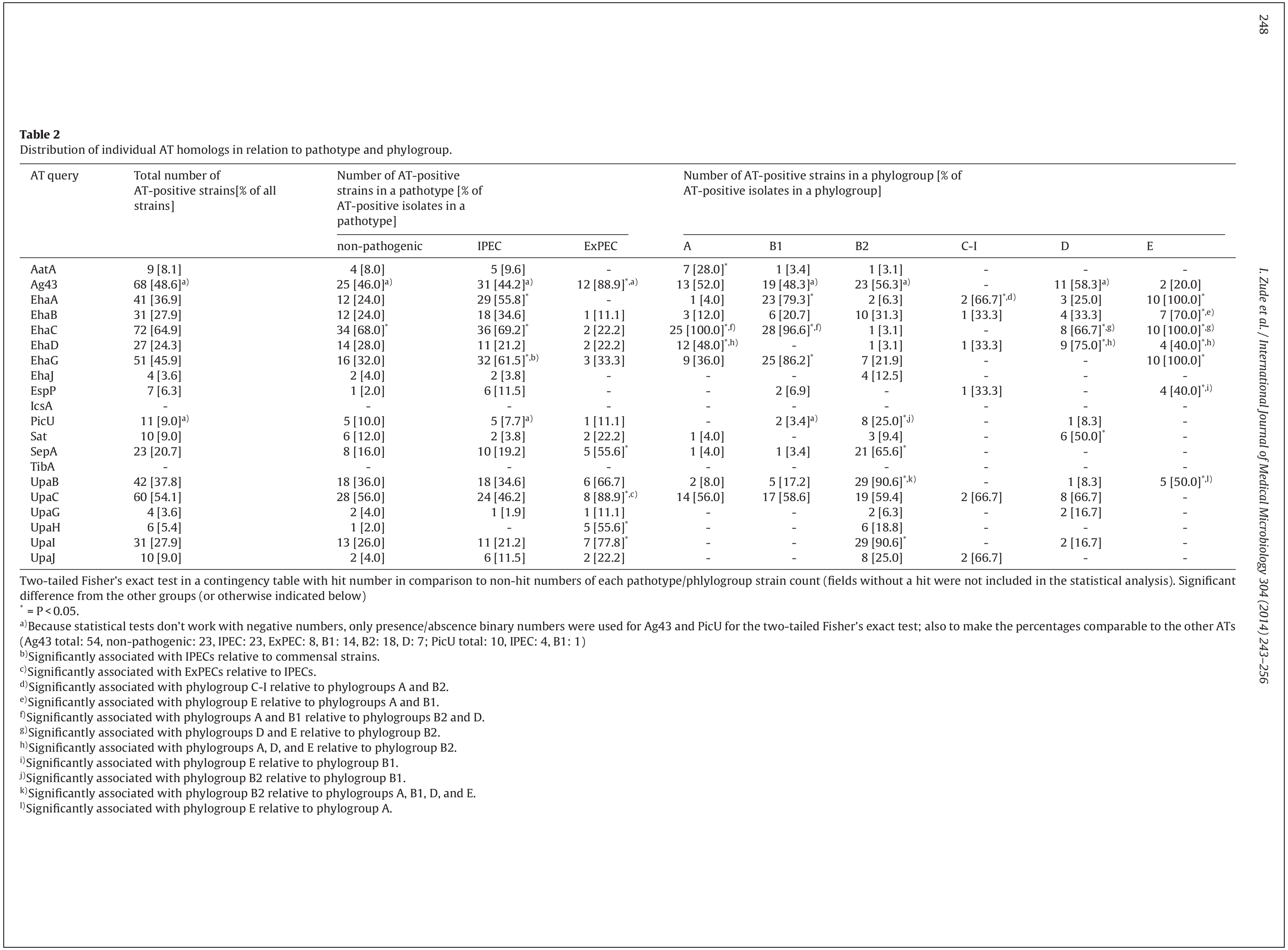


Table 2
Distribution of individual AT homologs in relation to pathotype and phylogroup.

AT query	Total number of AT-positive strains[% of all strains]	Number of AT-positive strains in a pathotype [% of AT-positive isolates in a pathotype]			Number of AT-positive strains in a phylogroup [% of AT-positive isolates in a phylogroup]					
		non-pathogenic	IPEC	ExPEC	A	B1	B2	C-I	D	E
AatA	9 [8.1]	4 [8.0]	5 [9.6]	-	7 [28.0]*	1 [3.4]	1 [3.1]	-	-	-
Ag43	68 [48.6]$^{a)}$	25 [46.0]$^{a)}$	31 [44.2]$^{a)}$	12 [88.9]$^{*,a)}$	13 [52.0]	19 [48.3]$^{a)}$	23 [56.3]$^{a)}$	-	11 [58.3]$^{a)}$	2 [20.0]
EhaA	41 [36.9]	12 [24.0]	29 [55.8]*	-	1 [4.0]	23 [79.3]*	2 [6.3]	2 [66.7]$^{*,d)}$	3 [25.0]	10 [100.0]*
EhaB	31 [27.9]	12 [24.0]	18 [34.6]	1 [11.1]	3 [12.0]	6 [20.7]	10 [31.3]	1 [33.3]	4 [33.3]	7 [70.0]$^{*,e)}$
EhaC	72 [64.9]	34 [68.0]*	36 [69.2]*	2 [22.2]	25 [100.0]$^{*,f)}$	28 [96.6]$^{*,f)}$	1 [3.1]	-	8 [66.7]$^{*,g)}$	10 [100.0]$^{*,g)}$
EhaD	27 [24.3]	14 [28.0]	11 [21.2]	2 [22.2]	12 [48.0]$^{*,h)}$	-	1 [3.1]	1 [33.3]	9 [75.0]$^{*,h)}$	4 [40.0]$^{*,h)}$
EhaG	51 [45.9]	16 [32.0]	32 [61.5]$^{*,b)}$	3 [33.3]	9 [36.0]	25 [86.2]*	7 [21.9]	-	-	10 [100.0]*
EhaJ	4 [3.6]	2 [4.0]	2 [3.8]	-	-	-	4 [12.5]	-	-	-
EspP	7 [6.3]	1 [2.0]	6 [11.5]	-	-	2 [6.9]	-	1 [33.3]	-	4 [40.0]$^{*,i)}$
IcsA	-	-	-	-	-	-	-	-	-	-
PicU	11 [9.0]$^{a)}$	5 [10.0]	5 [7.7]$^{a)}$	1 [11.1]	-	2 [3.4]$^{a)}$	8 [25.0]$^{*,j)}$	-	1 [8.3]	-
Sat	10 [9.0]	6 [12.0]	2 [3.8]	2 [22.2]	1 [4.0]	-	3 [9.4]	-	6 [50.0]*	-
SepA	23 [20.7]	8 [16.0]	10 [19.2]	5 [55.6]*	1 [4.0]	1 [3.4]	21 [65.6]*	-	-	-
TibA	-	-	-	-	-	-	-	-	-	-
UpaB	42 [37.8]	18 [36.0]	18 [34.6]	6 [66.7]	2 [8.0]	5 [17.2]	29 [90.6]$^{*,k)}$	-	1 [8.3]	5 [50.0]$^{*,l)}$
UpaC	60 [54.1]	28 [56.0]	24 [46.2]	8 [88.9]$^{*,c)}$	14 [56.0]	17 [58.6]	19 [59.4]	2 [66.7]	8 [66.7]	-
UpaG	4 [3.6]	2 [4.0]	1 [1.9]	1 [11.1]	-	-	2 [6.3]	-	2 [16.7]	-
UpaH	6 [5.4]	1 [2.0]	-	5 [55.6]*	-	-	6 [18.8]	-	-	-
UpaI	31 [27.9]	13 [26.0]	11 [21.2]	7 [77.8]*	-	-	29 [90.6]*	-	2 [16.7]	-
UpaJ	10 [9.0]	2 [4.0]	6 [11.5]	2 [22.2]	-	-	8 [25.0]	2 [66.7]	-	-

Two-tailed Fisher's exact test in a contingency table with hit number in comparison to non-hit numbers of each pathotype/phlylogroup strain count (fields without a hit were not included in the statistical analysis). Significant difference from the other groups (or otherwise indicated below)

* = P < 0.05.

$^{a)}$Because statistical tests don't work with negative numbers, only presence/abscence binary numbers were used for Ag43 and PicU for the two-tailed Fisher's exact test; also to make the percentages comparable to the other ATs (Ag43 total: 54, non-pathogenic: 23, IPEC: 23, ExPEC: 8, B1: 14, B2: 18, D: 7; PicU total: 10, IPEC: 4, B1: 1)

$^{b)}$Significantly associated with IPECs relative to commensal strains.

$^{c)}$Significantly associated with ExPECs relative to IPECs.

$^{d)}$Significantly associated with phylogroup C-I relative to phylogroups A and B2.

$^{e)}$Significantly associated with phylogroup E relative to phylogroups A and B1.

$^{f)}$Significantly associated with phylogroups A and B1 relative to phylogroups B2 and D.

$^{g)}$Significantly associated with phylogroups D and E relative to phylogroup B2.

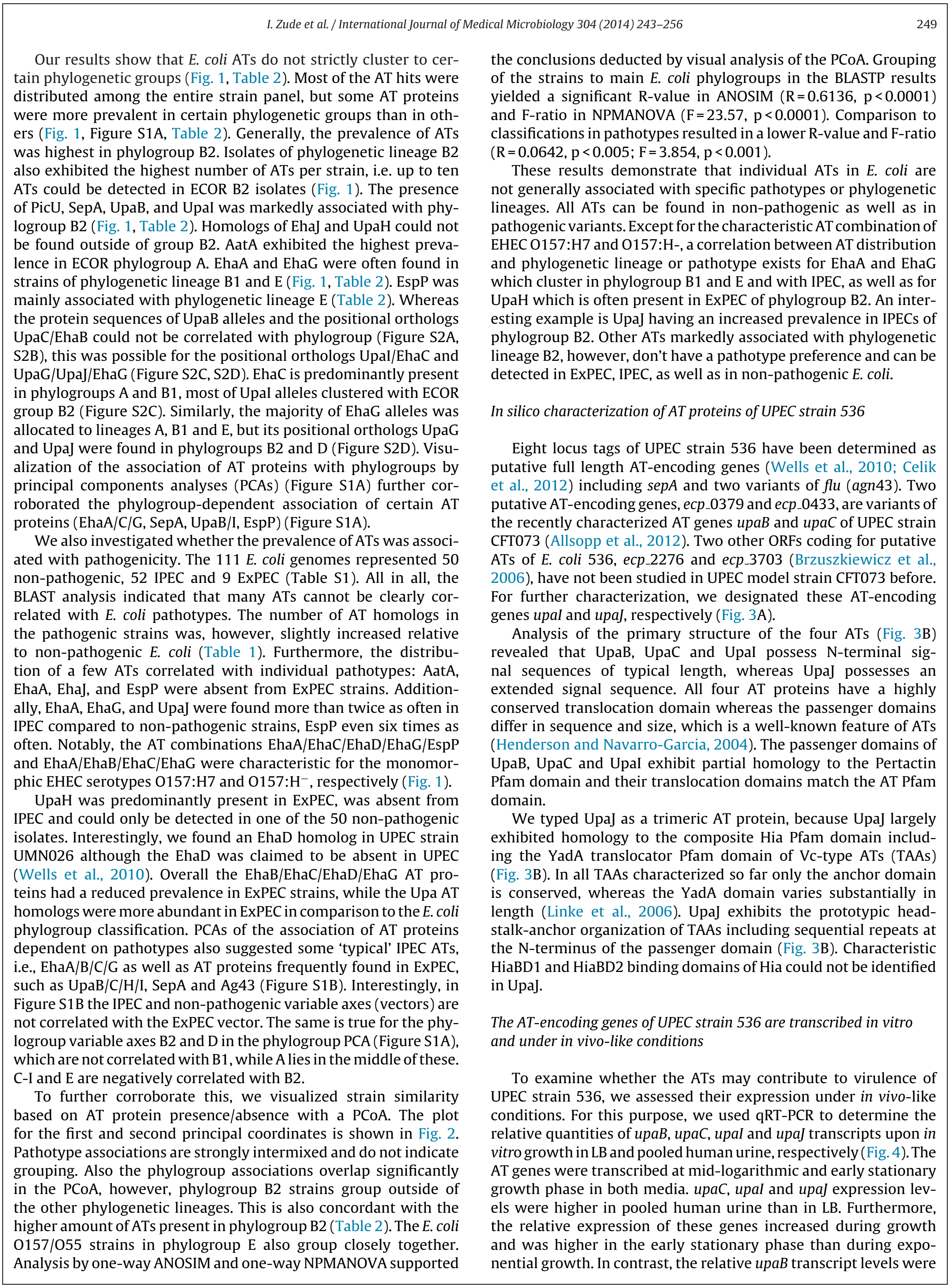
$^{h)}$Significantly associated with phylogroups A, D, and E relative to phylogroup B2.

$^{i)}$Significantly associated with phylogroup E relative to phylogroup B1.

$^{j)}$Significantly associated with phylogroup B2 relative to phylogroup B1.

$^{k)}$Significantly associated with phylogroup B2 relative to phylogroups A, B1, D, and E.

$^{l)}$Significantly associated with phylogroup E relative to phylogroup A.



Our results show that *E. coli* ATs do not strictly cluster to certain phylogenetic groups (Fig. 1, Table 2). Most of the AT hits were distributed among the entire strain panel, but some AT proteins were more prevalent in certain phylogenetic groups than in others (Fig. 1, Figure S1A, Table 2). Generally, the prevalence of ATs was highest in phylogroup B2. Isolates of phylogenetic lineage B2 also exhibited the highest number of ATs per strain, i.e. up to ten ATs could be detected in ECOR B2 isolates (Fig. 1). The presence of PicU, SepA, UpaB, and UpaI was markedly associated with phylogroup B2 (Fig. 1, Table 2). Homologs of EhaJ and UpaH could not be found outside of group B2. AatA exhibited the highest prevalence in ECOR phylogroup A. EhaA and EhaG were often found in strains of phylogenetic lineage B1 and E (Fig. 1, Table 2). EspP was mainly associated with phylogenetic lineage E (Table 2). Whereas the protein sequences of UpaB alleles and the positional orthologs UpaC/EhaB could not be correlated with phylogroup (Figure S2A, S2B), this was possible for the positional orthologs UpaI/EhaC and UpaG/UpaJ/EhaG (Figure S2C, S2D). EhaC is predominantly present in phylogroups A and B1, most of UpaI alleles clustered with ECOR group B2 (Figure S2C). Similarly, the majority of EhaG alleles was allocated to lineages A, B1 and E, but its positional orthologs UpaG and UpaJ were found in phylogroups B2 and D (Figure S2D). Visualization of the association of AT proteins with phylogroups by principal components analyses (PCAs) (Figure S1A) further corroborated the phylogroup-dependent association of certain AT proteins (EhaA/C/G, SepA, UpaB/I, EspP) (Figure S1A).

We also investigated whether the prevalence of ATs was associated with pathogenicity. The 111 *E. coli* genomes represented 50 non-pathogenic, 52 IPEC and 9 ExPEC (Table S1). All in all, the BLAST analysis indicated that many ATs cannot be clearly correlated with *E. coli* pathotypes. The number of AT homologs in the pathogenic strains was, however, slightly increased relative to non-pathogenic *E. coli* (Table 1). Furthermore, the distribution of a few ATs correlated with individual pathotypes: AatA, EhaA, EhaJ, and EspP were absent from ExPEC strains. Additionally, EhaA, EhaG, and UpaJ were found more than twice as often in IPEC compared to non-pathogenic strains, EspP even six times as often. Notably, the AT combinations EhaA/EhaC/EhaD/EhaG/EspP and EhaA/EhaB/EhaC/EhaG were characteristic for the monomorphic EHEC serotypes O157:H7 and O157:H^{-}, respectively (Fig. 1).

UpaH was predominantly present in ExPEC, was absent from IPEC and could only be detected in one of the 50 non-pathogenic isolates. Interestingly, we found an EhaD homolog in UPEC strain UMN026 although the EhaD was claimed to be absent in UPEC (Wells et al., 2010). Overall the EhaB/EhaC/EhaD/EhaG AT proteins had a reduced prevalence in ExPEC strains, while the Upa AT homologs were more abundant in ExPEC in comparison to the *E. coli* phylogroup classification. PCAs of the association of AT proteins dependent on pathotypes also suggested some 'typical' IPEC ATs, i.e., EhaA/B/C/G as well as AT proteins frequently found in ExPEC, such as UpaB/C/H/I, SepA and Ag43 (Figure S1B). Interestingly, in Figure S1B the IPEC and non-pathogenic variable axes (vectors) are not correlated with the ExPEC vector. The same is true for the phylogroup variable axes B2 and D in the phylogroup PCA (Figure S1A), which are not correlated with B1, while A lies in the middle of these. C-I and E are negatively correlated with B2.

To further corroborate this, we visualized strain similarity based on AT protein presence/absence with a PCoA. The plot for the first and second principal coordinates is shown in Fig. 2. Pathotype associations are strongly intermixed and do not indicate grouping. Also the phylogroup associations overlap significantly in the PCoA, however, phylogroup B2 strains group outside of the other phylogenetic lineages. This is also concordant with the higher amount of ATs present in phylogroup B2 (Table 2). The *E. coli* O157/O55 strains in phylogroup E also group closely together. Analysis by one-way ANOSIM and one-way NPMANOVA supported the conclusions deducted by visual analysis of the PCoA. Grouping of the strains to main *E. coli* phylogroups in the BLASTP results yielded a significant R-value in ANOSIM ($R = 0.6136$, $p < 0.0001$) and F-ratio in NPMANOVA ($F = 23.57$, $p < 0.0001$). Comparison to classifications in pathotypes resulted in a lower R-value and F-ratio ($R = 0.0642$, $p < 0.005$; $F = 3.854$, $p < 0.001$).

These results demonstrate that individual ATs in *E. coli* are not generally associated with specific pathotypes or phylogenetic lineages. All ATs can be found in non-pathogenic as well as in pathogenic variants. Except for the characteristic AT combination of EHEC O157:H7 and O157:H-, a correlation between AT distribution and phylogenetic lineage or pathotype exists for EhaA and EhaG which cluster in phylogroup B1 and E and with IPEC, as well as for UpaH which is often present in ExPEC of phylogroup B2. An interesting example is UpaJ having an increased prevalence in IPECs of phylogroup B2. Other ATs markedly associated with phylogenetic lineage B2, however, don't have a pathotype preference and can be detected in ExPEC, IPEC, as well as in non-pathogenic *E. coli*.

In silico characterization of AT proteins of UPEC strain 536

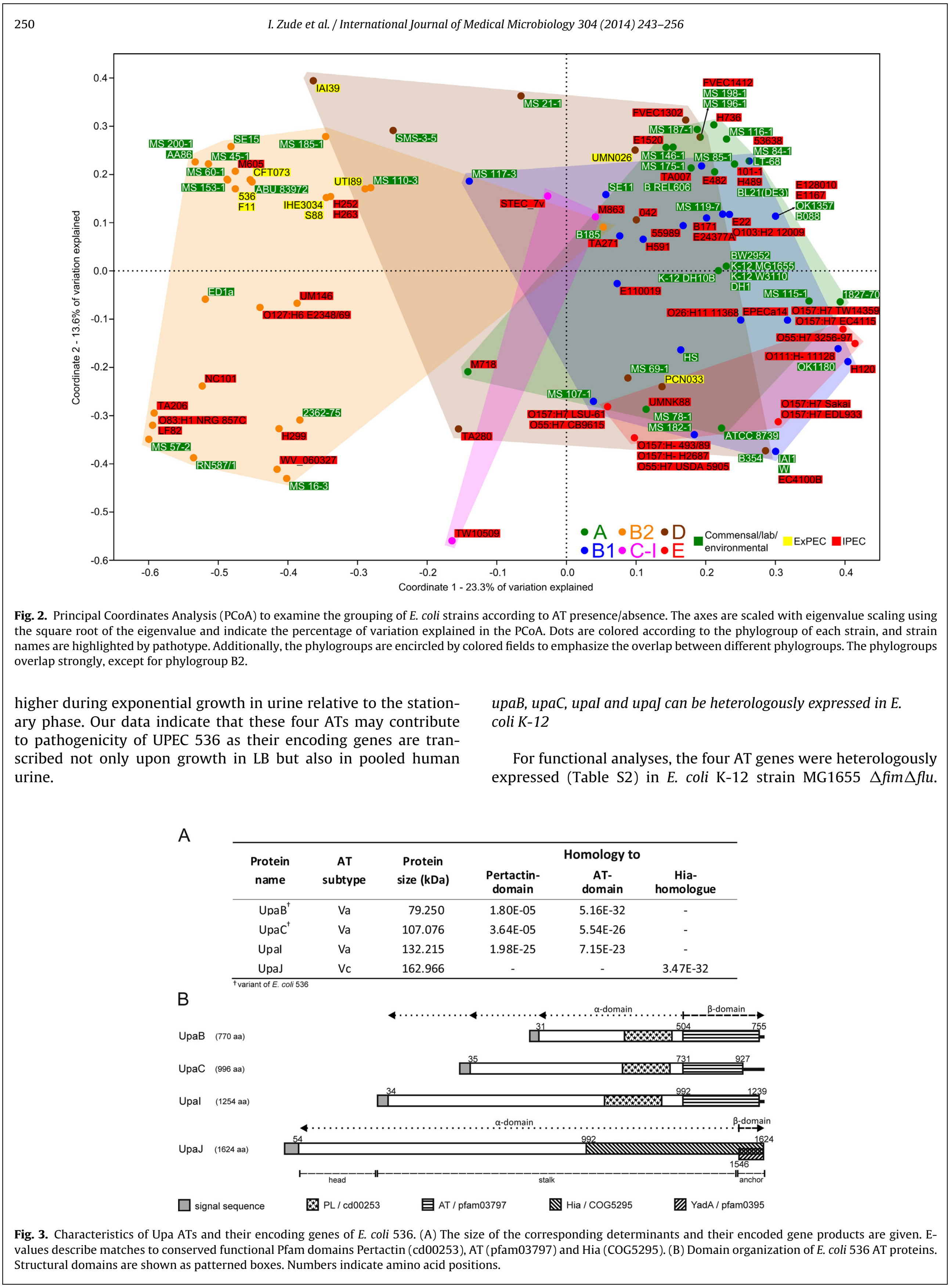
Eight locus tags of UPEC strain 536 have been determined as putative full length AT-encoding genes (Wells et al., 2010; Celik et al., 2012) including *sepA* and two variants of *flu* (*agn*43). Two putative AT-encoding genes, *ecp*_0379 and *ecp*_0433, are variants of the recently characterized AT genes *upaB* and *upaC* of UPEC strain CFT073 (Allsopp et al., 2012). Two other ORFs coding for putative ATs of *E. coli* 536, *ecp*_2276 and *ecp*_3703 (Brzuszkiewicz et al., 2006), have not been studied in UPEC model strain CFT073 before. For further characterization, we designated these AT-encoding genes *upaI* and *upaJ*, respectively (Fig. 3A).

Analysis of the primary structure of the four ATs (Fig. 3B) revealed that UpaB, UpaC and UpaI possess N-terminal signal sequences of typical length, whereas UpaJ possesses an extended signal sequence. All four AT proteins have a highly conserved translocation domain whereas the passenger domains differ in sequence and size, which is a well-known feature of ATs (Henderson and Navarro-Garcia, 2004). The passenger domains of UpaB, UpaC and UpaI exhibit partial homology to the Pertactin Pfam domain and their translocation domains match the AT Pfam domain.

We typed UpaJ as a trimeric AT protein, because UpaJ largely exhibited homology to the composite Hia Pfam domain including the YadA translocator Pfam domain of Vc-type ATs (TAAs) (Fig. 3B). In all TAAs characterized so far only the anchor domain is conserved, whereas the YadA domain varies substantially in length (Linke et al., 2006). UpaJ exhibits the prototypic head-stalk-anchor organization of TAAs including sequential repeats at the N-terminus of the passenger domain (Fig. 3B). Characteristic HiaBD1 and HiaBD2 binding domains of Hia could not be identified in UpaJ.

The AT-encoding genes of UPEC strain 536 are transcribed in vitro and under in vivo-like conditions

To examine whether the ATs may contribute to virulence of UPEC strain 536, we assessed their expression under *in vivo*-like conditions. For this purpose, we used qRT-PCR to determine the relative quantities of *upaB*, *upaC*, *upaI* and *upaJ* transcripts upon *in vitro* growth in LB and pooled human urine, respectively (Fig. 4). The AT genes were transcribed at mid-logarithmic and early stationary growth phase in both media. *upaC*, *upaI* and *upaJ* expression levels were higher in pooled human urine than in LB. Furthermore, the relative expression of these genes increased during growth and was higher in the early stationary phase than during exponential growth. In contrast, the relative *upaB* transcript levels were



Coordinate 2 - 13.6% of variation explained
Coordinate 1 - 23.3% of variation explained
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
IAI39
MS 21-1
FVEC1412
MS 198-1
MS 196-1
FVEC1302
H736
MS 187-1
MS 116-1
E1520
53638
MS 84-1
UMN026
MS 146-1
MS 85-1
LT-68
MS 175-1
101-1
TA007
E482
H489
SE11
B REL606
BL21(DE3)
E128010
E1167
MS 119-7
OK1357
B088
M863
042
E22
B171
O103:H2 12009
B185
55989
E24377A
TA271
H591
BW2952
K-12 MG1655
K-12 W3110
K-12 DH10B
DH1
E110019
MS 115-1
1827-70
O26:H11 11368
EPECa14
O157:H7 TW14359
O157:H7 EC4115
O55:H7 3256-97
O111:H- 11128
OK1180
HS
H120
MS 69-1
PCN033
MS 107-1
UMNK88
O157:H7 Sakai
O157:H7 EDL933
O157:H7 LSU-61
MS 78-1
O55:H7 CB9615
MS 182-1
ATCC 8739
O157:H- 493/89
O157:H- H2687
B354
IAI1
O55:H7 USDA 5905
W
EC4100B
MS 117-3
STEC_7v
SMS-3-5
MS 200-1
SE15
MS 185-1
AA86
MS 45-1
M605
MS 60-1
CFT073
UTI89
MS 110-3
MS 153-1
ABU 83972
536
F11
IHE3034
S88
H252
H263
ED1a
UM146
O127:H6 E2348/69
M718
NC101
TA206
O83:H1 NRG 857C
LF82
2362-75
H299
TA280
MS 57-2
RN587/1
WV_060327
MS 16-3
TW10509
● A ● B2 ● D
● B1 ● C-I ● E
🟩 Commensal/lab/ environmental
🟨 ExPEC
🟥 IPEC

Fig. 2. Principal Coordinates Analysis (PCoA) to examine the grouping of *E. coli* strains according to AT presence/absence. The axes are scaled with eigenvalue scaling using the square root of the eigenvalue and indicate the percentage of variation explained in the PCoA. Dots are colored according to the phylogroup of each strain, and strain names are highlighted by pathotype. Additionally, the phylogroups are encircled by colored fields to emphasize the overlap between different phylogroups. The phylogroups overlap strongly, except for phylogroup B2.

higher during exponential growth in urine relative to the stationary phase. Our data indicate that these four ATs may contribute to pathogenicity of UPEC 536 as their encoding genes are transcribed not only upon growth in LB but also in pooled human urine.

upaB, upaC, upaI and upaJ can be heterologously expressed in E. coli K-12

For functional analyses, the four AT genes were heterologously expressed (Table S2) in *E. coli* K-12 strain MG1655 Δ*fim*Δ*flu*.

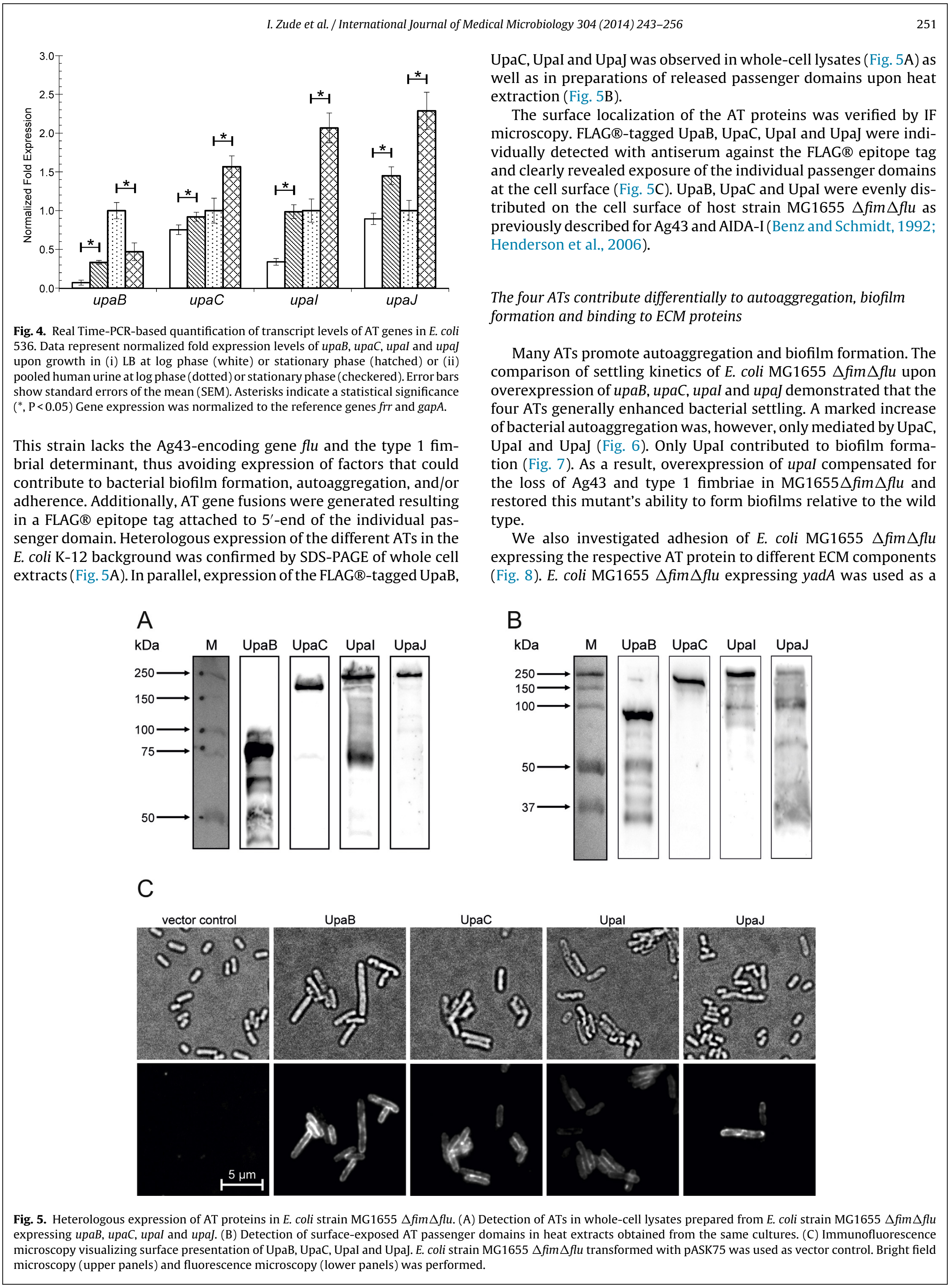
A

Protein name	AT subtype	Protein size (kDa)	Homology to		
			Pertactin-domain	AT-domain	Hia-homologue
UpaB†	Va	79.250	1.80E-05	5.16E-32	-
UpaC†	Va	107.076	3.64E-05	5.54E-26	-
UpaI	Va	132.215	1.98E-25	7.15E-23	-
UpaJ	Vc	162.966	-	-	3.47E-32

† variant of *E. coli* 536

B

Fig. 3. Characteristics of Upa ATs and their encoding genes of *E. coli* 536. (A) The size of the corresponding determinants and their encoded gene products are given. E-values describe matches to conserved functional Pfam domains Pertactin (cd00253), AT (pfam03797) and Hia (COG5295). (B) Domain organization of *E. coli* 536 AT proteins. Structural domains are shown as patterned boxes. Numbers indicate amino acid positions.


Normalized Fold Expression
0.0
0.5
1.0
1.5
2.0
2.5
3.0
*
upaB
upaC
upaI
upaJ
A
B
kDa
M
UpaB
UpaC
UpaI
UpaJ
250
150
100
75
50
37
C
vector control
5 µm

I. Zude et al. / International Journal of Medical Microbiology 304 (2014) 243–256 251

Fig. 4. Real Time-PCR-based quantification of transcript levels of AT genes in *E. coli* 536. Data represent normalized fold expression levels of *upaB*, *upaC*, *upaI* and *upaJ* upon growth in (i) LB at log phase (white) or stationary phase (hatched) or (ii) pooled human urine at log phase (dotted) or stationary phase (checkered). Error bars show standard errors of the mean (SEM). Asterisks indicate a statistical significance (*, P < 0.05) Gene expression was normalized to the reference genes *frr* and *gapA*.

This strain lacks the Ag43-encoding gene *flu* and the type 1 fimbrial determinant, thus avoiding expression of factors that could contribute to bacterial biofilm formation, autoaggregation, and/or adherence. Additionally, AT gene fusions were generated resulting in a FLAG® epitope tag attached to 5′-end of the individual passenger domain. Heterologous expression of the different ATs in the *E. coli* K-12 background was confirmed by SDS-PAGE of whole cell extracts (Fig. 5A). In parallel, expression of the FLAG®-tagged UpaB, UpaC, UpaI and UpaJ was observed in whole-cell lysates (Fig. 5A) as well as in preparations of released passenger domains upon heat extraction (Fig. 5B).

The surface localization of the AT proteins was verified by IF microscopy. FLAG®-tagged UpaB, UpaC, UpaI and UpaJ were individually detected with antiserum against the FLAG® epitope tag and clearly revealed exposure of the individual passenger domains at the cell surface (Fig. 5C). UpaB, UpaC and UpaI were evenly distributed on the cell surface of host strain MG1655 Δ*fim*Δ*flu* as previously described for Ag43 and AIDA-I (Benz and Schmidt, 1992; Henderson et al., 2006).

The four ATs contribute differentially to autoaggregation, biofilm formation and binding to ECM proteins

Many ATs promote autoaggregation and biofilm formation. The comparison of settling kinetics of *E. coli* MG1655 Δ*fim*Δ*flu* upon overexpression of *upaB*, *upaC*, *upaI* and *upaJ* demonstrated that the four ATs generally enhanced bacterial settling. A marked increase of bacterial autoaggregation was, however, only mediated by UpaC, UpaI and UpaJ (Fig. 6). Only UpaI contributed to biofilm formation (Fig. 7). As a result, overexpression of *upaI* compensated for the loss of Ag43 and type 1 fimbriae in MG1655Δ*fim*Δ*flu* and restored this mutant's ability to form biofilms relative to the wild type.

We also investigated adhesion of *E. coli* MG1655 Δ*fim*Δ*flu* expressing the respective AT protein to different ECM components (Fig. 8). *E. coli* MG1655 Δ*fim*Δ*flu* expressing *yadA* was used as a

Fig. 5. Heterologous expression of AT proteins in *E. coli* strain MG1655 Δ*fim*Δ*flu*. (A) Detection of ATs in whole-cell lysates prepared from *E. coli* strain MG1655 Δ*fim*Δ*flu* expressing *upaB*, *upaC*, *upaI* and *upaJ*. (B) Detection of surface-exposed AT passenger domains in heat extracts obtained from the same cultures. (C) Immunofluorescence microscopy visualizing surface presentation of UpaB, UpaC, UpaI and UpaJ. *E. coli* strain MG1655 Δ*fim*Δ*flu* transformed with pASK75 was used as vector control. Bright field microscopy (upper panels) and fluorescence microscopy (lower panels) was performed.

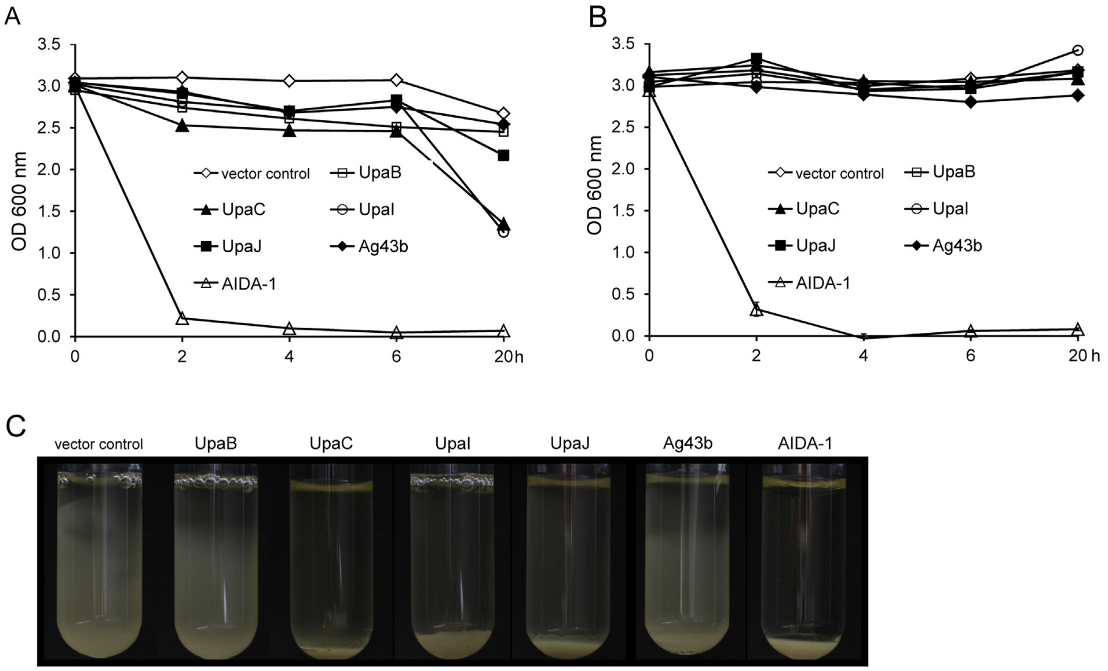


Fig. 6. Autoaggregation mediated by autotransporters of *E. coli* strain 536. (A) Induced expression of *upaC* and *upaI* led to significant bacterial settling ($P < 0.005$) compared to non-induced cultures (B). AIDA-1 was used as a positive control. The data represent the average of three independent experiments using an *E. coli* strain MG1655 Δ*fim*Δ*flu* background. (C) Expression of AT proteins promotes autoaggregation, visible as bacterial accumulation at the bottom of the test tubes after 20 h.

positive control, because this TAA is known to mediate binding to collagens and laminin and to a lesser extent to vitronectin (Ackermann et al., 2008). Expression of *upaB* resulted in a modest but significant binding to actin and collagen IV and clearly mediated binding to vitronectin. *upaC* and *upaJ* expression, on the other hand, did not promote bacterial binding to any of the tested ECM proteins. Interestingly *upaI* expression significantly enhanced binding to actin, collagen type I and type IV, laminin and vitronectin. Binding of UpaB, UpaC, UpaI, and UpaJ to fetuin and fibronectin was not observed (data not shown). Therefore, UpaB and UpaI promote specific binding to some ECM proteins. The overexpression of UpaB, UpaC, UpaI and UpaJ in *E. coli* MG1655Δ*fim*Δ*flu* did not result in significant adherence to T24 human bladder epithelial or T84 or human intestinal epithelial cells (data not shown). Thus, UpaB, UpaC, UpaI and UpaJ of UPEC 536 possess typical traits of ATs and can contribute to varying degrees of bacterial autoaggregation, biofilm formation and adhesion to

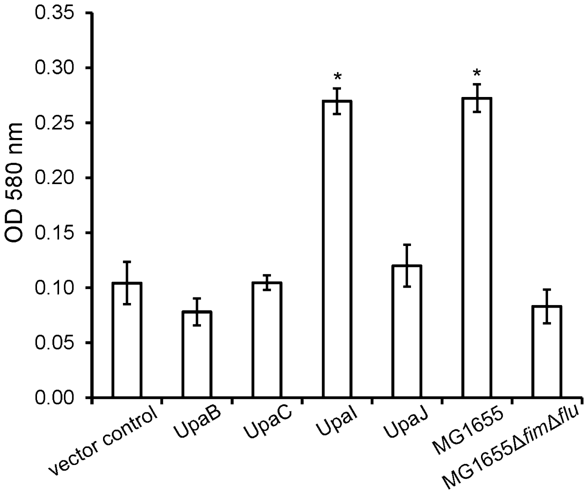


Fig. 7. Biofilm formation mediated by AT proteins of *E. coli* 536. The ATs UpaB, UpaC, UpaI, and UpaJ were overexpressed in *E. coli* strain MG1655 Δ*fim*Δ*flu*. Expression of *upaI* led to significantly increased biofilm formation ($P < 0.001$) relative to control strain MG1655 Δ*fim*Δ*flu*. The data represent averages from three independent experiments and SEM.

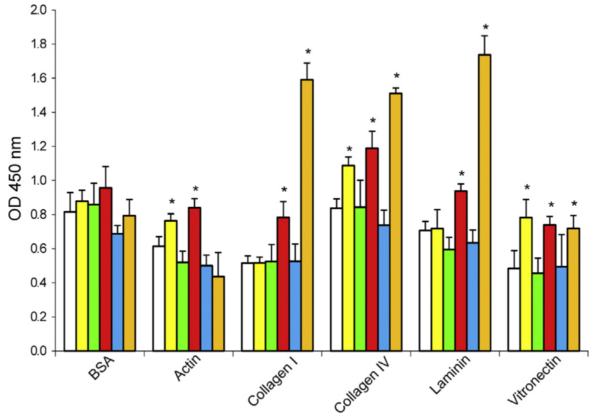
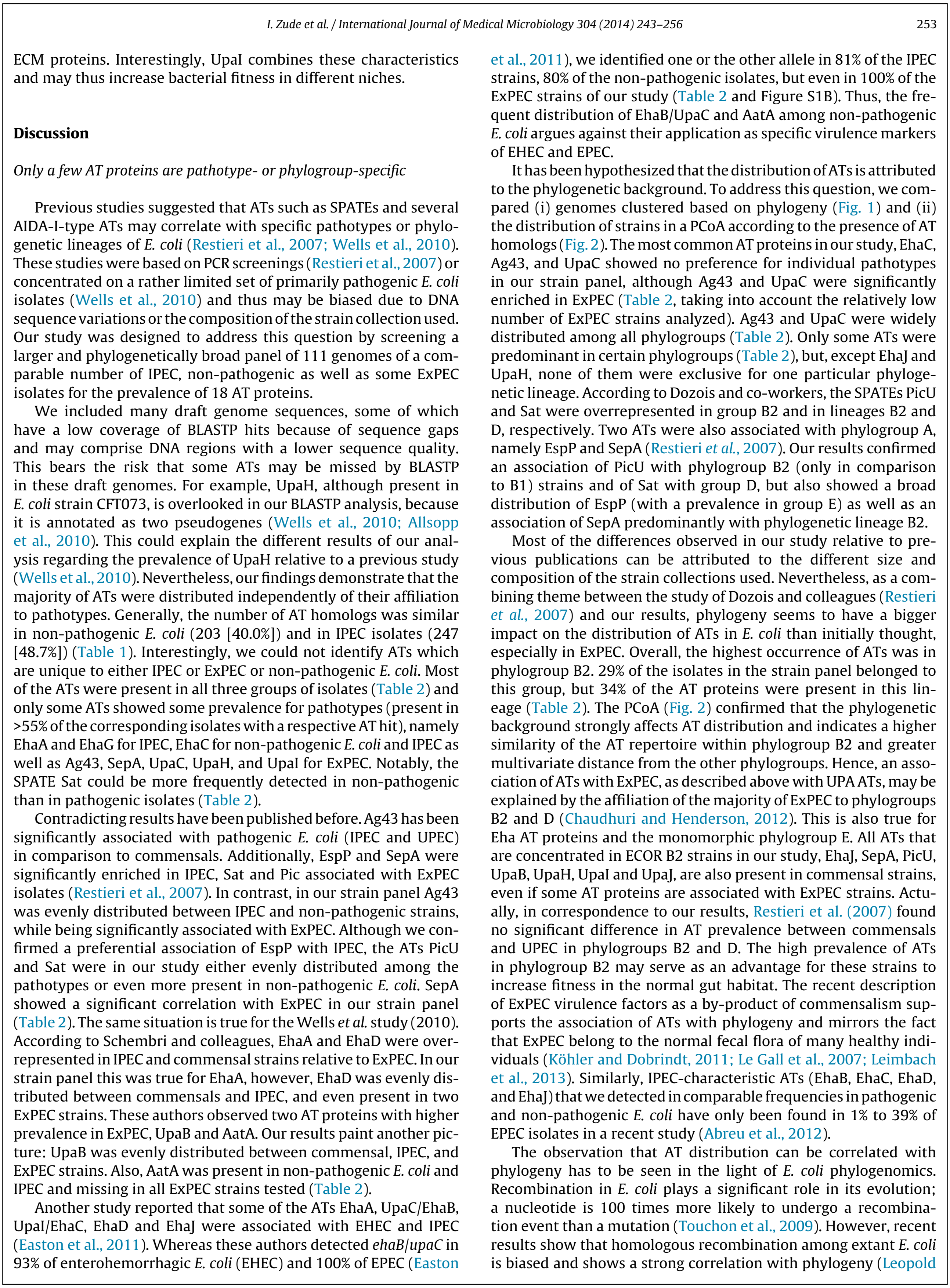


Fig. 8. Adhesion to ECM proteins mediated by AT proteins of *E. coli* 536. Adhesion was quantified in an ELISA-based binding assay upon expression of the UpaB, UpaC, UpaI, and UpaJ ATs in *E. coli* strain MG1655 Δ*fim*Δ*flu*. White bars, vector control (pASK75); yellow bars, UpaB; green bars, UpaC; red bars, UpaI; blue bars, UpaJ; orange bars, YadA. Data represent averages from three independent experiments and SEM. Expression of *upaB* led to significant adhesion to actin, collagen type IV and vitronectin ($P < 0.05$) and expression of *upaI* led to significant adhesion to actin, collagen type I and type IV, laminin and vitronectin ($P < 0.05$) compared to the vector control (pASK75).



ECM proteins. Interestingly, UpaI combines these characteristics and may thus increase bacterial fitness in different niches.

Discussion

Only a few AT proteins are pathotype- or phylogroup-specific

Previous studies suggested that ATs such as SPATEs and several AIDA-I-type ATs may correlate with specific pathotypes or phylogenetic lineages of *E. coli* (Restieri et al., 2007; Wells et al., 2010). These studies were based on PCR screenings (Restieri et al., 2007) or concentrated on a rather limited set of primarily pathogenic *E. coli* isolates (Wells et al., 2010) and thus may be biased due to DNA sequence variations or the composition of the strain collection used. Our study was designed to address this question by screening a larger and phylogenetically broad panel of 111 genomes of a comparable number of IPEC, non-pathogenic as well as some ExPEC isolates for the prevalence of 18 AT proteins.

We included many draft genome sequences, some of which have a low coverage of BLASTP hits because of sequence gaps and may comprise DNA regions with a lower sequence quality. This bears the risk that some ATs may be missed by BLASTP in these draft genomes. For example, UpaH, although present in *E. coli* strain CFT073, is overlooked in our BLASTP analysis, because it is annotated as two pseudogenes (Wells et al., 2010; Allsopp et al., 2010). This could explain the different results of our analysis regarding the prevalence of UpaH relative to a previous study (Wells et al., 2010). Nevertheless, our findings demonstrate that the majority of ATs were distributed independently of their affiliation to pathotypes. Generally, the number of AT homologs was similar in non-pathogenic *E. coli* (203 [40.0%]) and in IPEC isolates (247 [48.7%]) (Table 1). Interestingly, we could not identify ATs which are unique to either IPEC or ExPEC or non-pathogenic *E. coli*. Most of the ATs were present in all three groups of isolates (Table 2) and only some ATs showed some prevalence for pathotypes (present in >55% of the corresponding isolates with a respective AT hit), namely EhaA and EhaG for IPEC, EhaC for non-pathogenic *E. coli* and IPEC as well as Ag43, SepA, UpaC, UpaH, and UpaI for ExPEC. Notably, the SPATE Sat could be more frequently detected in non-pathogenic than in pathogenic isolates (Table 2).

Contradicting results have been published before. Ag43 has been significantly associated with pathogenic *E. coli* (IPEC and UPEC) in comparison to commensals. Additionally, EspP and SepA were significantly enriched in IPEC, Sat and Pic associated with ExPEC isolates (Restieri et al., 2007). In contrast, in our strain panel Ag43 was evenly distributed between IPEC and non-pathogenic strains, while being significantly associated with ExPEC. Although we confirmed a preferential association of EspP with IPEC, the ATs PicU and Sat were in our study either evenly distributed among the pathotypes or even more present in non-pathogenic *E. coli*. SepA showed a significant correlation with ExPEC in our strain panel (Table 2). The same situation is true for the Wells *et al.* study (2010). According to Schembri and colleagues, EhaA and EhaD were overrepresented in IPEC and commensal strains relative to ExPEC. In our strain panel this was true for EhaA, however, EhaD was evenly distributed between commensals and IPEC, and even present in two ExPEC strains. These authors observed two AT proteins with higher prevalence in ExPEC, UpaB and AatA. Our results paint another picture: UpaB was evenly distributed between commensal, IPEC, and ExPEC strains. Also, AatA was present in non-pathogenic *E. coli* and IPEC and missing in all ExPEC strains tested (Table 2).

Another study reported that some of the ATs EhaA, UpaC/EhaB, UpaI/EhaC, EhaD and EhaJ were associated with EHEC and IPEC (Easton et al., 2011). Whereas these authors detected *ehaB*/*upaC* in 93% of enterohemorrhagic *E. coli* (EHEC) and 100% of EPEC (Easton et al., 2011), we identified one or the other allele in 81% of the IPEC strains, 80% of the non-pathogenic isolates, but even in 100% of the ExPEC strains of our study (Table 2 and Figure S1B). Thus, the frequent distribution of EhaB/UpaC and AatA among non-pathogenic *E. coli* argues against their application as specific virulence markers of EHEC and EPEC.

It has been hypothesized that the distribution of ATs is attributed to the phylogenetic background. To address this question, we compared (i) genomes clustered based on phylogeny (Fig. 1) and (ii) the distribution of strains in a PCoA according to the presence of AT homologs (Fig. 2). The most common AT proteins in our study, EhaC, Ag43, and UpaC showed no preference for individual pathotypes in our strain panel, although Ag43 and UpaC were significantly enriched in ExPEC (Table 2, taking into account the relatively low number of ExPEC strains analyzed). Ag43 and UpaC were widely distributed among all phylogroups (Table 2). Only some ATs were predominant in certain phylogroups (Table 2), but, except EhaJ and UpaH, none of them were exclusive for one particular phylogenetic lineage. According to Dozois and co-workers, the SPATEs PicU and Sat were overrepresented in group B2 and in lineages B2 and D, respectively. Two ATs were also associated with phylogroup A, namely EspP and SepA (Restieri *et al.*, 2007). Our results confirmed an association of PicU with phylogroup B2 (only in comparison to B1) strains and of Sat with group D, but also showed a broad distribution of EspP (with a prevalence in group E) as well as an association of SepA predominantly with phylogenetic lineage B2.

Most of the differences observed in our study relative to previous publications can be attributed to the different size and composition of the strain collections used. Nevertheless, as a combining theme between the study of Dozois and colleagues (Restieri *et al.*, 2007) and our results, phylogeny seems to have a bigger impact on the distribution of ATs in *E. coli* than initially thought, especially in ExPEC. Overall, the highest occurrence of ATs was in phylogroup B2. 29% of the isolates in the strain panel belonged to this group, but 34% of the AT proteins were present in this lineage (Table 2). The PCoA (Fig. 2) confirmed that the phylogenetic background strongly affects AT distribution and indicates a higher similarity of the AT repertoire within phylogroup B2 and greater multivariate distance from the other phylogroups. Hence, an association of ATs with ExPEC, as described above with UPA ATs, may be explained by the affiliation of the majority of ExPEC to phylogroups B2 and D (Chaudhuri and Henderson, 2012). This is also true for Eha AT proteins and the monomorphic phylogroup E. All ATs that are concentrated in ECOR B2 strains in our study, EhaJ, SepA, PicU, UpaB, UpaH, UpaI and UpaJ, are also present in commensal strains, even if some AT proteins are associated with ExPEC strains. Actually, in correspondence to our results, Restieri et al. (2007) found no significant difference in AT prevalence between commensals and UPEC in phylogroups B2 and D. The high prevalence of ATs in phylogroup B2 may serve as an advantage for these strains to increase fitness in the normal gut habitat. The recent description of ExPEC virulence factors as a by-product of commensalism supports the association of ATs with phylogeny and mirrors the fact that ExPEC belong to the normal fecal flora of many healthy individuals (Köhler and Dobrindt, 2011; Le Gall et al., 2007; Leimbach et al., 2013). Similarly, IPEC-characteristic ATs (EhaB, EhaC, EhaD, and EhaJ) that we detected in comparable frequencies in pathogenic and non-pathogenic *E. coli* have only been found in 1% to 39% of EPEC isolates in a recent study (Abreu et al., 2012).

The observation that AT distribution can be correlated with phylogeny has to be seen in the light of *E. coli* phylogenomics. Recombination in *E. coli* plays a significant role in its evolution; a nucleotide is 100 times more likely to undergo a recombination event than a mutation (Touchon et al., 2009). However, recent results show that homologous recombination among extant *E. coli* is biased and shows a strong correlation with phylogeny (Leopold et al., 2011; Didelot et al., 2012). Whole genome phylogeny of *E. coli* shows a split in two branches with ECOR group D being polyphyletic at the root (Touchon et al., 2009; Chaudhuri and Henderson, 2012). The phylogenetic history of *E. coli* impacts homologous recombination, in that strains of closely related phylogroups have a preference for recombination with and especially within each other. Recombination between phylogroups B2 and D, and between A and B1 is high. Group E is a special case separate from the others with the background of EHEC/EPEC specialization. Restrictions for recombination exist between phylogroups B2 and A/B1/E.

Interestingly, the PCoA reflects these barriers for homologous recombination (Fig. 2), with B2 strains grouping separate from the other phylogroups. Phylogroups A, B1, and E overlap in total. Phylogroup D lies between the B2 and A/B1/E clusters, reflecting its intermediate role in recombination. Phylogroup association of the strains in the BLASTP results indicated that the groups are separated but overlap in the ANOSIM and NPMANOVA analyses. In contrast, pathotype classification resulted in only low group significance in the ANOSIM and NPMANOVA tests. Additionally, the PCA based on phylogenetic classification showed variable axes that also reflect recombination (Figure S1A). The B2 vector is closest to D, while the B1 vector is not correlated to that of phylogroup B2. The phylogroup E vector indicates a negative correlation with B2.

These findings were corroborated by an analysis of the *E. coli* genetic population structure based on a large MLST database in comparison to whole-genome phylogeny. The rate of genotype (sequence type) and DNA admixture in the core genome of subpopulations of *E. coli* was determined. Phylogroup B2 of *E. coli* could be associated with a single population and this population had the lowest rate of admixture, which can be attributed to restricted homologous recombination with other phylogroups. This is also true for the *E. coli* O55 and O157 isolates of phylogroup E. On the other hand, the K-12 strains in phylogroup A show the highest rate of homologous core genome recombination (McNally et al., 2013). It was even hypothesized, that *E. coli* may be in the process of a speciation event, if the restrictions of recombination are further enforced between the two phylogenetic branches or populations of *E. coli* (Leopold et al., 2011; Didelot et al., 2012; McNally et al., 2013). An example of such enforcement is the recently emerged multidrug-resistant ExPEC sequence type 131 in phylogroup B2 with a strong reduction in detectable homologous core genome recombination with other lineages of the species and even other closely related B2 strains (McNally et al., 2013).

Inter- and especially intragroup recombination is important to keep the genetic information of closely related strains similar, a cohesive force to counteract divergence. As a consequence, these recombination restrictions may impact the distribution of ATs. Indeed, several studies have shown a link between extraintestinal virulence and phylogroup B2 (Picard et al., 1999; Johnson et al., 2001). Finally, horizontal gene transfer (HGT) can also contribute to the spread of AT genes, involving mobile genetic elements (MGEs), like plasmids, phages, and genomic islands. This leads to a model of sympatric speciation in that HGT acts as a founder effect and might lead to the proliferation of a successful clone, e.g. through genetic factors like novel restriction/modification systems, resistance determinants, or ecological fitness in a specific habitat. Restriction of recombination with closely related *E. coli* and high intragroup recombination results in a steady divergence of the clone, a higher niche adaptation and finally in a possible speciation (McNally et al., 2013; Corander et al., 2012).

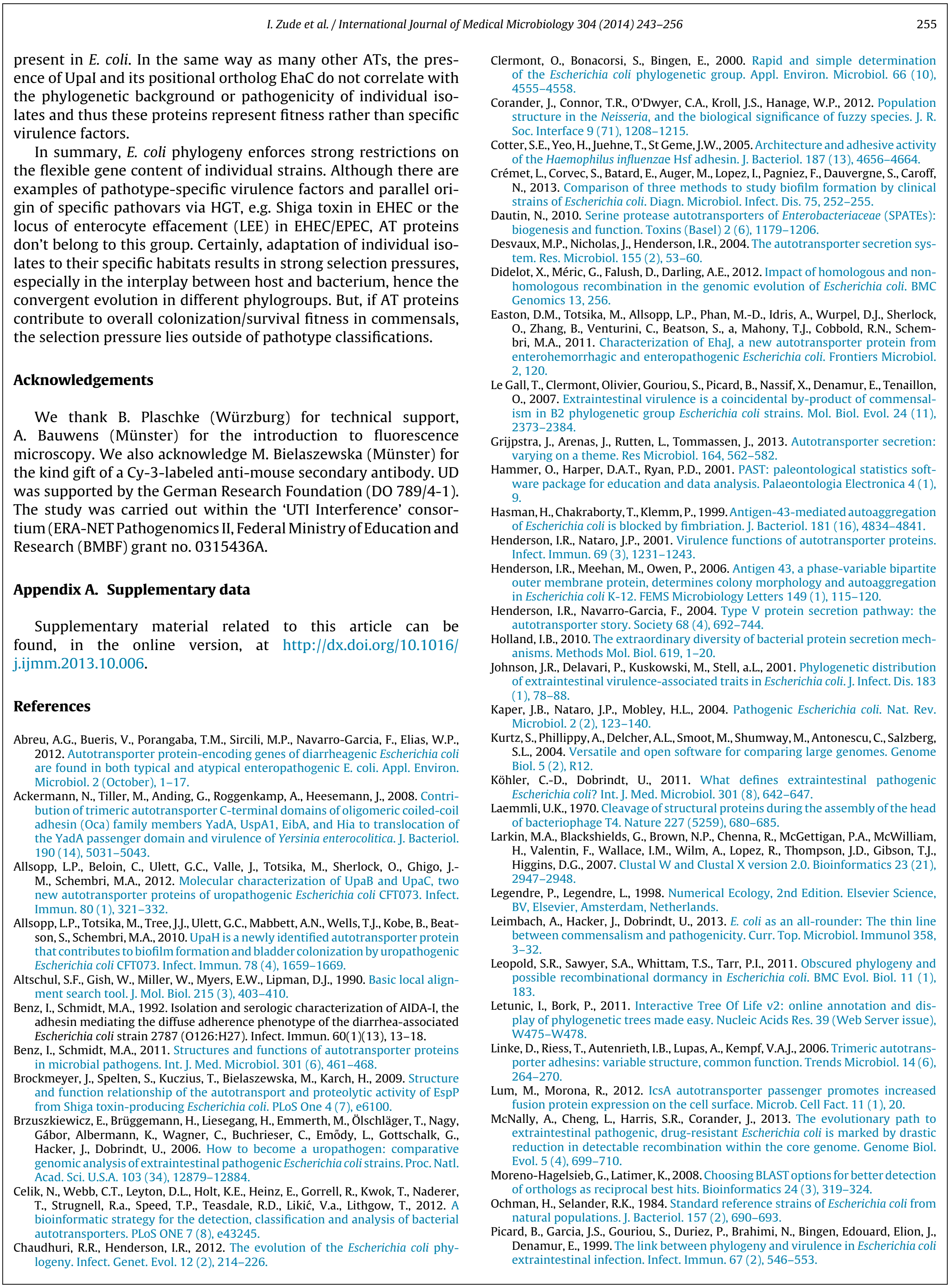
Characterization of AT proteins of UPEC 536

Extraintestinal virulence of *E. coli* is considered to be a by-product of commensalism in the phylogenetic group B2 (Le Gall et al., 2007) and several Upa ATs were more frequently present among ExPEC strains than among IPEC. We characterized four ATs of UPEC strain 536 with respect to their contribution to virulence- and fitness-associated traits. Two of them were variants of UpaB and UpaC of UPEC strain CFT073. The novel ATs, designated UpaI and UpaJ, show similarity to EhaC and UpaG, but because of different properties still represent different proteins (Table S4). Although EhaC (82% identity to UpaI) has been described in EHEC before, it has never been functionally characterized, probably because the correct N-terminal sequence was unknown and thus the protein could not be heterologously expressed (Wells et al., 2008). Interestingly, we identified UpaI together with its positional ortholog EhaC as the most common AT among all *E. coli* strains examined in our study (Table 1). We demonstrated that UpaI mediates cell aggregation, biofilm formation and binding to ECM proteins.

UpaB, UpaC, UpaI and UpaJ were expressed in UPEC strain 536 *in vitro* in LB as well as in pooled human urine mimicking *in vivo*-like conditions (Fig. 4). When heterologously expressed in *E. coli* MG1655 Δ*flu*Δ*fim*, all four ATs were displayed on the bacterial surface (Fig. 5C). We noticed that detection of cell surface-exposed UpaJ was lower relative to the other ATs tested, suggesting that an N-terminal portion of the protein including the FLAG® tag was cleaved, either during or after translocation to the cell surface. This is supported by breakdown products of UpaJ observed upon heat extraction (Fig. 5B). Similar observations were recently reported for the IcsA AT fusion protein expression on the cell surface (Lum and Morona, 2012). TAAs usually mediate adherence due to binding domains as described for Hia and Hsf (Yeo et al., 2004; Cotter et al., 2005). No such binding domains were found in UpaJ and probably therefore UpaJ exhibited no typical functional properties of TAAs except weak autoaggregation. Our sequence comparison indicated that UpaJ is a shorter variant of UpaG, but lacks HiaBD1 and HiaBD2 binding domains of TAAs. In contrast, UpaG mediates adhesion to human bladder cells (Valle et al., 2008), and has been shown, together with its positional ortholog EhaG, to promote autoaggregation, biofilm formation and adherence to various ECM proteins (Totsika et al., 2012).

Previously, UpaB and UpaC variants of UPEC CFT073 have been characterized (Allsopp et al., 2012). UpaB conferred binding to fibronectin, fibrinogen and laminin and contributed to uropathogenesis, whereas UpaC mediated increased biofilm formation (Allsopp et al., 2012). Our results generally confirmed the functional features of UpaB and UpaC, but also indicated discrepancies. The UpaB variant of UPEC 536 mediated adhesion to actin, collagen IV and vitronectin, but not to laminin. While having a high overall identity (93%), most of the amino acid sequence differences of UpaB in UPEC 536 and CFT073 are in the passenger domain (Table S4). These mutations may explain different functional properties. In contrast to UpaC of strain CFT073, UpaC of *E. coli* 536 did not promote biofilm formation, but contributed to autoaggregation, although both proteins/alleles are almost identical (99.8% identity). These discrepancies obviously result from different experimental conditions used. It has been shown that biofilm formation varies considerably according to the environmental conditions as well as the methods used (Crémet et al., 2013). Autoaggregation mediated by UpaC of UPEC 536 may be explained by the lack of the *fim* gene cluster in the K-12 host strain used for the autoaggregation assay. Several studies demonstrated that autoaggregation is blocked by fimbriae expression (Hasman et al., 1999; Schembri et al., 2004; Ulett et al., 2006).

For the first time, we functionally characterized UpaI. As a typical member of subtype Va ATs, UpaI promotes autoaggregation, biofilm formation as well as binding to different ECM proteins. Like UpaB, UpaC and UpaJ, also UpaI is expressed in pooled human urine and may thus contribute to uropathogenesis, although its intrinsic function may be during intestinal colonization. Taken together, UpaI further extends the number of ATs with redundant functions



present in *E. coli*. In the same way as many other ATs, the presence of UpaI and its positional ortholog EhaC do not correlate with the phylogenetic background or pathogenicity of individual isolates and thus these proteins represent fitness rather than specific virulence factors.

In summary, *E. coli* phylogeny enforces strong restrictions on the flexible gene content of individual strains. Although there are examples of pathotype-specific virulence factors and parallel origin of specific pathovars via HGT, e.g. Shiga toxin in EHEC or the locus of enterocyte effacement (LEE) in EHEC/EPEC, AT proteins don't belong to this group. Certainly, adaptation of individual isolates to their specific habitats results in strong selection pressures, especially in the interplay between host and bacterium, hence the convergent evolution in different phylogroups. But, if AT proteins contribute to overall colonization/survival fitness in commensals, the selection pressure lies outside of pathotype classifications.

Acknowledgements

We thank B. Plaschke (Würzburg) for technical support, A. Bauwens (Münster) for the introduction to fluorescence microscopy. We also acknowledge M. Bielaszewska (Münster) for the kind gift of a Cy-3-labeled anti-mouse secondary antibody. UD was supported by the German Research Foundation (DO 789/4-1). The study was carried out within the 'UTI Interference' consortium (ERA-NET Pathogenomics II, Federal Ministry of Education and Research (BMBF) grant no. 0315436A.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ijmm.2013.10.006.

References

Abreu, A.G., Bueris, V., Porangaba, T.M., Sircili, M.P., Navarro-Garcia, F., Elias, W.P., 2012. Autotransporter protein-encoding genes of diarrheagenic *Escherichia coli* are found in both typical and atypical enteropathogenic E. coli. Appl. Environ. Microbiol. 2 (October), 1–17.

Ackermann, N., Tiller, M., Anding, G., Roggenkamp, A., Heesemann, J., 2008. Contribution of trimeric autotransporter C-terminal domains of oligomeric coiled-coil adhesin (Oca) family members YadA, UspA1, EibA, and Hia to translocation of the YadA passenger domain and virulence of *Yersinia enterocolitica*. J. Bacteriol. 190 (14), 5031–5043.

Allsopp, L.P., Beloin, C., Ulett, G.C., Valle, J., Totsika, M., Sherlock, O., Ghigo, J.-M., Schembri, M.A., 2012. Molecular characterization of UpaB and UpaC, two new autotransporter proteins of uropathogenic *Escherichia coli* CFT073. Infect. Immun. 80 (1), 321–332.

Allsopp, L.P., Totsika, M., Tree, J.J., Ulett, G.C., Mabbett, A.N., Wells, T.J., Kobe, B., Beatson, S., Schembri, M.A., 2010. UpaH is a newly identified autotransporter protein that contributes to biofilm formation and bladder colonization by uropathogenic *Escherichia coli* CFT073. Infect. Immun. 78 (4), 1659–1669.

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.

Benz, I., Schmidt, M.A., 1992. Isolation and serologic characterization of AIDA-I, the adhesin mediating the diffuse adherence phenotype of the diarrhea-associated *Escherichia coli* strain 2787 (O126:H27). Infect. Immun. 60(1)(13), 13–18.

Benz, I., Schmidt, M.A., 2011. Structures and functions of autotransporter proteins in microbial pathogens. Int. J. Med. Microbiol. 301 (6), 461–468.

Brockmeyer, J., Spelten, S., Kuczius, T., Bielaszewska, M., Karch, H., 2009. Structure and function relationship of the autotransport and proteolytic activity of EspP from Shiga toxin-producing *Escherichia coli*. PLoS One 4 (7), e6100.

Brzuszkiewicz, E., Brüggemann, H., Liesegang, H., Emmerth, M., Ölschläger, T., Nagy, Gábor, Albermann, K., Wagner, C., Buchrieser, C., Emődy, L., Gottschalk, G., Hacker, J., Dobrindt, U., 2006. How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains. Proc. Natl. Acad. Sci. U.S.A. 103 (34), 12879–12884.

Celik, N., Webb, C.T., Leyton, D.L., Holt, K.E., Heinz, E., Gorrell, R., Kwok, T., Naderer, T., Strugnell, R.a., Speed, T.P., Teasdale, R.D., Likić, V.a., Lithgow, T., 2012. A bioinformatic strategy for the detection, classification and analysis of bacterial autotransporters. PLoS ONE 7 (8), e43245.

Chaudhuri, R.R., Henderson, I.R., 2012. The evolution of the *Escherichia coli* phylogeny. Infect. Genet. Evol. 12 (2), 214–226.

Clermont, O., Bonacorsi, S., Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl. Environ. Microbiol. 66 (10), 4555–4558.

Corander, J., Connor, T.R., O'Dwyer, C.A., Kroll, J.S., Hanage, W.P., 2012. Population structure in the *Neisseria*, and the biological significance of fuzzy species. J. R. Soc. Interface 9 (71), 1208–1215.

Cotter, S.E., Yeo, H., Juehne, T., St Geme, J.W., 2005. Architecture and adhesive activity of the *Haemophilus influenzae* Hsf adhesin. J. Bacteriol. 187 (13), 4656–4664.

Crémet, L., Corvec, S., Batard, E., Auger, M., Lopez, I., Pagniez, F., Dauvergne, S., Caroff, N., 2013. Comparison of three methods to study biofilm formation by clinical strains of *Escherichia coli*. Diagn. Microbiol. Infect. Dis. 75, 252–255.

Dautin, N., 2010. Serine protease autotransporters of *Enterobacteriaceae* (SPATEs): biogenesis and function. Toxins (Basel) 2 (6), 1179–1206.

Desvaux, M.P., Nicholas, J., Henderson, I.R., 2004. The autotransporter secretion system. Res. Microbiol. 155 (2), 53–60.

Didelot, X., Méric, G., Falush, D., Darling, A.E., 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics 13, 256.

Easton, D.M., Totsika, M., Allsopp, L.P., Phan, M.-D., Idris, A., Wurpel, D.J., Sherlock, O., Zhang, B., Venturini, C., Beatson, S., a, Mahony, T.J., Cobbold, R.N., Schembri, M.A., 2011. Characterization of EhaJ, a new autotransporter protein from enterohemorrhagic and enteropathogenic *Escherichia coli*. Frontiers Microbiol. 2, 120.

Le Gall, T., Clermont, Olivier, Gouriou, S., Picard, B., Nassif, X., Denamur, E., Tenaillon, O., 2007. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol. Biol. Evol. 24 (11), 2373–2384.

Grijpstra, J., Arenas, J., Rutten, L., Tommassen, J., 2013. Autotransporter secretion: varying on a theme. Res Microbiol. 164, 562–582.

Hammer, O., Harper, D.A.T., Ryan, P.D., 2001. PAST: paleontological statistics software package for education and data analysis. Palaeontologia Electronica 4 (1), 9.

Hasman, H., Chakraborty, T., Klemm, P., 1999. Antigen-43-mediated autoaggregation of *Escherichia coli* is blocked by fimbriation. J. Bacteriol. 181 (16), 4834–4841.

Henderson, I.R., Nataro, J.P., 2001. Virulence functions of autotransporter proteins. Infect. Immun. 69 (3), 1231–1243.

Henderson, I.R., Meehan, M., Owen, P., 2006. Antigen 43, a phase-variable bipartite outer membrane protein, determines colony morphology and autoaggregation in *Escherichia coli* K-12. FEMS Microbiology Letters 149 (1), 115–120.

Henderson, I.R., Navarro-Garcia, F., 2004. Type V protein secretion pathway: the autotransporter story. Society 68 (4), 692–744.

Holland, I.B., 2010. The extraordinary diversity of bacterial protein secretion mechanisms. Methods Mol. Biol. 619, 1–20.

Johnson, J.R., Delavari, P., Kuskowski, M., Stell, a.L., 2001. Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*. J. Infect. Dis. 183 (1), 78–88.

Kaper, J.B., Nataro, J.P., Mobley, H.L., 2004. Pathogenic *Escherichia coli*. Nat. Rev. Microbiol. 2 (2), 123–140.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., Salzberg, S.L., 2004. Versatile and open software for comparing large genomes. Genome Biol. 5 (2), R12.

Köhler, C.-D., Dobrindt, U., 2011. What defines extraintestinal pathogenic *Escherichia coli*? Int. J. Med. Microbiol. 301 (8), 642–647.

Laemmli, U.K., 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227 (5259), 680–685.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23 (21), 2947–2948.

Legendre, P., Legendre, L., 1998. Numerical Ecology, 2nd Edition. Elsevier Science, BV, Elsevier, Amsterdam, Netherlands.

Leimbach, A., Hacker, J., Dobrindt, U., 2013. *E. coli* as an all-rounder: The thin line between commensalism and pathogenicity. Curr. Top. Microbiol. Immunol 358, 3–32.

Leopold, S.R., Sawyer, S.A., Whittam, T.S., Tarr, P.I., 2011. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. BMC Evol. Biol. 11 (1), 183.

Letunic, I., Bork, P., 2011. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 39 (Web Server issue), W475–W478.

Linke, D., Riess, T., Autenrieth, I.B., Lupas, A., Kempf, V.A.J., 2006. Trimeric autotransporter adhesins: variable structure, common function. Trends Microbiol. 14 (6), 264–270.

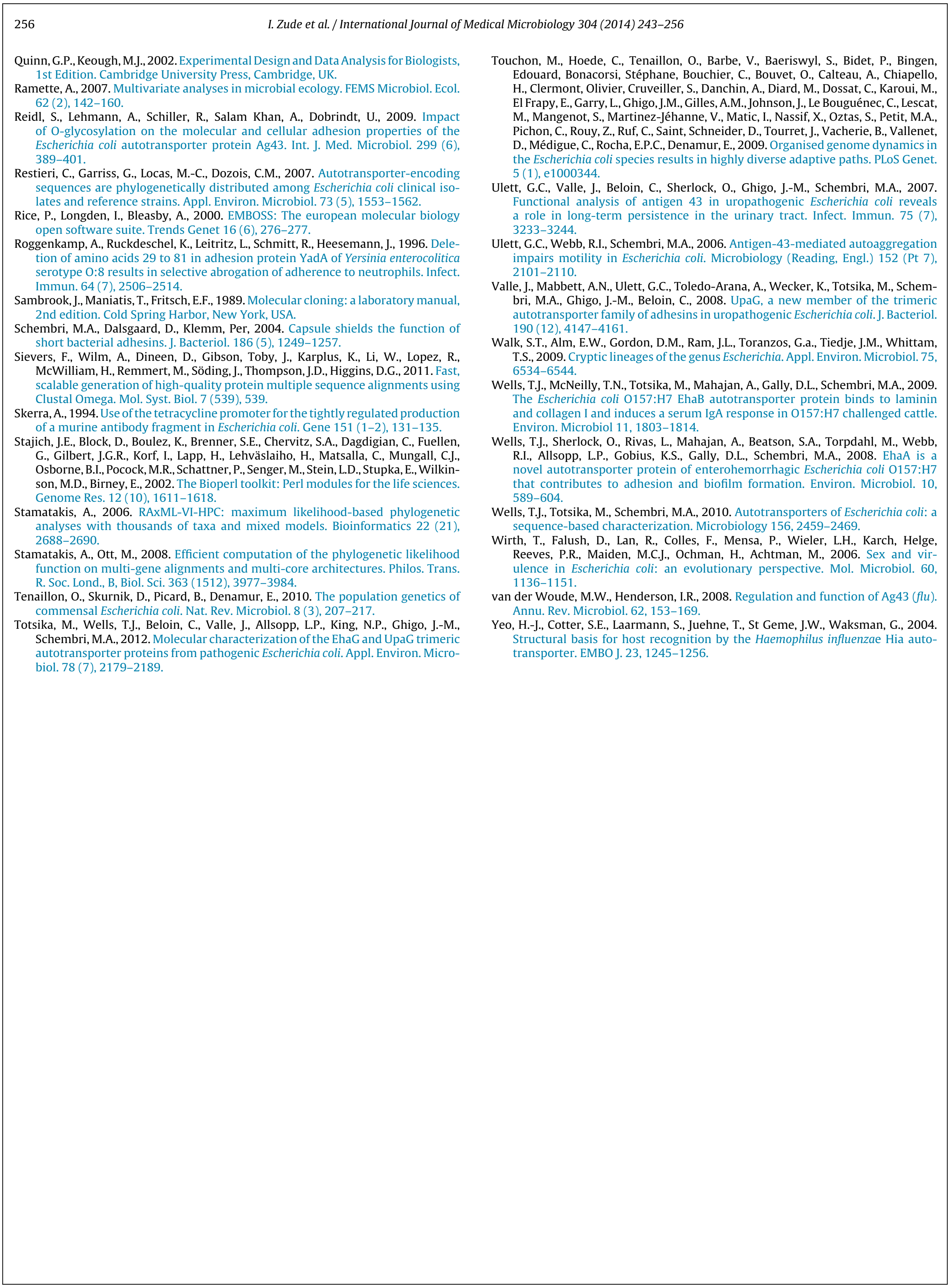
Lum, M., Morona, R., 2012. IcsA autotransporter passenger promotes increased fusion protein expression on the cell surface. Microb. Cell Fact. 11 (1), 20.

McNally, A., Cheng, L., Harris, S.R., Corander, J., 2013. The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome. Genome Biol. Evol. 5 (4), 699–710.

Moreno-Hagelsieb, G., Latimer, K., 2008. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics 24 (3), 319–324.

Ochman, H., Selander, R.K., 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. 157 (2), 690–693.

Picard, B., Garcia, J.S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, Edouard, Elion, J., Denamur, E., 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. Infect. Immun. 67 (2), 546–553.



Quinn, G.P., Keough, M.J., 2002. Experimental Design and Data Analysis for Biologists, 1st Edition. Cambridge University Press, Cambridge, UK.

Ramette, A., 2007. Multivariate analyses in microbial ecology. FEMS Microbiol. Ecol. 62 (2), 142–160.

Reidl, S., Lehmann, A., Schiller, R., Salam Khan, A., Dobrindt, U., 2009. Impact of O-glycosylation on the molecular and cellular adhesion properties of the *Escherichia coli* autotransporter protein Ag43. Int. J. Med. Microbiol. 299 (6), 389–401.

Restieri, C., Garriss, G., Locas, M.-C., Dozois, C.M., 2007. Autotransporter-encoding sequences are phylogenetically distributed among *Escherichia coli* clinical isolates and reference strains. Appl. Environ. Microbiol. 73 (5), 1553–1562.

Rice, P., Longden, I., Bleasby, A., 2000. EMBOSS: The european molecular biology open software suite. Trends Genet 16 (6), 276–277.

Roggenkamp, A., Ruckdeschel, K., Leitritz, L., Schmitt, R., Heesemann, J., 1996. Deletion of amino acids 29 to 81 in adhesion protein YadA of *Yersinia enterocolitica* serotype O:8 results in selective abrogation of adherence to neutrophils. Infect. Immun. 64 (7), 2506–2514.

Sambrook, J., Maniatis, T., Fritsch, E.F., 1989. Molecular cloning: a laboratory manual, 2nd edition. Cold Spring Harbor, New York, USA.

Schembri, M.A., Dalsgaard, D., Klemm, Per, 2004. Capsule shields the function of short bacterial adhesins. J. Bacteriol. 186 (5), 1249–1257.

Sievers, F., Wilm, A., Dineen, D., Gibson, Toby, J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7 (539), 539.

Skerra, A., 1994. Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in *Escherichia coli*. Gene 151 (1–2), 131–135.

Stajich, J.E., Block, D., Boulez, K., Brenner, S.E., Chervitz, S.A., Dagdigian, C., Fuellen, G., Gilbert, J.G.R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C.J., Osborne, B.I., Pocock, M.R., Schattner, P., Senger, M., Stein, L.D., Stupka, E., Wilkinson, M.D., Birney, E., 2002. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 12 (10), 1611–1618.

Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22 (21), 2688–2690.

Stamatakis, A., Ott, M., 2008. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 363 (1512), 3977–3984.

Tenaillon, O., Skurnik, D., Picard, B., Denamur, E., 2010. The population genetics of commensal *Escherichia coli*. Nat. Rev. Microbiol. 8 (3), 207–217.

Totsika, M., Wells, T.J., Beloin, C., Valle, J., Allsopp, L.P., King, N.P., Ghigo, J.-M., Schembri, M.A., 2012. Molecular characterization of the EhaG and UpaG trimeric autotransporter proteins from pathogenic *Escherichia coli*. Appl. Environ. Microbiol. 78 (7), 2179–2189.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, Edouard, Bonacorsi, Stéphane, Bouchier, C., Bouvet, O., Calteau, A., Chiapello, H., Clermont, Olivier, Cruveiller, S., Danchin, A., Diard, M., Dossat, C., Karoui, M., El Frapy, E., Garry, L., Ghigo, J.M., Gilles, A.M., Johnson, J., Le Bouguénec, C., Lescat, M., Mangenot, S., Martinez-Jéhanne, V., Matic, I., Nassif, X., Oztas, S., Petit, M.A., Pichon, C., Rouy, Z., Ruf, C., Saint, Schneider, D., Tourret, J., Vacherie, B., Vallenet, D., Médigue, C., Rocha, E.P.C., Denamur, E., 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 5 (1), e1000344.

Ulett, G.C., Valle, J., Beloin, C., Sherlock, O., Ghigo, J.-M., Schembri, M.A., 2007. Functional analysis of antigen 43 in uropathogenic *Escherichia coli* reveals a role in long-term persistence in the urinary tract. Infect. Immun. 75 (7), 3233–3244.

Ulett, G.C., Webb, R.I., Schembri, M.A., 2006. Antigen-43-mediated autoaggregation impairs motility in *Escherichia coli*. Microbiology (Reading, Engl.) 152 (Pt 7), 2101–2110.

Valle, J., Mabbett, A.N., Ulett, G.C., Toledo-Arana, A., Wecker, K., Totsika, M., Schembri, M.A., Ghigo, J.-M., Beloin, C., 2008. UpaG, a new member of the trimeric autotransporter family of adhesins in uropathogenic *Escherichia coli*. J. Bacteriol. 190 (12), 4147–4161.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.a., Tiedje, J.M., Whittam, T.S., 2009. Cryptic lineages of the genus *Escherichia*. Appl. Environ. Microbiol. 75, 6534–6544.

Wells, T.J., McNeilly, T.N., Totsika, M., Mahajan, A., Gally, D.L., Schembri, M.A., 2009. The *Escherichia coli* O157:H7 EhaB autotransporter protein binds to laminin and collagen I and induces a serum IgA response in O157:H7 challenged cattle. Environ. Microbiol 11, 1803–1814.

Wells, T.J., Sherlock, O., Rivas, L., Mahajan, A., Beatson, S.A., Torpdahl, M., Webb, R.I., Allsopp, L.P., Gobius, K.S., Gally, D.L., Schembri, M.A., 2008. EhaA is a novel autotransporter protein of enterohemorrhagic *Escherichia coli* O157:H7 that contributes to adhesion and biofilm formation. Environ. Microbiol. 10, 589–604.

Wells, T.J., Totsika, M., Schembri, M.A., 2010. Autotransporters of *Escherichia coli*: a sequence-based characterization. Microbiology 156, 2459–2469.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., Karch, Helge, Reeves, P.R., Maiden, M.C.J., Ochman, H., Achtman, M., 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol. Microbiol. 60, 1136–1151.

van der Woude, M.W., Henderson, I.R., 2008. Regulation and function of Ag43 (*flu*). Annu. Rev. Microbiol. 62, 153–169.

Yeo, H.-J., Cotter, S.E., Laarmann, S., Juehne, T., St Geme, J.W., Waksman, G., 2004. Structural basis for host recognition by the *Haemophilus influenzae* Hia autotransporter. EMBO J. 23, 1245–1256.

5.1.1.3 *Supplementary information*

The supplementary figures and tables for Zude et al. (2014) are presented on pages 96–103. Supplementary Table S1 (overview of the used *E. coli* strain panel) and Table S5 (statistics of the BLASTP hits) are too extensive for a reprint, but can be found with the original sources here: http://www.sciencedirect.com/science/article/pii/S1438422113001562

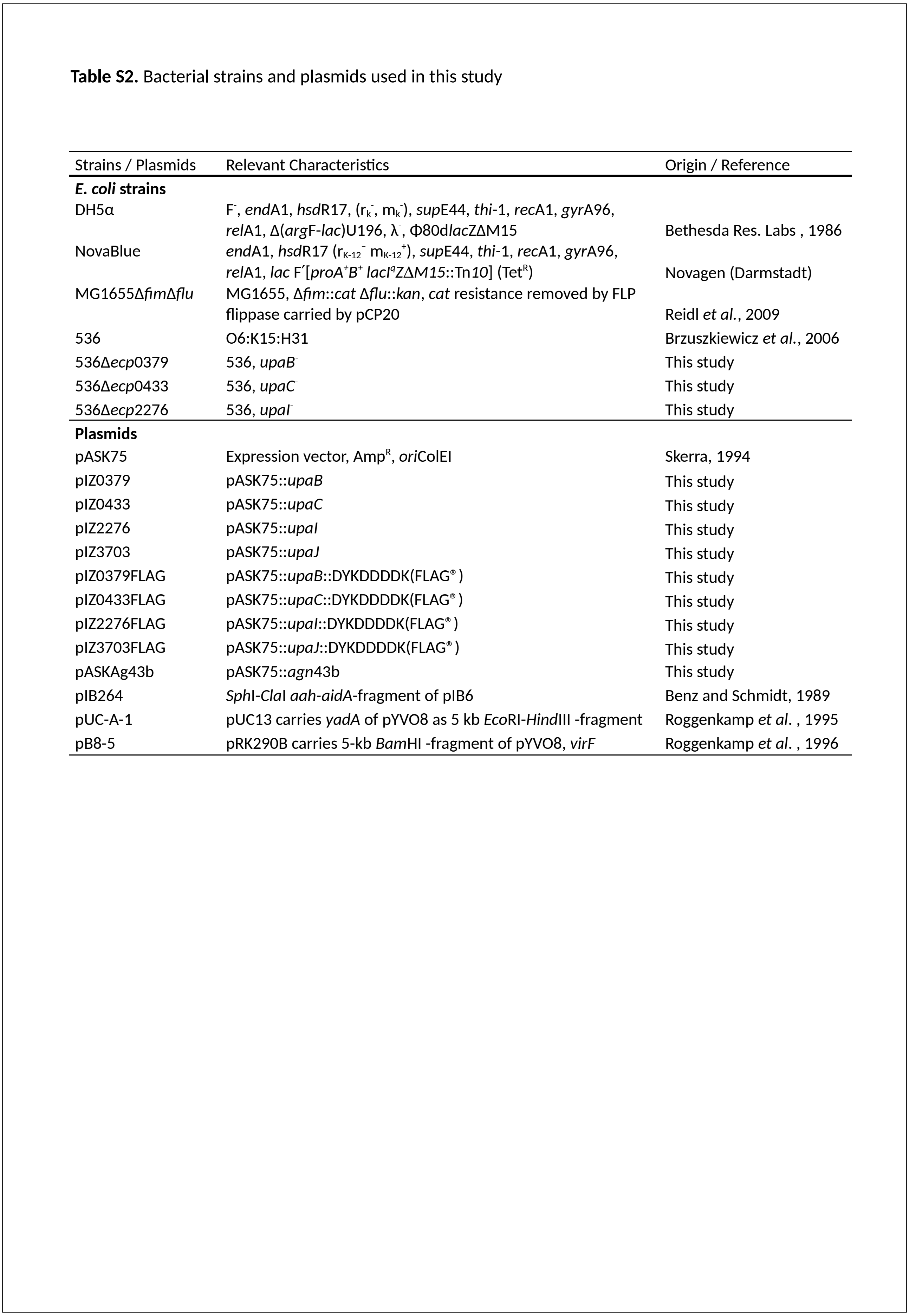


Table S2. Bacterial strains and plasmids used in this study

Strains / Plasmids	Relevant Characteristics	Origin / Reference
***E. coli* strains**		
DH5α	F^{-}, *end*A1, *hsd*R17, (r$_{k}^{-}$, m$_{k}^{-}$), *sup*E44, *thi*-1, *rec*A1, *gyr*A96, *rel*A1, Δ(*arg*F-*lac*)U196, λ^{-}, Φ80d*lac*ZΔM15	Bethesda Res. Labs , 1986
NovaBlue	*end*A1, *hsd*R17 (r$_{K\text{-}12}^{-}$ m$_{K\text{-}12}^{+}$), *sup*E44, *thi*-1, *rec*A1, *gyr*A96, *rel*A1, *lac* F′[*proA*$^{+}$*B*$^{+}$ *lacI*^{q}ZΔ*M15*::Tn*10*] (Tet^{R})	Novagen (Darmstadt)
MG1655Δ*fim*Δ*flu*	MG1655, Δ*fim*::*cat* Δ*flu*::*kan*, *cat* resistance removed by FLP flippase carried by pCP20	Reidl *et al.*, 2009
536	O6:K15:H31	Brzuszkiewicz *et al.*, 2006
536Δ*ecp*0379	536, *upaB*$^{-}$	This study
536Δ*ecp*0433	536, *upaC*$^{-}$	This study
536Δ*ecp*2276	536, *upaI*$^{-}$	This study
Plasmids		
pASK75	Expression vector, Amp^{R}, *ori*ColEI	Skerra, 1994
pIZ0379	pASK75::*upaB*	This study
pIZ0433	pASK75::*upaC*	This study
pIZ2276	pASK75::*upaI*	This study
pIZ3703	pASK75::*upaJ*	This study
pIZ0379FLAG	pASK75::*upaB*::DYKDDDDK(FLAG®)	This study
pIZ0433FLAG	pASK75::*upaC*::DYKDDDDK(FLAG®)	This study
pIZ2276FLAG	pASK75::*upaI*::DYKDDDDK(FLAG®)	This study
pIZ3703FLAG	pASK75::*upaJ*::DYKDDDDK(FLAG®)	This study
pASKAg43b	pASK75::*agn*43b	This study
pIB264	*Sph*I-*Cla*I *aah-aidA*-fragment of pIB6	Benz and Schmidt, 1989
pUC-A-1	pUC13 carries *yadA* of pYVO8 as 5 kb *Eco*RI-*Hind*III -fragment	Roggenkamp *et al.* , 1995
pB8-5	pRK290B carries 5-kb *Bam*HI -fragment of pYVO8, *virF*	Roggenkamp *et al.* , 1996

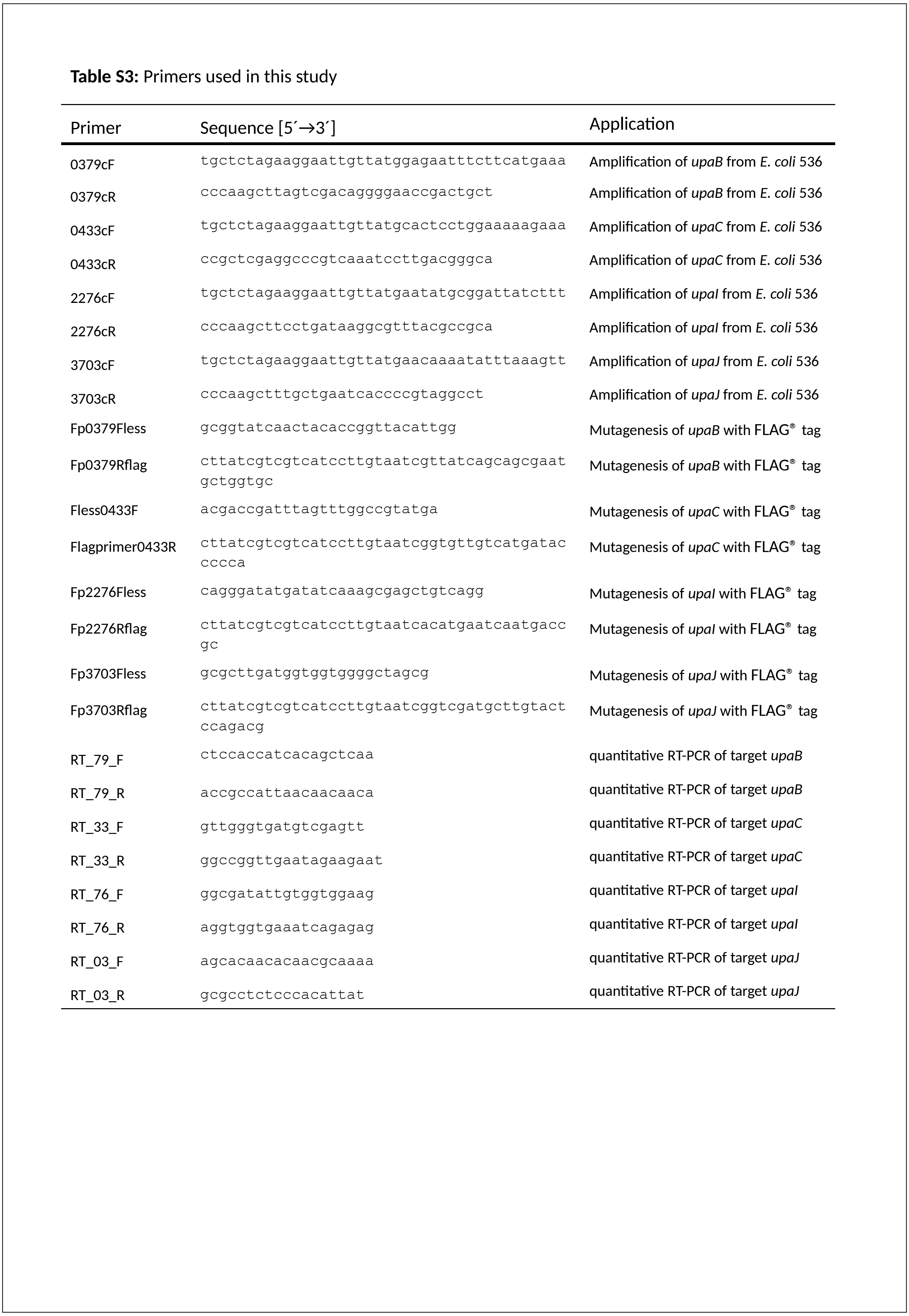


Table S3: Primers used in this study

Primer	Sequence [5´→3´]	Application
0379cF	tgctctagaaggaattgttatggagaatttcttcatgaaa	Amplification of *upaB* from *E. coli* 536
0379cR	cccaagcttagtcgacaggggaaccgactgct	Amplification of *upaB* from *E. coli* 536
0433cF	tgctctagaaggaattgttatgcactcctggaaaaagaaa	Amplification of *upaC* from *E. coli* 536
0433cR	ccgctcgaggcccgtcaaatccttgacgggca	Amplification of *upaC* from *E. coli* 536
2276cF	tgctctagaaggaattgttatgaatatgcggattatcttt	Amplification of *upaI* from *E. coli* 536
2276cR	cccaagcttcctgataaggcgtttacgccgca	Amplification of *upaI* from *E. coli* 536
3703cF	tgctctagaaggaattgttatgaacaaaatatttaaagtt	Amplification of *upaJ* from *E. coli* 536
3703cR	cccaagctttgctgaatcaccccgtaggcct	Amplification of *upaJ* from *E. coli* 536
Fp0379Fless	gcggtatcaactacaccggttacattgg	Mutagenesis of *upaB* with FLAG® tag
Fp0379Rflag	cttatcgtcgtcatccttgtaatcgttatcagcagcgaatgctggtgc	Mutagenesis of *upaB* with FLAG® tag
Fless0433F	acgaccgatttagtttggccgtatga	Mutagenesis of *upaC* with FLAG® tag
Flagprimer0433R	cttatcgtcgtcatccttgtaatcggtgttgtcatgataccccca	Mutagenesis of *upaC* with FLAG® tag
Fp2276Fless	cagggatatgatatcaaagcgagctgtcagg	Mutagenesis of *upaI* with FLAG® tag
Fp2276Rflag	cttatcgtcgtcatccttgtaatcacatgaatcaatgaccgc	Mutagenesis of *upaI* with FLAG® tag
Fp3703Fless	gcgcttgatggtggtggggctagcg	Mutagenesis of *upaJ* with FLAG® tag
Fp3703Rflag	cttatcgtcgtcatccttgtaatcggtcgatgcttgtactccagacg	Mutagenesis of *upaJ* with FLAG® tag
RT_79_F	ctccaccatcacagctcaa	quantitative RT-PCR of target *upaB*
RT_79_R	accgccattaacaacaaca	quantitative RT-PCR of target *upaB*
RT_33_F	gttgggtgatgtcgagtt	quantitative RT-PCR of target *upaC*
RT_33_R	ggccggttgaatagaagaat	quantitative RT-PCR of target *upaC*
RT_76_F	ggcgatattgtggtggaag	quantitative RT-PCR of target *upaI*
RT_76_R	aggtggtgaaatcagagag	quantitative RT-PCR of target *upaI*
RT_03_F	agcacaacacaacgcaaaa	quantitative RT-PCR of target *upaJ*
RT_03_R	gcgcctctcccacattat	quantitative RT-PCR of target *upaJ*

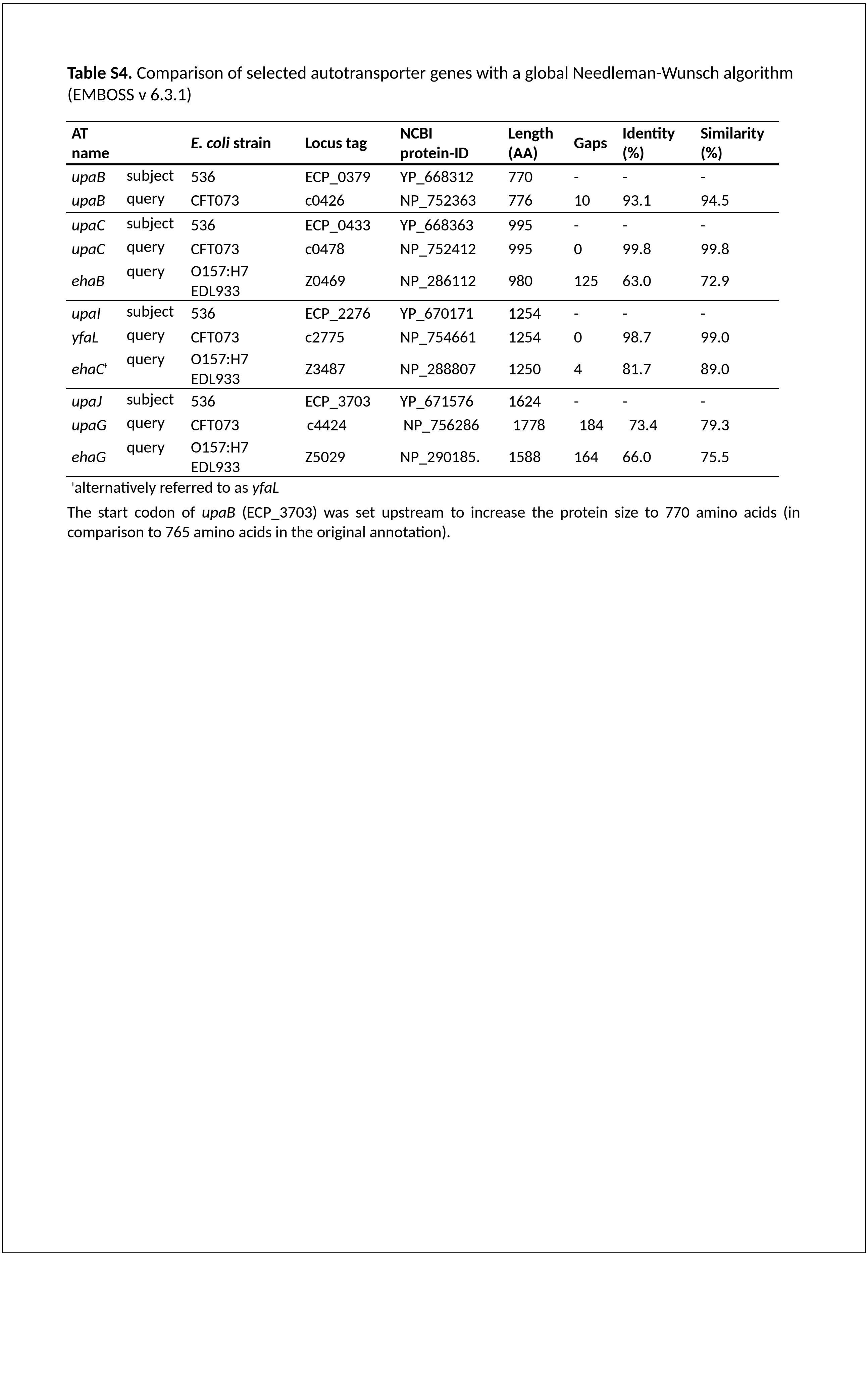


Table S4. Comparison of selected autotransporter genes with a global Needleman-Wunsch algorithm (EMBOSS v 6.3.1)

AT name		*E. coli* strain	Locus tag	NCBI protein-ID	Length (AA)	Gaps	Identity (%)	Similarity (%)
upaB	subject	536	ECP_0379	YP_668312	770	-	-	-
upaB	query	CFT073	c0426	NP_752363	776	10	93.1	94.5
upaC	subject	536	ECP_0433	YP_668363	995	-	-	-
upaC	query	CFT073	c0478	NP_752412	995	0	99.8	99.8
ehaB	query	O157:H7 EDL933	Z0469	NP_286112	980	125	63.0	72.9
upaI	subject	536	ECP_2276	YP_670171	1254	-	-	-
yfaL	query	CFT073	c2775	NP_754661	1254	0	98.7	99.0
*ehaC*ˈ	query	O157:H7 EDL933	Z3487	NP_288807	1250	4	81.7	89.0
upaJ	subject	536	ECP_3703	YP_671576	1624	-	-	-
upaG	query	CFT073	c4424	NP_756286	1778	184	73.4	79.3
ehaG	query	O157:H7 EDL933	Z5029	NP_290185.	1588	164	66.0	75.5

ˈalternatively referred to as *yfaL*

The start codon of *upaB* (ECP_3703) was set upstream to increase the protein size to 770 amino acids (in comparison to 765 amino acids in the original annotation).

Figure S1A

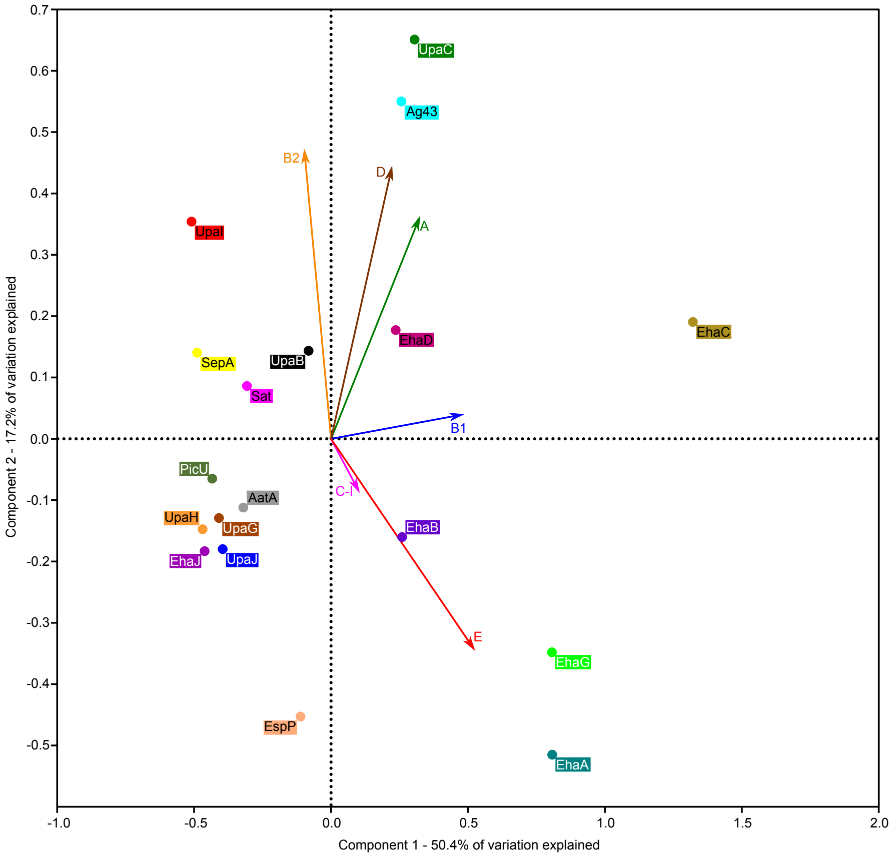


Figure S1B

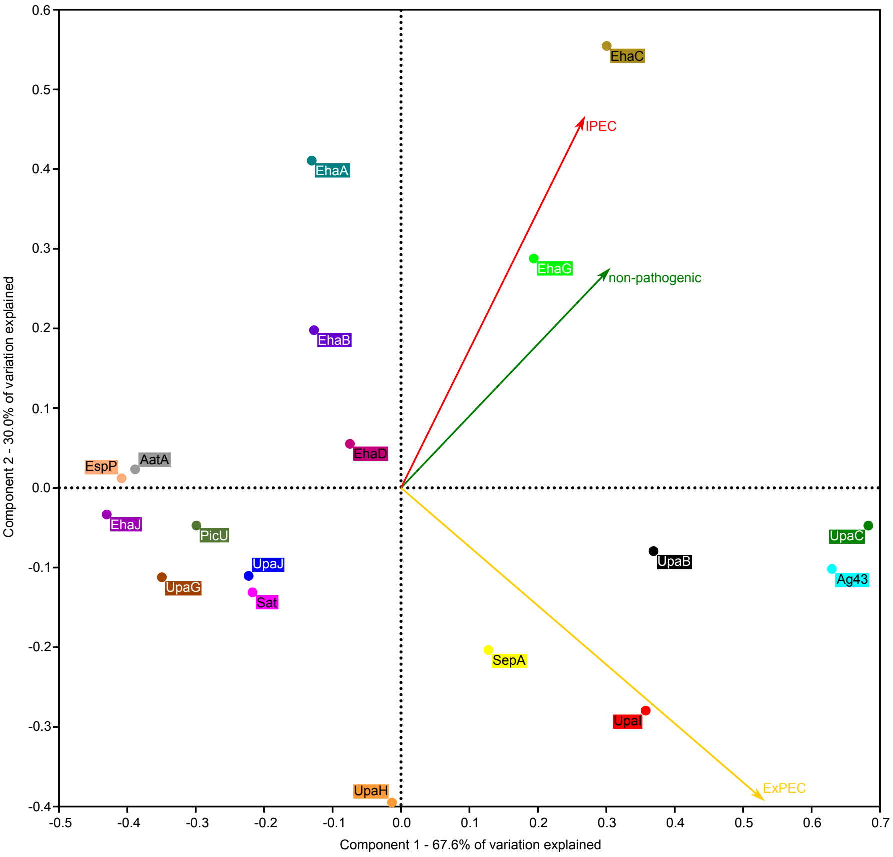


Figure S2A

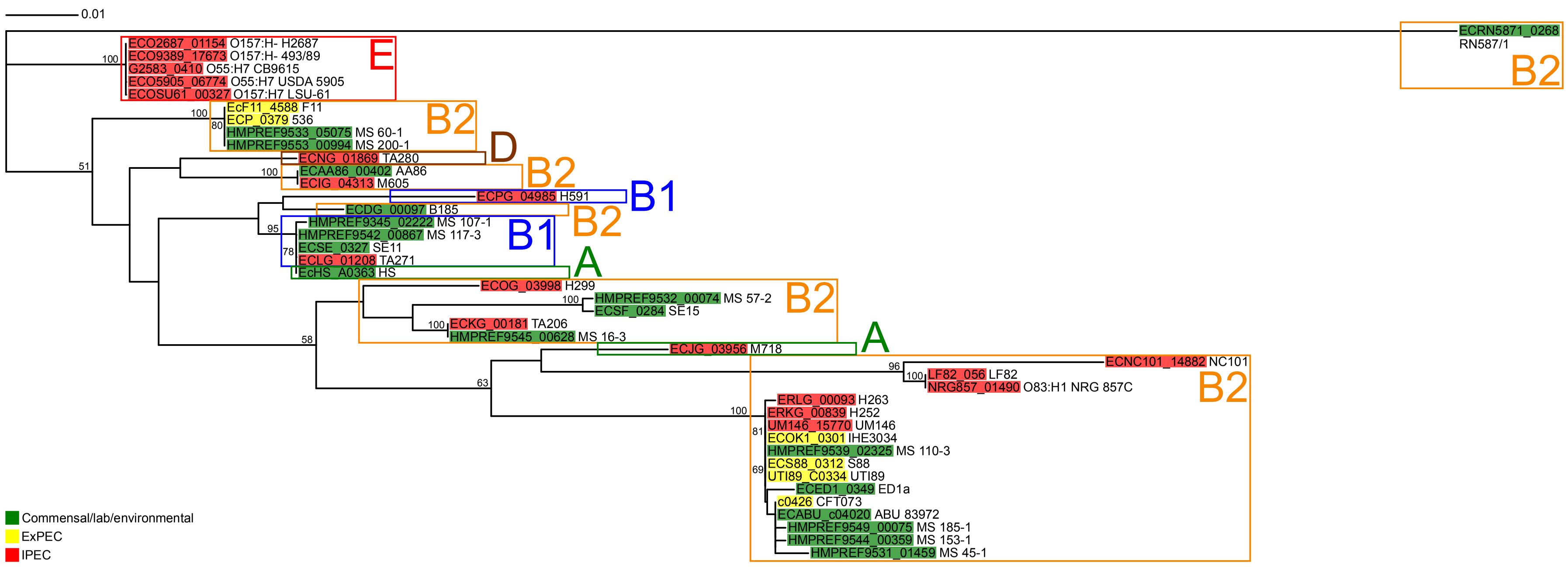

0.01
ECO2687_01154 O157:H- H2687
ECO9389_17673 O157:H- 493/89
G2583_0410 O55:H7 CB9615
ECO5905_06774 O55:H7 USDA 5905
ECOSU61_00327 O157:H7 LSU-61
E
EcF11_4588 F11
ECP_0379 536
HMPREF9533_05075 MS 60-1
HMPREF9553_00994 MS 200-1
B2
ECNG_01869 TA280
D
ECAA86_00402 AA86
ECIG_04313 M605
B2
ECPG_04985 H591
B1
ECDG_00097 B185
B2
HMPREF9345_02222 MS 107-1
HMPREF9542_00867 MS 117-3
ECSE_0327 SE11
ECLG_01208 TA271
B1
EcHS_A0363 HS
A
ECOG_03998 H299
HMPREF9532_00074 MS 57-2
ECSF_0284 SE15
ECKG_00181 TA206
HMPREF9545_00628 MS 16-3
B2
ECJG_03956 M718
A
ECNC101_14882 NC101
LF82_056 LF82
NRG857_01490 O83:H1 NRG 857C
ERLG_00093 H263
ERKG_00839 H252
UM146_15770 UM146
ECOK1_0301 IHE3034
HMPREF9539_02325 MS 110-3
ECS88_0312 S88
UTI89_C0334 UTI89
ECED1_0349 ED1a
c0426 CFT073
ECABU_c04020 ABU 83972
HMPREF9549_00075 MS 185-1
HMPREF9544_00359 MS 153-1
HMPREF9531_01459 MS 45-1
B2
ECRN5871_0268
RN587/1
B2
100
100
80
51
100
95
78
100
100
58
96
100
63
100
81
69
🟩 Commensal/lab/environmental
🟨 ExPEC
🟥 IPEC

Figure S2B

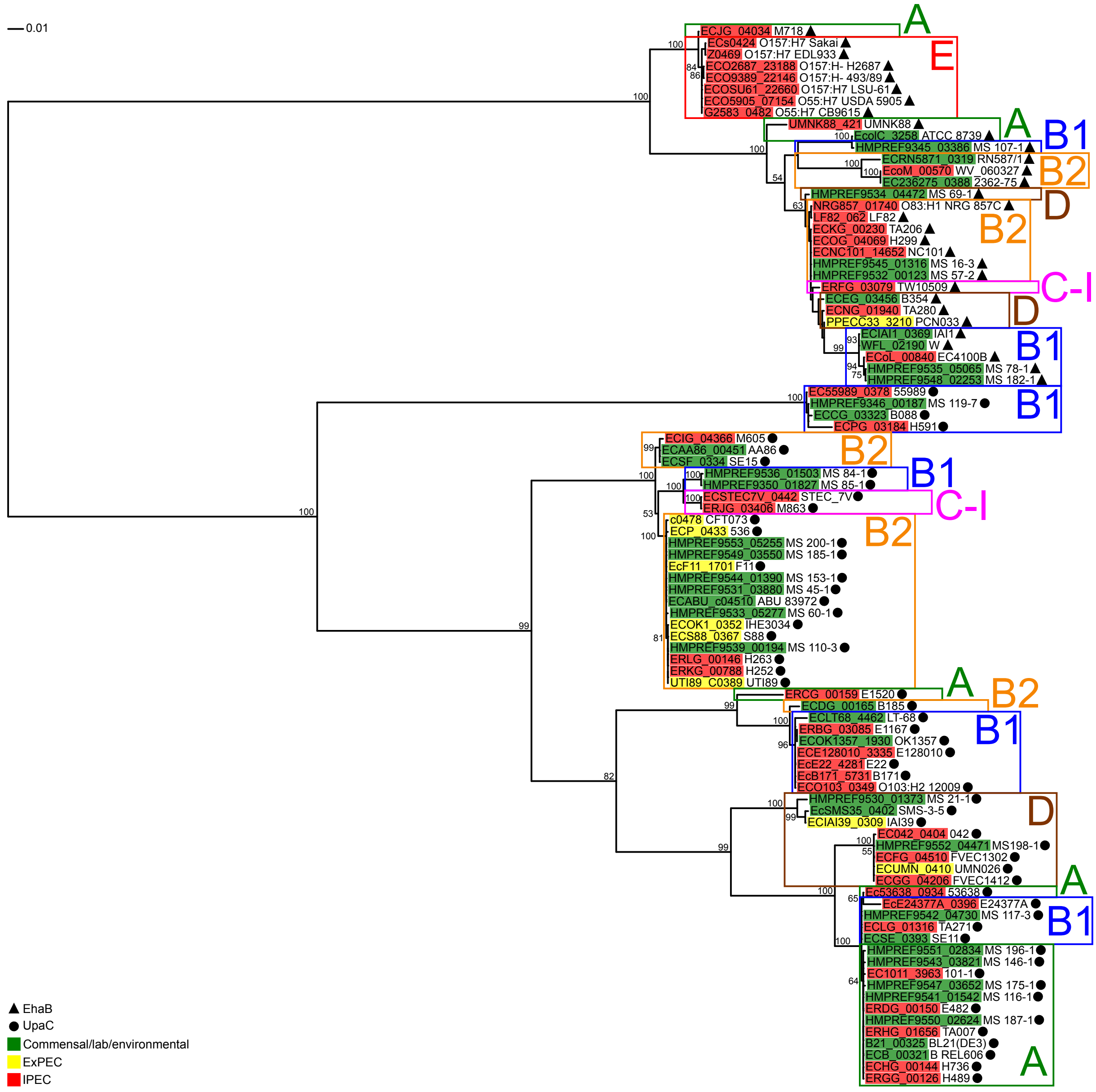

0.01
A
E
ECJG_04034 M718 ▲
ECs0424 O157:H7 Sakai ▲
Z0469 O157:H7 EDL933 ▲
ECO2687_23188 O157:H- H2687 ▲
ECO9389_22146 O157:H- 493/89 ▲
ECOSU61_22660 O157:H7 LSU-61 ▲
ECO5905_07154 O55:H7 USDA 5905 ▲
G2583_0482 O55:H7 CB9615 ▲
100
84
86
100
A
B1
UMNK88_421 UMNK88 ▲
EcolC_3258 ATCC 8739 ▲
HMPREF9345_03386 MS 107-1 ▲
B2
ECRN5871_0319 RN587/1 ▲
EcoM_00570 WV_060327 ▲
EC236275_0388 2362-75 ▲
100
100
100
54
D
HMPREF9534_04472 MS 69-1 ▲
63
B2
NRG857_01740 O83:H1 NRG 857C ▲
LF82_062 LF82 ▲
ECKG_00230 TA206 ▲
ECOG_04069 H299 ▲
ECNC101_14652 NC101 ▲
HMPREF9545_01316 MS 16-3 ▲
HMPREF9532_00123 MS 57-2 ▲
C-I
ERFG_03079 TW10509 ▲
D
ECEG_03456 B354 ▲
ECNG_01940 TA280 ▲
PPECC33_3210 PCN033 ▲
B1
ECIAI1_0369 IAI1 ▲
WFL_02190 W ▲
ECoL_00840 EC4100B ▲
HMPREF9535_05065 MS 78-1 ▲
HMPREF9548_02253 MS 182-1 ▲
93
99
94
75
B1
EC55989_0378 55989 ●
HMPREF9346_00187 MS 119-7 ●
ECCG_03323 B088 ●
ECPG_03184 H591 ●
100
B2
ECIG_04366 M605 ●
ECAA86_00451 AA86 ●
ECSF_0334 SE15 ●
99
B1
HMPREF9536_01503 MS 84-1 ●
HMPREF9350_01827 MS 85-1 ●
C-I
ECSTEC7V_0442 STEC_7V ●
ERJG_03406 M863 ●
100
100
100
100
53
B2
c0478 CFT073 ●
ECP_0433 536 ●
HMPREF9553_05255 MS 200-1 ●
HMPREF9549_03550 MS 185-1 ●
EcF11_1701 F11 ●
HMPREF9544_01390 MS 153-1 ●
HMPREF9531_03880 MS 45-1 ●
ECABU_c04510 ABU 83972 ●
HMPREF9533_05277 MS 60-1 ●
ECOK1_0352 IHE3034 ●
ECS88_0367 S88 ●
HMPREF9539_00194 MS 110-3 ●
ERLG_00146 H263 ●
ERKG_00788 H252 ●
UTI89_C0389 UTI89 ●
100
81
100
99
A
B2
ERCG_00159 E1520 ●
ECDG_00165 B185 ●
B1
ECLT68_4462 LT-68 ●
ERBG_03085 E1167 ●
ECOK1357_1930 OK1357 ●
ECE128010_3335 E128010 ●
EcE22_4281 E22 ●
EcB171_5731 B171 ●
ECO103_0349 O103:H2 12009 ●
99
100
96
82
D
HMPREF9530_01373 MS 21-1 ●
EcSMS35_0402 SMS-3-5 ●
ECIAI39_0309 IAI39 ●
EC042_0404 042 ●
HMPREF9552_04471 MS198-1 ●
ECFG_04510 FVEC1302 ●
ECUMN_0410 UMN026 ●
ECGG_04206 FVEC1412 ●
100
99
100
55
99
A
Ec53638_0934 53638 ●
B1
EcE24377A_0396 E24377A ●
HMPREF9542_04730 MS 117-3 ●
ECLG_01316 TA271 ●
ECSE_0393 SE11 ●
100
65
100
HMPREF9551_02834 MS 196-1 ●
HMPREF9543_03821 MS 146-1 ●
EC1011_3963 101-1 ●
HMPREF9547_03652 MS 175-1 ●
HMPREF9541_01542 MS 116-1 ●
ERDG_00150 E482 ●
HMPREF9550_02624 MS 187-1 ●
ERHG_01656 TA007 ●
B21_00325 BL21(DE3) ●
ECB_00321 B REL606 ●
ECHG_00144 H736 ●
ERGG_00126 H489 ●
64
A
▲ EhaB
● UpaC
🟩 Commensal/lab/environmental
🟨 ExPEC
🟥 IPEC

Figure S2C

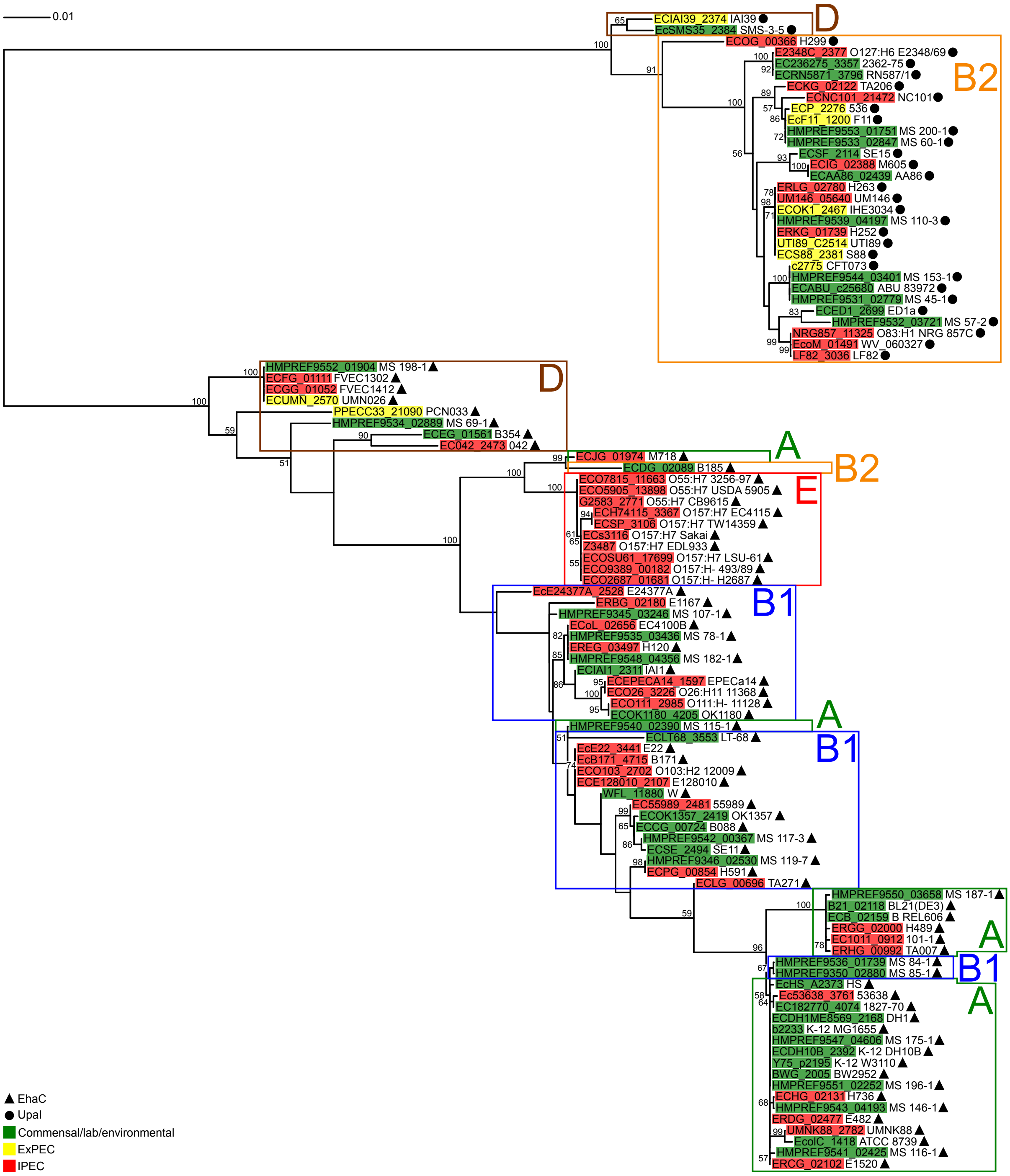
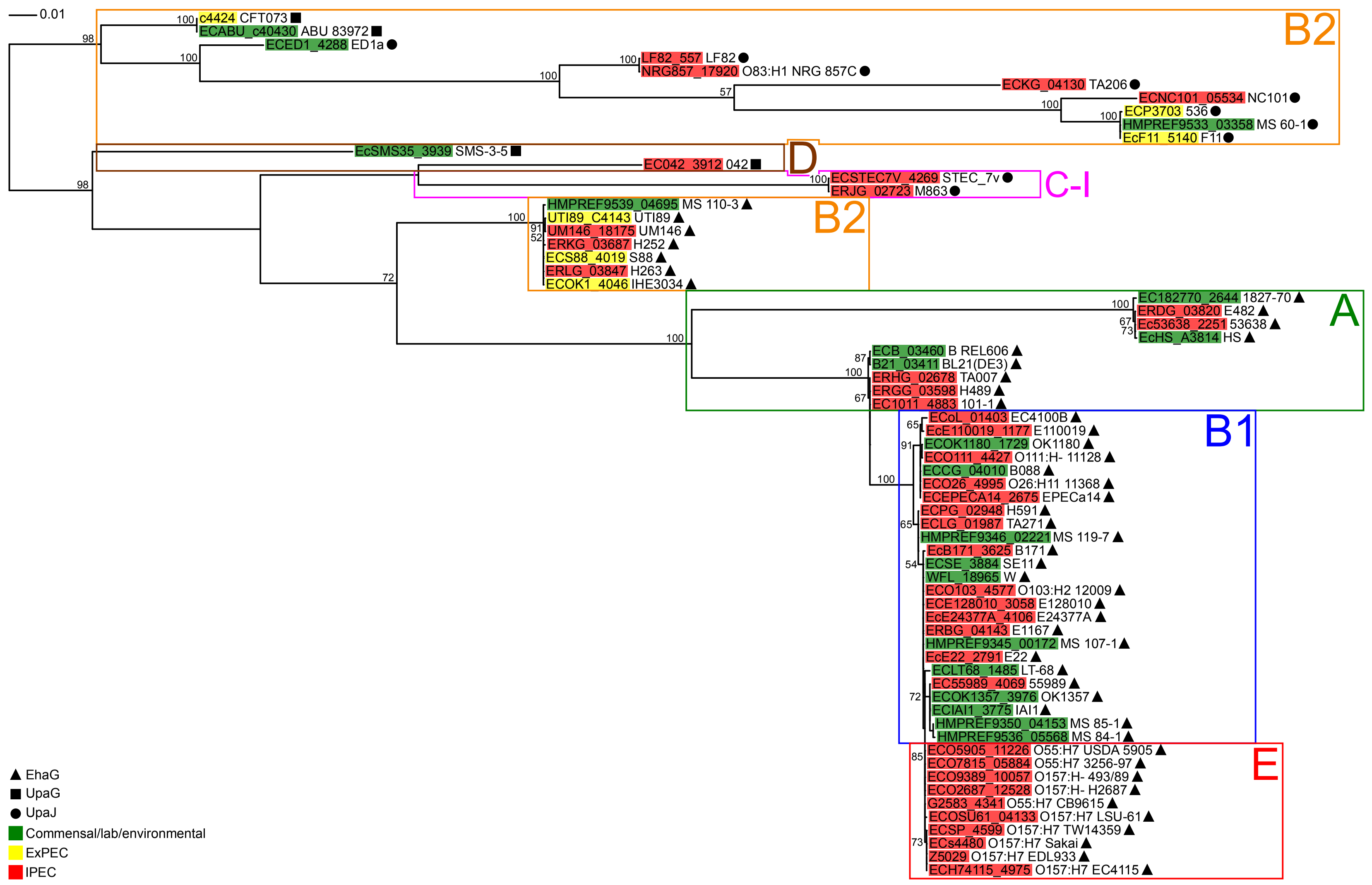

0.01
65
ECIAI39_2374 IAI39 ●
EcSMS35_2384 SMS-3-5 ●
D
100
ECOG_00366 H299 ●
E2348C_2377 O127:H6 E2348/69 ●
100
EC236275_3357 2362-75 ●
92
ECRN5871_3796 RN587/1 ●
B2
91
ECKG_02122 TA206 ●
89
ECNC101_21472 NC101 ●
100
57
ECP_2276 536 ●
86
EcF11_1200 F11 ●
HMPREF9553_01751 MS 200-1 ●
72
HMPREF9533_02847 MS 60-1 ●
56
ECSF_2114 SE15 ●
93
ECIG_02388 M605 ●
100
ECAA86_02439 AA86 ●
ERLG_02780 H263 ●
78
UM146_05640 UM146 ●
98
ECOK1_2467 IHE3034 ●
71
HMPREF9539_04197 MS 110-3 ●
ERKG_01739 H252 ●
UTI89_C2514 UTI89 ●
ECS88_2381 S88 ●
c2775 CFT073 ●
100
HMPREF9544_03401 MS 153-1 ●
ECABU_c25680 ABU 83972 ●
HMPREF9531_02779 MS 45-1 ●
83
ECED1_2699 ED1a ●
HMPREF9532_03721 MS 57-2 ●
NRG857_11325 O83:H1 NRG 857C ●
99
EcoM_01491 WV_060327 ●
99
LF82_3036 LF82 ●
100
HMPREF9552_01904 MS 198-1 ▲
ECFG_01111 FVEC1302 ▲
ECGG_01052 FVEC1412 ▲
D
100
ECUMN_2570 UMN026 ▲
PPECC33_21090 PCN033 ▲
HMPREF9534_02889 MS 69-1 ▲
59
90
ECEG_01561 B354 ▲
EC042_2473 042 ▲
51
A
99
ECJG_01974 M718 ▲
ECDG_02089 B185 ▲
B2
100
ECO7815_11663 O55:H7 3256-97 ▲
100
ECO5905_13898 O55:H7 USDA 5905 ▲
E
G2583_2771 O55:H7 CB9615 ▲
94
ECH74115_3367 O157:H7 EC4115 ▲
ECSP_3106 O157:H7 TW14359 ▲
61
ECs3116 O157:H7 Sakai ▲
65
Z3487 O157:H7 EDL933 ▲
100
ECOSU61_17699 O157:H7 LSU-61 ▲
55
ECO9389_00182 O157:H- 493/89 ▲
ECO2687_01681 O157:H- H2687 ▲
EcE24377A_2528 E24377A ▲
ERBG_02180 E1167 ▲
B1
HMPREF9345_03246 MS 107-1 ▲
ECoL_02656 EC4100B ▲
82
HMPREF9535_03436 MS 78-1 ▲
EREG_03497 H120 ▲
85
HMPREF9548_04356 MS 182-1 ▲
ECIAI1_2311 IAI1 ▲
86
95
ECEPECA14_1597 EPECa14 ▲
100
ECO26_3226 O26:H11 11368 ▲
95
ECO111_2985 O111:H- 11128 ▲
ECOK1180_4205 OK1180 ▲
A
HMPREF9540_02390 MS 115-1 ▲
51
ECLT68_3553 LT-68 ▲
B1
EcE22_3441 E22 ▲
EcB171_4715 B171 ▲
74
ECO103_2702 O103:H2 12009 ▲
ECE128010_2107 E128010 ▲
WFL_11880 W ▲
EC55989_2481 55989 ▲
99
ECOK1357_2419 OK1357 ▲
65
ECCG_00724 B088 ▲
HMPREF9542_00367 MS 117-3 ▲
86
ECSE_2494 SE11 ▲
98
HMPREF9346_02530 MS 119-7 ▲
ECPG_00854 H591 ▲
ECLG_00696 TA271 ▲
HMPREF9550_03658 MS 187-1 ▲
100
B21_02118 BL21(DE3) ▲
59
ECB_02159 B REL606 ▲
ERGG_02000 H489 ▲
A
EC1011_0912 101-1 ▲
96
78
ERHG_00992 TA007 ▲
67
HMPREF9536_01739 MS 84-1 ▲
HMPREF9350_02880 MS 85-1 ▲
B1
EcHS_A2373 HS ▲
58
Ec53638_3761 53638 ▲
64
EC182770_4074 1827-70 ▲
A
ECDH1ME8569_2168 DH1 ▲
b2233 K-12 MG1655 ▲
HMPREF9547_04606 MS 175-1 ▲
ECDH10B_2392 K-12 DH10B ▲
Y75_p2195 K-12 W3110 ▲
BWG_2005 BW2952 ▲
HMPREF9551_02252 MS 196-1 ▲
ECHG_02131 H736 ▲
68
HMPREF9543_04193 MS 146-1 ▲
ERDG_02477 E482 ▲
99
UMNK88_2782 UMNK88 ▲
EcolC_1418 ATCC 8739 ▲
HMPREF9541_02425 MS 116-1 ▲
57
ERCG_02102 E1520 ▲
▲ EhaC
● UpaI
🟩 Commensal/lab/environmental
🟨 ExPEC
🟥 IPEC

Figure S2D


0.01
B2
c4424 CFT073 ■
ECABU_c40430 ABU 83972 ■
ECED1_4288 ED1a ●
LF82_557 LF82 ●
NRG857_17920 O83:H1 NRG 857C ●
ECKG_04130 TA206 ●
ECNC101_05534 NC101 ●
ECP3703 536 ●
HMPREF9533_03358 MS 60-1 ●
EcF11_5140 F11 ●
EcSMS35_3939 SMS-3-5 ■
D
EC042_3912 042 ■
ECSTEC7V_4269 STEC_7v ●
ERJG_02723 M863 ●
C-I
B2
HMPREF9539_04695 MS 110-3 ▲
UTI89_C4143 UTI89 ▲
UM146_18175 UM146 ▲
ERKG_03687 H252 ▲
ECS88_4019 S88 ▲
ERLG_03847 H263 ▲
ECOK1_4046 IHE3034 ▲
A
EC182770_2644 1827-70 ▲
ERDG_03820 E482 ▲
Ec53638_2251 53638 ▲
EcHS_A3814 HS ▲
ECB_03460 B REL606 ▲
B21_03411 BL21(DE3) ▲
ERHG_02678 TA007 ▲
ERGG_03598 H489 ▲
EC1011_4883 101-1 ▲
B1
ECoL_01403 EC4100B ▲
EcE110019_1177 E110019 ▲
ECOK1180_1729 OK1180 ▲
ECO111_4427 O111:H- 11128 ▲
ECCG_04010 B088 ▲
ECO26_4995 O26:H11 11368 ▲
ECEPECA14_2675 EPECa14 ▲
ECPG_02948 H591 ▲
ECLG_01987 TA271 ▲
HMPREF9346_02221 MS 119-7 ▲
EcB171_3625 B171 ▲
ECSE_3884 SE11 ▲
WFL_18965 W ▲
ECO103_4577 O103:H2 12009 ▲
ECE128010_3058 E128010 ▲
EcE24377A_4106 E24377A ▲
ERBG_04143 E1167 ▲
HMPREF9345_00172 MS 107-1 ▲
EcE22_2791 E22 ▲
ECLT68_1485 LT-68 ▲
EC55989_4069 55989 ▲
ECOK1357_3976 OK1357 ▲
ECIAI1_3775 IAI1 ▲
HMPREF9350_04153 MS 85-1 ▲
HMPREF9536_05568 MS 84-1 ▲
E
ECO5905_11226 O55:H7 USDA 5905 ▲
ECO7815_05884 O55:H7 3256-97 ▲
ECO9389_10057 O157:H- 493/89 ▲
ECO2687_12528 O157:H- H2687 ▲
G2583_4341 O55:H7 CB9615 ▲
ECOSU61_04133 O157:H7 LSU-61 ▲
ECSP_4599 O157:H7 TW14359 ▲
ECs4480 O157:H7 Sakai ▲
Z5029 O157:H7 EDL933 ▲
ECH74115_4975 O157:H7 EC4115 ▲
98
100
100
100
100
57
100
100
98
100
72
100
91
52
100
100
67
73
87
100
67
65
91
100
65
54
72
85
73
▲ EhaG
■ UpaG
● UpaJ
🟩 Commensal/lab/environmental
🟨 ExPEC
🟥 IPEC

5.2 GENOMIC ANALYSES OF TWO STEC ISOLATES FROM THE GERMAN 2011 EPIDEMIC

In 2011 the largest and deadliest food-borne STEC epidemic to date descended upon Germany (Section 1.4 on page 48). Because of the enormous development in sequencing technologies and accompanying dramatic cost reduction (Section 1.1.1 on page 5), it was the first opportunity to analyze the genome of a bacterial pathogen in an *ongoing* epidemic. We were one of the teams that analyzed the virulence potential of the O104:H4 STEC genome in detail (Table 6 on page 198).

5.2.1 *Genome sequence analyses of two isolates from the recent* Escherichia coli *outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic* Escherichia coli *(EAHEC)*

** Authors contributed equally*

Brzuszkiewicz E*, Thürmer A*, Schuldes J*, LEIMBACH A*, Liesegang H*, Meyer F-D, Boelter J, Petersen H, Gottschalk G, Daniel R. 2011. Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC). *Arch. Microbiol.* 193:883–891.
DOI: 10.1007/s00203-011-0725-6

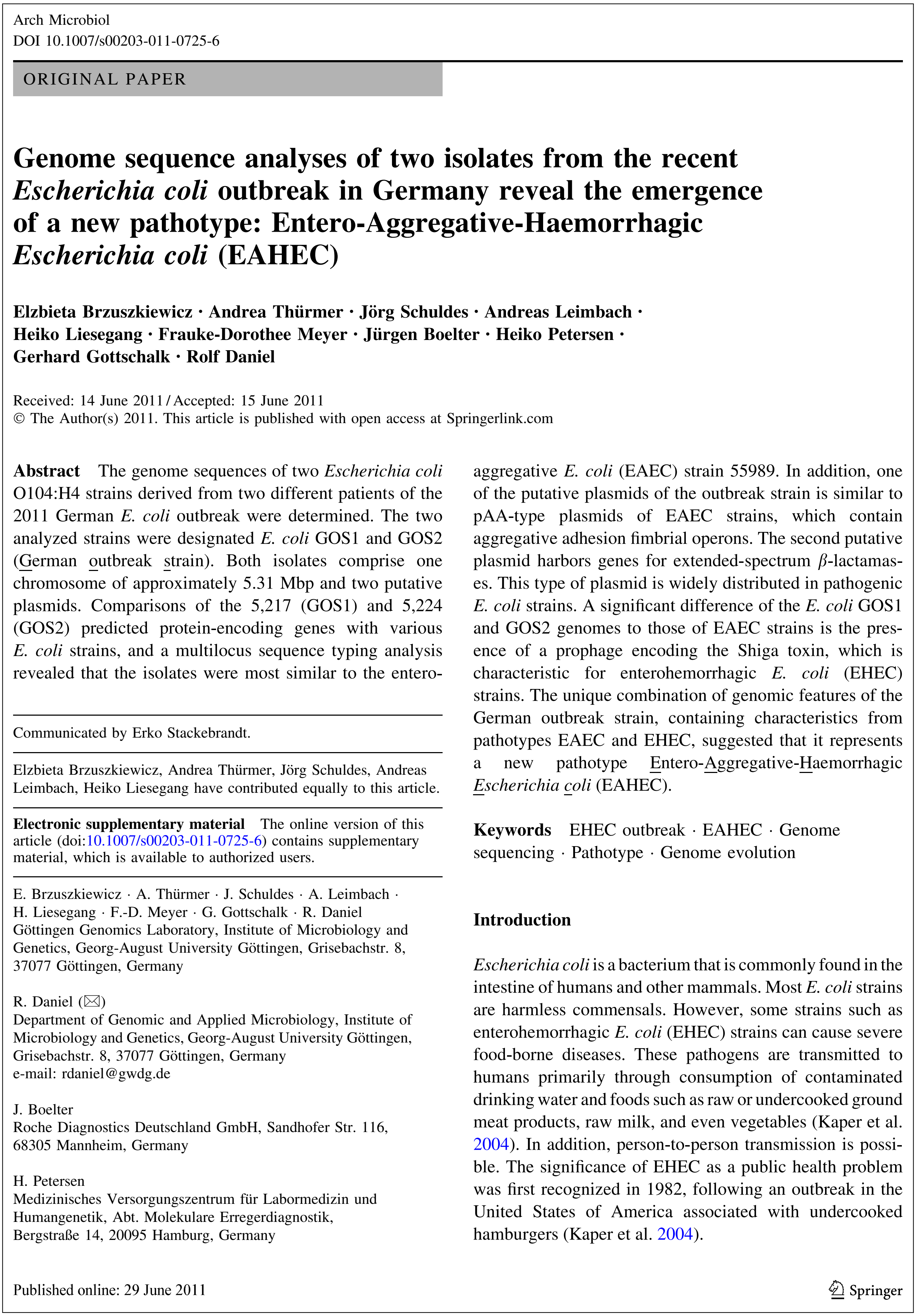
5.2.1.1 *Contributions*

Brzuszkiewicz et al. (2011) presents the genomic sequences and analyses of two STEC isolates (GOS1 and GOS2) from the food-borne epidemic (Section 1.4 on page 48). Because the isolates contained virulence characteristics from two *E. coli* pathotypes, EAEC and EHEC, we suggested a new *E. coli* pathotype acronym: entero-aggregative-haemorrhagic *E. coli* (EAHEC).

I contributed to the bioinformatical analyses and study design of the publication, especially identification and visualization of the respective VFs and prophages, and the phylogeny based on MLST. I was involved in all parts of the writing process of the manuscript. Detailed individual author contributions for each part of the paper and each figure/table can be found in Table 12 and Table 13 on page 231, respectively.

5.2.1.2 *Main paper*

This open access publication can be found on pages 105–113 or freely available and to reuse (licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License (CC BY-NC 3.0) cc BY-NC) at: `https://link.springer.com/article/10.1007%2Fs00203-011-0725-6`



Arch Microbiol
DOI 10.1007/s00203-011-0725-6

ORIGINAL PAPER

Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC)

Elzbieta Brzuszkiewicz · Andrea Thürmer · Jörg Schuldes · Andreas Leimbach · Heiko Liesegang · Frauke-Dorothee Meyer · Jürgen Boelter · Heiko Petersen · Gerhard Gottschalk · Rolf Daniel

Received: 14 June 2011 / Accepted: 15 June 2011

Abstract The genome sequences of two *Escherichia coli* O104:H4 strains derived from two different patients of the 2011 German *E. coli* outbreak were determined. The two analyzed strains were designated *E. coli* GOS1 and GOS2 (German outbreak strain). Both isolates comprise one chromosome of approximately 5.31 Mbp and two putative plasmids. Comparisons of the 5,217 (GOS1) and 5,224 (GOS2) predicted protein-encoding genes with various *E. coli* strains, and a multilocus sequence typing analysis revealed that the isolates were most similar to the entero-aggregative *E. coli* (EAEC) strain 55989. In addition, one of the putative plasmids of the outbreak strain is similar to pAA-type plasmids of EAEC strains, which contain aggregative adhesion fimbrial operons. The second putative plasmid harbors genes for extended-spectrum β-lactamases. This type of plasmid is widely distributed in pathogenic *E. coli* strains. A significant difference of the *E. coli* GOS1 and GOS2 genomes to those of EAEC strains is the presence of a prophage encoding the Shiga toxin, which is characteristic for enterohemorrhagic *E. coli* (EHEC) strains. The unique combination of genomic features of the German outbreak strain, containing characteristics from pathotypes EAEC and EHEC, suggested that it represents a new pathotype Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC).

Keywords EHEC outbreak · EAHEC · Genome sequencing · Pathotype · Genome evolution

Introduction

Escherichia coli is a bacterium that is commonly found in the intestine of humans and other mammals. Most *E. coli* strains are harmless commensals. However, some strains such as enterohemorrhagic *E. coli* (EHEC) strains can cause severe food-borne diseases. These pathogens are transmitted to humans primarily through consumption of contaminated drinking water and foods such as raw or undercooked ground meat products, raw milk, and even vegetables (Kaper et al. 2004). In addition, person-to-person transmission is possible. The significance of EHEC as a public health problem was first recognized in 1982, following an outbreak in the United States of America associated with undercooked hamburgers (Kaper et al. 2004).

Communicated by Erko Stackebrandt.

Elzbieta Brzuszkiewicz, Andrea Thürmer, Jörg Schuldes, Andreas Leimbach, Heiko Liesegang have contributed equally to this article.

Electronic supplementary material The online version of this article (doi:10.1007/s00203-011-0725-6) contains supplementary material, which is available to authorized users.

E. Brzuszkiewicz · A. Thürmer · J. Schuldes · A. Leimbach · H. Liesegang · F.-D. Meyer · G. Gottschalk · R. Daniel
Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August University Göttingen, Grisebachstr. 8, 37077 Göttingen, Germany

R. Daniel (✉)
Department of Genomic and Applied Microbiology, Institute of Microbiology and Genetics, Georg-August University Göttingen, Grisebachstr. 8, 37077 Göttingen, Germany
e-mail: rdaniel@gwdg.de

J. Boelter
Roche Diagnostics Deutschland GmbH, Sandhofer Str. 116, 68305 Mannheim, Germany

H. Petersen
Medizinisches Versorgungszentrum für Labormedizin und Humangenetik, Abt. Molekulare Erregerdiagnostik, Bergstraße 14, 20095 Hamburg, Germany

Published online: 29 June 2011

Infections caused by EHEC may lead to severe diarrhea and hemorrhagic colitis with complications such as microangiopathic hemolytic anemia, thrombocytopenia, and fatal acute renal failure, which are summarized as hemolytic uremic syndrome (HUS) (Karmali et al. 1983, 1985; Law et al. 1992). Ruminants, predominantly cows, are the natural reservoir of EHEC strains (Kaper et al. 2004).

EHEC is known to produce characteristic toxins, which are similar to toxins produced by *Shigella dysenteriae* and are known as verocytotoxins or Shiga toxins (STX) (Kaper et al. 2004; Karch et al. 2005; Tarr et al. 2005). Absorption of these toxins by the bloodstream leads to damage to the kidneys and to HUS. The most significant serogroups among EHEC strains are O26, O103, O111, and O157. *E. coli* O157:H7 is the most important EHEC serotype with respect to public health in North America, the United Kingdom, and Japan (Kaper et al. 2004). Typical EHEC strains produce STX but also encode a LEE (locus of enterocyte effacement) pathogenicity island, which is important for adherence in the colon (Jores et al. 2004). *E. coli* strains that encode a Shiga toxin, but do not contain the LEE pathogenicity island, are designated as STEC (Shiga toxin-producing *E. coli*) strains. Approximately 200 different serogroups of STEC strains are known and more than 100 harbor a virulence potential. Up to 50% of infections with STEC strains are linked to non-O157 serogroups (Kaper et al. 2004).

The EHEC outbreak started in Germany in May 2011 with 3,368 cases including 36 deaths (as of June 14th, 2011, European Centre for Disease Prevention and Control; http://www.ecdc.europa.eu/en/Pages/home.aspx). This is the second largest food-borne *E. coli* outbreak in history. The enterohemorrhagic *E. coli* strain O104:H4 was identified as the causative agent of the EHEC infection outbreak. This strain was found in humans before but never as causative agent of an EHEC outbreak (Robert Koch Institute, Berlin, Germany; http://www.rki.de). Only one case of infection with strain O104:H4 has been documented in the literature prior to the 2011 outbreak. In this case, the strain was isolated from a 29-year-old Korean woman, who suffered from HUS (Bae et al. 2006).

In this study, we report on the genome sequences of two O104:H4 isolates, which were derived from two patients of the 2011 EHEC outbreak in Germany. The determination of the genomic features of the isolates provides insights into the genomic potential, pathogenicity, and evolution of the O104:H4 strain. Comparison of our *E. coli* O104:H4 genome sequences with that of other pathogenic *E. coli* suggests that strain O104:H4 represents a new *E. coli* pathotype, which we named Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC).

Results

General features of *E. coli* GOS1 and GOS2 genome sequences

The genome sequences of two *E. coli* O104:H4 strains derived from two different patients, a 75-year-old woman and 48-year-old man, from the 2011 German EHEC outbreak were determined using 454 pyrosequencing technology (Margulies et al. 2005). The two analyzed strains were designated *E. coli* GOS1 and GOS2 (German outbreak strain). PCR-based detection of four specific marker genes (*stx2, terD, rfb0104,* and *fliC* H4) confirmed that both were O104:H4 strains (Fig. S1). The general genomic features of the genomes of *E. coli* GOS1 and GOS2 are presented together with features of already sequenced and selected *E. coli* reference genomes in Table S1. The assembly of the draft genomes of *E. coli* GOS1 and GOS2 yielded 171 and 204 large contigs, respectively (Table 1). The estimated genome size of both isolates is 5.31 Mbp. In addition, a total of 5,217 (GOS1) and 5,224 (GOS2) protein-encoding genes were predicted.

Genome comparison of GOS1 and GOS2 with selected *E. coli* genomes

Sequence alignment of *E. coli* GOS1 and GOS2 genome sequences using the MUMmer software tool (Kurtz et al. 2003) revealed 99.9% identity of both sequences. We could not find a single-nucleotide polymorphism when we compared the draft genomes of *E. coli* GOS1 and GOS2 by employing the GS Mapper Reference software (Roche 454, Branford, USA). Thus, as these isolates derived from patients showing different gender and age, it appears that the genome of *E. coli* O104:H4 is stable during its infection in different hosts. This assumption was supported by comparison of the *E. coli* GOS1 and GOS2 genomes with the three

Table 1 Assembly data of the *Escherichia coli* GOS1 and GOS2 genome sequences

	E. coli GOS1	*E. coli* GOS2
Genome size (Mbp)	5.31	5.31
GC content (%)	50.6	50.6
Coverage	24-fold	21-fold
Number of large contigs (>500 bp)	171	204
Average contig size (kbp)	30.99	25.96
N50 contig size (kbp)	109.54	88
Largest contig size (kbp)	337.55	247.7
Q40 value (%)	99.41	99.42

The genomes of *E. coli* GOS1 and *E. coli* GOS2 were assembled de novo from 349.788 and 311.478 shotgun reads, respectively, by employing the Roche Newbler assembly software other available draft genome sequences of *E. coli* O104:H4 isolates derived from the German outbreak. The sequence identities of *E. coli* GOS1 to the genome sequences of *E. coli* O104:H4 isolates TY-2482 (Beijing Genomics Institute, China), LB226692 (Life Technologies, Germany; University of Münster, Germany), and H112180280 (Health Protection Agency, Cambridge, United Kingdom) were 99.8, 99.5, and 99.9%, respectively. Taking into account the overall high similarity of all five genome sequences and the different sequencing approaches used, we assume that the recorded differences of the genome sequences are mainly due to sequencing errors and not to changes within the genome of the different isolates. In addition, as all analyzed chromosomal *E. coli* sequences share synteny over the whole chromosome length, we could align chromosomal contigs of all available sequences of the German outbreak to the chromosome of EAEC 55989 and obtain the contig order for the genomes of *E. coli* GOS1 and GOS2 (Fig. S2).

Comparison of the complete gene content of *E. coli* GOS1 and GOS2 with selected *E. coli* genomes showed that the chromosome of both isolates is most similar to that of the entero-aggregative *E. coli* (EAEC) strain 55989 (Fig. S2). *E. coli* strain 55989 was originally isolated from the diarrheagenic stools of an HIV-positive adult suffering from persistent watery diarrhea (Mossoro et al. 2002). Genome wide BiBag comparisons revealed a set of 4,606 (GOS1) and 4,607 (GOS2) orthologous genes that are shared by at least one chromosome of the selected reference *E. coli* strains (Table S1). Among the remaining 611 (GOS1) and 617 (GOS2) genes 122 and 211, respectively, genes were orthologous to genes located on plasmids.

Comparisons of the *E. coli* GOS1 and GOS2 chromosomes with those of EAEC 55989 and EHEC O157:H7 Sakai using the Artemis comparison tool (Carver et al. 2005) revealed that the chromosomal backbone of the German outbreak strain is different from that of typical *E. coli* EHEC or EAEC strain. Most important differences are the lack of the LEE pathogenicity island and the presence of a Stx-phage in the genomes of *E. coli* GOS1 and GOS2 (Fig. 1).

A multilocus sequence typing (MLST) analysis of seven housekeeping genes *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA* of the two *E. coli* isolates GOS1 and GOS2 was done according to Wirth et al. (2006). *E. coli* GOS1 and GOS2 share the same sequence for all seven genes. By interrogation of the Achtman's MLST scheme database (Wirth et al. 2006), the outbreak strain could be assigned to the sequence type 678 (ST678) complex (*adk* 6, *fumC* 6, *gyrB* 5, *icd* 136, *mdh* 9, *purA* 7, *recA* 7). This complex belongs to the ECOR ancestral group B1, which is a very heterogeneous group with respect to included pathotypes (Tenaillon et al. 2010). The group B1 includes non-O157, EHEC, ETEC, and commensal *E. coli* strains. In addition, EAEC strain 55989 is grouped in B1. A Maximum Likelihood tree of completely sequenced *E. coli* genomes confirmed the close relationship of the German outbreak strain to EAEC 55989 (Fig. 2).

Plasmids

We identified two genes encoding plasmid replication proteins in each dataset (GOS1, RGOS01291, and RGOS00376; GOS2, RGOT04762, and RGOT01786). Therefore, it is assumed that the outbreak strain harbors at least two extrachromosomal replicons. In order to identify the potential plasmid-encoded proteins, our sequence data were mapped on several reference plasmids (Table S2). A total of 169 potential plasmid-located genes were thereby identified. Further data analysis revealed the presence of a putative plasmid in *E. coli* GOS1 and GOS2, which is almost identical to the pEC_Bactec plasmid (Fig. 3). Contigs from our data spanned over 90% of the total pEC_Bactec plasmid length (84,221 bp out of 92,970 bp). Small contigs coding only for transposases or insertion elements were not included in the analysis. The reconstructed plasmids of *E. coli* GOS1 and GOS2 consist of only three contigs (Fig. 3). The resistance genes TEM-1 and CTX-M-15 are located on this plasmid. Extended-spectrum beta-lactamases (ESBLs) such as TEM-1 and

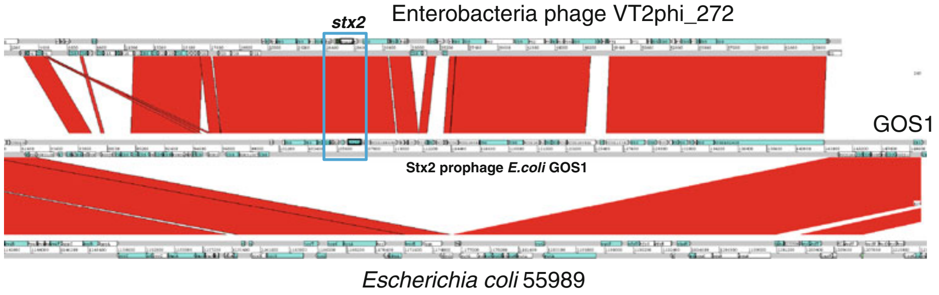
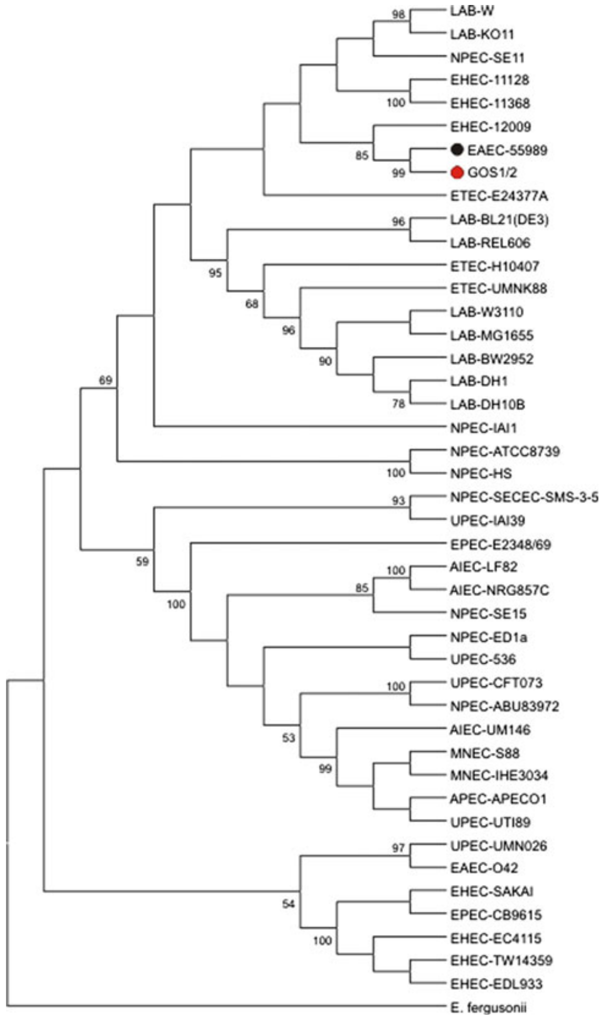


Fig. 1 Comparisons of enterobacteria phage VT2phi_272 with the corresponding genomic region of *E. coli* GOS1 and *E. coli* strain 55989. Analysis was performed by employing the ACT software tool (Sanger Institute, http://www.sanger.ac.uk). The relationship between each pair of sequences are depicted. Similar coding sequences are indicated by *red-colored lines*. The *stx* genes are *boxed*



◀**Fig. 2** Phylogenetic analysis of completely sequenced *E. coli* strains based on multilocus sequence typing. The phylogenetic analysis was conducted with MEGA 5.05 (Tamura et al. 2011). The resulting Maximum Likelihood tree illustrates the close relationship of the German outbreak strain (*red dot*) to EAEC 55989 (*black dot*). The pathotype of each *E. coli* strain is indicated in front of the strain name (see below for abbreviations). Bootstrap values were calculated from 100 resamplings. Bootstrap values below 50 were not shown. The following *E. coli* strains were used in the analysis: entero-aggregative *E. coli* (EAEC) 042 (FN554766), uropathogenic *E. coli* (UPEC) 536 (CP000247), EAEC 55989 (CU928145), commensal non-pathogenic *E. coli* (NPEC) ABU83972 (CP001671), avian pathogenic *E. coli* (APEC) O1 (CP000468), lab B strain BL21(DE3) (AM946981), lab B strain REL606 (CP000819), industrial production strain KO11 (CP002516), enteropathogenic *E. coli* (EPEC) CB9615 (CP001846), UPEC CFT073 (AE014075), EPEC E2348/69 (FM180568), enterotoxigenic *E. coli* (ETEC) E24377A (CP000800), commensal ED1a (CU928162), ETEC H10407 (FN649414), commensal HS (CP000802), commensal IAI1 (CU928160), UPEC IAI39 (CU928164), meningitis-associated *E. coli* (MNEC) IHE3034 (CP001969), commensal strain K-12 substrain ATCC 8739/Crooks (CP000946), lab strain K-12 substrain BW2952 (CP001396), lab strain K-12 substrain DH1 (CP001637), lab strain K-12 substrain DH10B (CP000948), lab strain K-12 substrain MG1655 (U00096), lab strain K-12 substrain W3110 (AP009048), adherent-invasive *E. coli* (AIEC) LF82 (CU651637), AIEC NRG 857C (CP001855), EHEC O103:H2 12009 (AP010958), EHEC O111:H- 11128 (AP010960), EHEC O157:H7 EC4115 (CP001164), EHEC O157:H7 EDL933 (AE005174), EHEC O157:H7 Sakai (BA000007), EHEC O157:H7 TW14359 (CP001368), EHEC O26:H11 11368 (AP010953), MNEC S88 (CU928161), commensal SE11 (AP009240), commensal SE15 (AP009378), environmental strain SECEC SMS-3-5 (CP000970), AIEC UM146 (CP002167), UPEC UMN026 (CU928163), porcine ETEC UMNK88 (CP002729), UPEC UTI89 (CP000243), and lab strain W (CP002185). *Escherichia fergusonii* ATCC 35469 was used as outgroup (CU928158)

CTX-M-15 are the most prevalent secondary beta-lactamases among clinical isolates of *Enterobacteriaceae* worldwide (Livermore 1995). ESBLs are a group of β-lactamases, which share the ability to hydrolyze third-generation cephalosporins and aztreonam (Paterson and Bonomo 2005).

A significant number of genes mapped to the plasmids p042 and 55989p, which are typical for EAEC strains (Fig. 4a; Table S2) (Touchon et al. 2009; Chaudhuri et al. 2010). The plasmids of GOS1 and GOS2 share a set of 46 genes with EAEC plasmid 55989p (Table S2) including the aggregative adhesion operon *aat* and the regulator *aggR*. Additionally, the toxin–antitoxin system *ccd* and the replication protein RepFIB were found. However, genes encoding for aggregative adherent fimbriae (AAF), a primary virulence factor of EAEC strains (Kaper et al. 2004), are different from the 55989p variant. Mapping *E. coli* GOS1 and GOS2 data on the second reference plasmid p042 showed also a significant number of homologous proteins (Fig. 4b; Table S2). Many potential virulence factors are shared with p042 plasmid such as the AAF (*agg3*) operon and the serine protease *pet.* Pet is secreted by many EAEC strains and exhibits enterotoxic activity (Navarro-García et al. 1998).

Phage analysis

We could identify 336 prophage-encoding genes for GOS1 and 334 for GOS2 (Tables S3, S4). The key virulence factor of EHEC, STX, is encoded on a lambda-like bacteriophage, the Stx-phage. Acquisition of this phage was a key step in the evolution of EHEC from EPEC (Reid et al. 2000). A Stx-phage is present in the outbreak strain (Fig. 1). This phage shows high identity to the *stx2*-containing enterobacteria phage VT2phi_272 from *E. coli* O157:H7 strain 71074 (HQ424691). The GOS1 Stx-prophage consists of 66 encoding genes and is identical to the GOS2 Stx-phage (Tables S3, S4). In addition to the Stx-phage, 70 prophage-encoding genes (Tables S3, S4) that are not present in *E. coli* 55989 could be identified in the genome of *E. coli* GOS1. These genes have high similarity to STX-producing prophages and also to the other

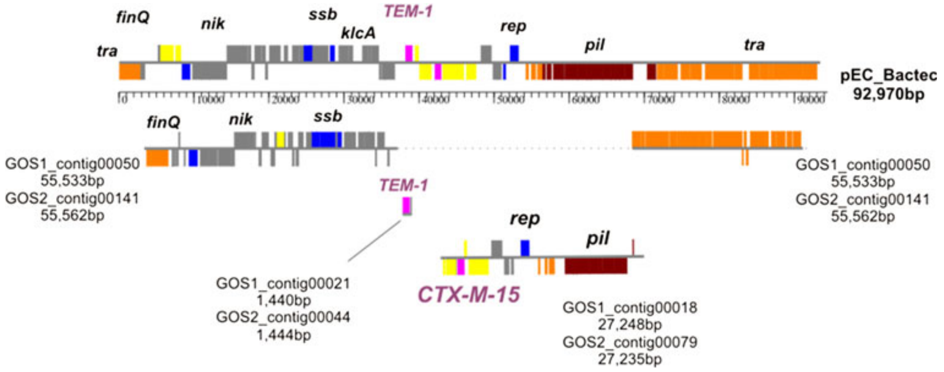


Fig. 3 Linear comparison of *E. coli* pEC_Bactec plasmid with corresponding GOS1 and GOS2 contigs. The *top map* represents the pEC_Bactec plasmid (GU371927.1), the resistance genes are highlighted in *pink*, IS-elements/transposases in *yellow*, plasmid replication/stabilization genes in *blue*, the *tra* operon in *orange*, *pil* operon in *brown*, and remaining genes in *gray*. The scale is in base pairs. All maps were done with GenVision software (http://www.dnastar.com/t-products-genvision.aspx)

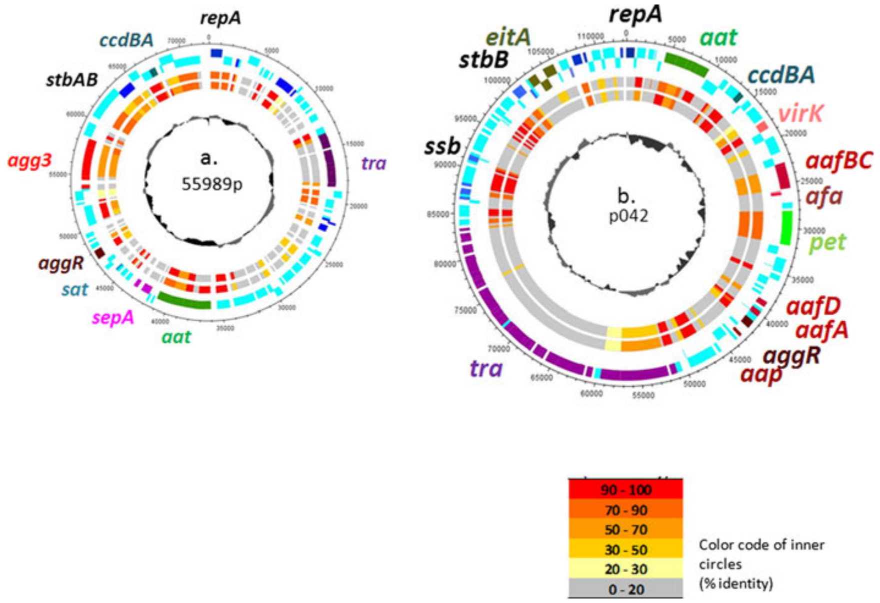


Fig. 4 Comparison of GOS1 and GOS2 genes with two different pAA-type plasmids. The two outermost rings represent maps of **a** 55989p and **b** p042 from strain *E. coli* 55989 and *E. coli* 042, respectively. Virulence factors and selected important genes are *highlighted* and *colored*. The *second* and the *third rings* represent presence (*colored*) or absence (*gray*) of GOS1 and GOS2 orthologs. The *inner rings* represent the GC contents of the plasmids

above-mentioned phage in the outbreak strain, but lack *stx2AB* (Fig. S3).

Resistance

EHEC O157:H7 strains resist the highly toxic tellurium oxyanion, tellurite (Tel) (Zadik et al. 1993; Taylor et al. 2002; Bielaszewska et al. 2005; Orth et al. 2007). Tellurite resistance (TelR) of EHEC O157:H7 is encoded by the chromosomal *terZABCDEF* gene cluster (Taylor et al. 2002; Bielaszewska et al. 2005), which is highly homologous to the *ter* cluster on plasmid R478 of *Serratia marcescens* (Whelan et al. 1995; Taylor et al. 2002). TelR is a common, but not obligatory, feature of EHEC O157:H7 strains, as tellurite-susceptible *E. coli* O157:H7 strains have been isolated in North America (Taylor et al. 2002) and Europe (Bielaszewska et al. 2005). We identified all proteins of the *terZABCDEF* operon in the outbreak strain (ORFs RGOS02836 to RGOS02842).

In addition, the German outbreak strain could bear a mercuric resistance plasmid, as in many bacteria resistance to mercury is associated with a plasmid (Smith 1967; Novick and Roth 1968; Summers and Silver 1972; Kondo et al. 1974). Correspondingly, the predicted proteins involved in mercury resistance were located all on one contig (GOS1_contig00023). These genes encode the putative mercuric ion transport proteins MerT, MerP, and MerC (RGOS00392, RGOS00393, and RGOS00394, respectively), the corresponding transcriptional regulators MerR (RGOS00391) and MerD (RGOS00396), and mercuric ion reductase MerA (RGOS00395). In addition to genes involved in mercuric resistance and tellurium resistance, we have predicted and annotated many genes involved in antibiotic resistance such as putative gene-encoding chloramphenicol (RGO00056), tetracycline (RGOS00387, RGOS00388), or streptomycin resistance (RGOS00359).

Discussion

Chromosomes and plasmids

The chromosomes of the *E. coli* isolates GOS1 and GOS2 are most similar to the chromosome of EAEC strain 55989 isolated in Africa over a decade ago. EAEC strains are the most recently emerged *E. coli* intestinal pathotype and the second most common agent of traveler's diarrhea (Huang et al. 2006). EAEC pathogenesis is thought to involve three primary steps. First, the bacteria adhere to the intestinal mucosa using aggregative adherent fimbriae (AAF). Second, these fimbriae cause autoaggregative adhesion, by which the bacteria adhere to each other in a 'stacked-brick' configuration producing a mucous-mediated biofilm on the enterocyte surface. Third, the bacteria release toxins that affect the inflammatory response, intestinal secretion, and mucosal cytotoxicity. Aspects of each of these steps involve plasmid-encoded traits but also chromosomal-encoded virulence factors (Kaper et al. 2004).

In addition to the chromosomal similarity, *E. coli* GOS1 and GOS2 share with EAEC strain 55989 part of the EAEC plasmid 55989p. This plasmid carries the AAF operon *aat* and the regulator *aggR*. Nevertheless, a different aggregative adhesion fimbrial complement was present in our strains. The AAF operon is usually localized on an approximately 100-kb plasmid, termed the "pAA plasmid" (Nataro et al. 1987). Four genetically distinct allelic variants of AAF have been identified previously, AAF/I from EAEC strain 17-2 (Nataro et al. 1992), AAF/II from strain O42 (Nataro et al. 1995), AAF/III from strain 55989 (Bernier et al. 2002), and Hda from strain C1010-00 (Boisen et al. 2008). All the identified AAF allelic types appear to be plasmid encoded, and most of the analyzed strains possess only a single AAF allelic type (Harrington et al. 2006). The outbreak strain is no exception and seems to contain the relatively rare AAF/I locus of EAEC. Additionally, the *ipd* gene encoding an extracellular serine protease and the gene encoding serine protease Pet were found in the German outbreak strain. Usually, these virulence factors are localized next to the AAF operon on the pAA plasmid. Another virulence feature, the *aatPABCD* operon (dispersin secretion locus), is a plasmid-borne characteristic of EAEC strains. This operon is also present in the genome of the German outbreak strain.

Two RepA proteins were found in the German outbreak strain. This suggests that this strain harbors at least two plasmids. In addition to the pAA-like plasmid, we identified contigs showing high similarity to the previously described plasmids pEC_Bactec, pCVM29188_101, and pEK204 (Fricke et al. 2009; Woodford et al. 2009; Smet et al. 2010). These plasmids encode the extended-spectrum β-lactamases blaCTX-M and blaTEM-1.

Evolution: horizontal gene transfer (HGT)

Escherichia coli virulence factors such as enterotoxins, invasion factors, adhesion factors, or Shiga toxins can be encoded by several mobile genetic elements, including transposons (Tn), plasmids, bacteriophages, or pathogenicity islands (e.g., LEE island). Bacterial plasmids play a key role in a variety of traits like drug resistance, virulence, and the metabolism of rare substrates under specific conditions (Actis et al. 1999). Plasmids are able to mobilize these traits between different strains and thus play an important role in horizontal gene transfer. The analyses indicate that a number of horizontal gene transfer events took place to create the genome of the German outbreak strain. This strain probably originated from an EAEC pathotype, which is suggested by the missing LEE island and the high similarity of the genome to the genome of EAEC strain 55989. In contrast to the EAEC strains, the German outbreak strain has acquired the Stx-phage, which is typical for EHEC strains (Fig. 1).

Another feature of the new outbreak strain is the acquisition of plasmid-encoded drug resistances. The strain has acquired a plasmid sharing high similarity with the plasmids pEC_Bactec, pCVM29188_10, and pEK204. The origin of this plasmid remains unclear, since the extended-spectrum β-lactamases (ESBLs) CTX-M and TEM-1 resistances seem to be located on a Tn3-type transposon that has been widely spread among enteric bacteria.

To conclude, *E. coli* O104:H4 possesses a Stx-phage typical for EHEC strains but is missing the characteristic LEE island. In addition to the high overall genome sequence similarity to EAEC strains, it harbors an AAF operon, which is a distinguishing feature for EAEC strains. The German outbreak strain harbors a unique combination of EHEC and EAEC genomic features (Fig. 5). These data suggest a new *E. coli* pathotype EAHEC that has EHEC and EAEC ancestors.

Materials and methods

Sample preparation and DNA extraction

The two *E. coli* O104:H4 isolates GOS1 and GOS2 were derived from stool samples of two different patients of the 2011 German outbreak. *E. coli* GOS1 and GOS2 were

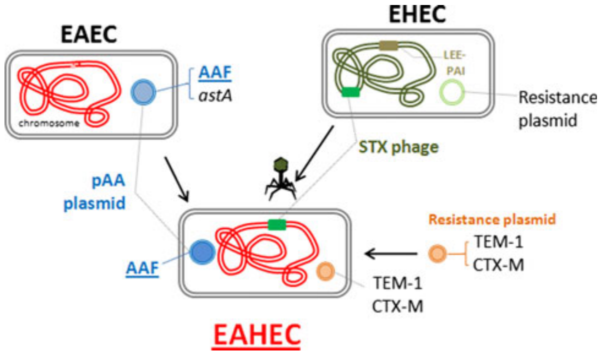


Fig. 5 Proposed scheme of the origin of the new *E. coli* pathotype—EAHEC

recovered from a 75-year-old woman and a 48-year-old man, respectively. To isolate these strains, stool samples were plated on Brilliance™ ESBL Agar plates (Oxoid, Wesel, Germany) and incubated for 24 h at 37°C. Initially, the *E. coli* O104:H4 strains were identified by the ability to produce STX2. For this purpose, the LightMix® kits *E. coli* EHEC Stx1 and Stx2 were applied as recommended by the manufacturer (TIB MOLBIOL, Berlin, Germany). A colony of each strain from the thereby recovered positive strains, *E. coli* GOS1 and GOS1, was grown in 4 ml EHEC-direct-media (Heipha Diagnostics, Eppelheim, Germany) overnight at 37°C. To isolate genomic DNA, the cultures were pelleted (5 min, 2,000*g*), resuspended in 1 ml S.T.A.R. Buffer (Roche, Molecular Diagnostics, Rotkreuz, Switzerland), and incubated for 5 min at 95°C. Subsequently, the suspension was subjected to centrifugation for 1 min at 1,100*g*. The cell-free supernatant (500 µl) was used for the preparation of the genomic DNA by employing the High Pure 16 System Viral Nucleic Acid kit as recommended by the manufacturer (Roche Applied Science, Mannheim, Germany). The resulting DNA solution (260 ng/µl) was used for further analysis.

To confirm that *E. coli* isolates GOS1 and GOS2 were O104:H4 serotype, a PCR-based detection of four specific marker genes (*stx2*, *terD*, *rfbO104*, and *fliC* H4) was performed according to the PCR typing scheme by the group of Prof. Karch at the National Consulting Laboratory on HUS at the University of Münster (see http://www.ehec.org/pdf/Laborinfo_01062011.pdf, 2011) with slight adaptations. Briefly, the PCR reaction mixture (25 µl) contained 2.5 µl tenfold reaction buffer (Bioline, Luckenwalde, Germany), 0.2 mM of each of the four deoxynucleoside triphosphates, 1.5 mM $MgCl_2$, 0.2 µM of each of the primers, 1 U of BIO-X-ACT™ DNA Polymerase (Bioline), and 100 ng of isolated genomic DNA as template. The *stx2, terD, rfbO104*, and *fliC* H4 were amplified with the following set of primers: *stx2*, 5′-ATCCTATTCCCGGGAGTTTACG-3′ and 5′-GCGTCATCGTATACACAGGAGC-3′; *terD*, 5′-AGTAAAGCAGCTCCGTCAAT-3′ and 5′-CCGAACAGCATGGCAGTCT-3′; *rfbO104*, 5′-TGAACTGATTTTTAGGATGG-3′ and 5′-AGAACCTCACTCAAATTATG-3′; and *fliC* H4, 5′-GGCGAAACTGACGGCTGCTG-3′ and 5′-GCACCAACAGTTACCGCCGC-3′. The following thermal cycling scheme was used: initial denaturation at 94°C for 5 min, 30 cycles of denaturation at 94°C for 45 s, annealing at 55°C (*stx2, terD, rfbO104*) or 63°C (*fliC* H4) for 45 s, and extension at 72°C for 60 s (*stx2, terD, rfbO104*) or 30 s (*fliC* H4) followed by a final extension period at 72°C for 5 min. Subsequently, PCR products were separated by agarose gel electrophoresis (1.5% gels) and analyzed. The analysis revealed that all four marker genes were present in *E. coli* isolates GOS1 and GOS2 in the expected sizes (Fig. S1).

Sequencing and assembly

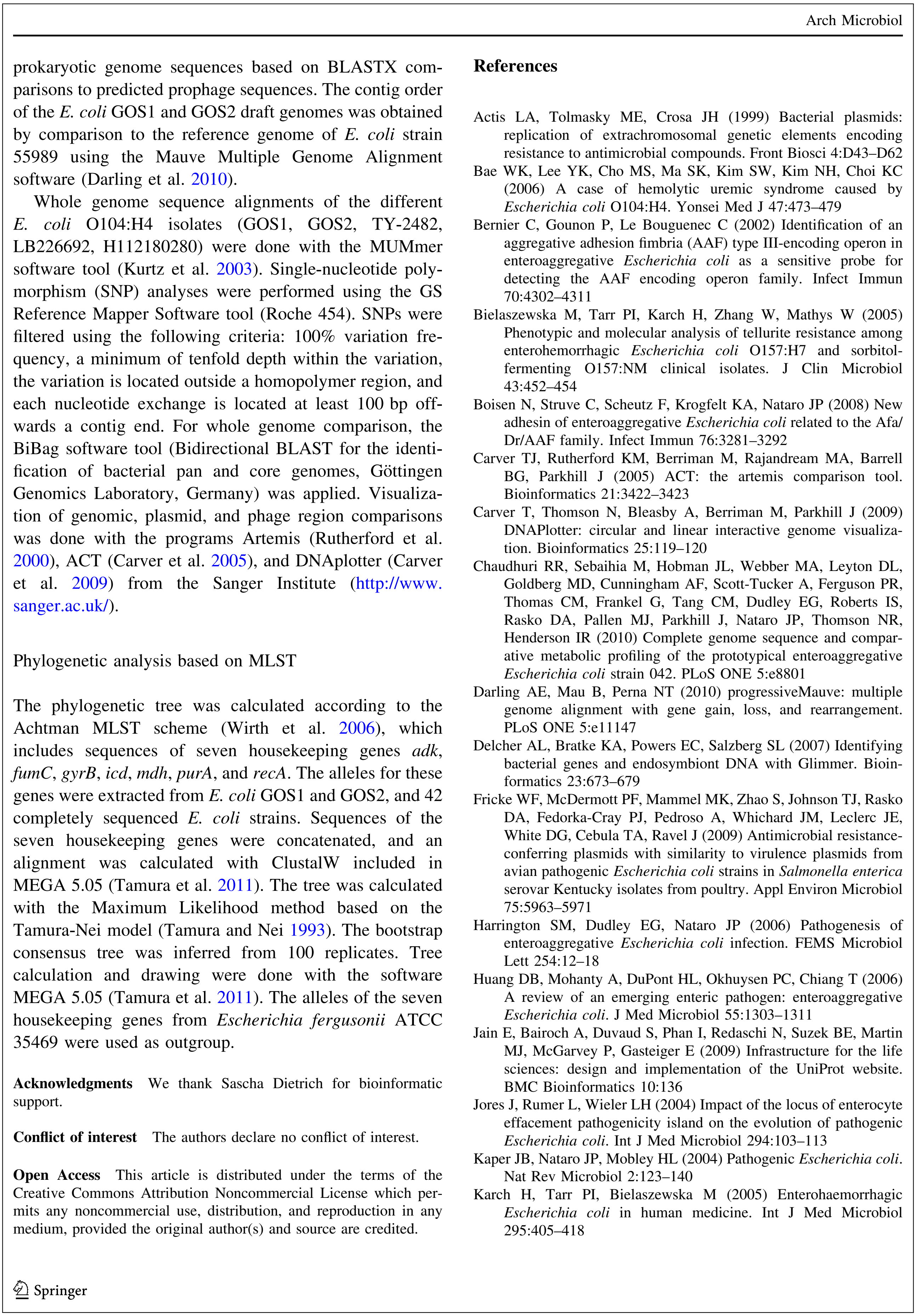
The isolated DNA from both strains was used to create 454-shotgun libraries following the GS Rapid library protocol (Roche 454, Branford, USA). The resulting two 454 DNA libraries were sequenced with the Genome Sequencer FLX (Roche 454) using Titanium chemistry. For sequencing of each sample, 1.5 medium lanes of a Titanium picotiter plate were used. A total of 349,788 and 311,478 shotgun reads were achieved for *E. coli* GOS1 and *E. coli* GOS2, respectively. Reads were assembled de novo using the Roche Newbler assembly software 2.3 (Roche 454) (Table 1).

Gene prediction and annotation

Gene prediction was performed with Glimmer3 (Delcher et al. 2007). Automatic gene annotation was done by transferring annotations from orthologous genes of reference strains (Table S1) available at the EMBL database. Orthologous genes were identified as described previously by bidirectional BLAST comparisons (Schmeisser et al. 2009). Proteins without orthologs in the reference strains were annotated according to their best BLAST hits to the SwissProt subset of the UniProt Database (Jain et al. 2009, http://www.uniprot.org). Sequence data of isolates GOS1 and GOS2 are publicly available and can be downloaded from the Göttingen Genomics Laboratory website (ftp://134.76.70.117; UserID: EAHEC_GOS; Password: EAHEC_GOS).

Genome analysis

In order to analyze the presence of prophage regions, the Prophage Finder software has been employed (http://131.210.201.64/~phage/ProphageFinder.php). This web application provides a quick prediction of prophage loci in



prokaryotic genome sequences based on BLASTX comparisons to predicted prophage sequences. The contig order of the *E. coli* GOS1 and GOS2 draft genomes was obtained by comparison to the reference genome of *E. coli* strain 55989 using the Mauve Multiple Genome Alignment software (Darling et al. 2010).

Whole genome sequence alignments of the different *E. coli* O104:H4 isolates (GOS1, GOS2, TY-2482, LB226692, H112180280) were done with the MUMmer software tool (Kurtz et al. 2003). Single-nucleotide polymorphism (SNP) analyses were performed using the GS Reference Mapper Software tool (Roche 454). SNPs were filtered using the following criteria: 100% variation frequency, a minimum of tenfold depth within the variation, the variation is located outside a homopolymer region, and each nucleotide exchange is located at least 100 bp offwards a contig end. For whole genome comparison, the BiBag software tool (Bidirectional BLAST for the identification of bacterial pan and core genomes, Göttingen Genomics Laboratory, Germany) was applied. Visualization of genomic, plasmid, and phage region comparisons was done with the programs Artemis (Rutherford et al. 2000), ACT (Carver et al. 2005), and DNAplotter (Carver et al. 2009) from the Sanger Institute (http://www.sanger.ac.uk/).

Phylogenetic analysis based on MLST

The phylogenetic tree was calculated according to the Achtman MLST scheme (Wirth et al. 2006), which includes sequences of seven housekeeping genes *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*. The alleles for these genes were extracted from *E. coli* GOS1 and GOS2, and 42 completely sequenced *E. coli* strains. Sequences of the seven housekeeping genes were concatenated, and an alignment was calculated with ClustalW included in MEGA 5.05 (Tamura et al. 2011). The tree was calculated with the Maximum Likelihood method based on the Tamura-Nei model (Tamura and Nei 1993). The bootstrap consensus tree was inferred from 100 replicates. Tree calculation and drawing were done with the software MEGA 5.05 (Tamura et al. 2011). The alleles of the seven housekeeping genes from *Escherichia fergusonii* ATCC 35469 were used as outgroup.

Acknowledgments We thank Sascha Dietrich for bioinformatic support.

Conflict of interest The authors declare no conflict of interest.

References

Actis LA, Tolmasky ME, Crosa JH (1999) Bacterial plasmids: replication of extrachromosomal genetic elements encoding resistance to antimicrobial compounds. Front Biosci 4:D43–D62

Bae WK, Lee YK, Cho MS, Ma SK, Kim SW, Kim NH, Choi KC (2006) A case of hemolytic uremic syndrome caused by *Escherichia coli* O104:H4. Yonsei Med J 47:473–479

Bernier C, Gounon P, Le Bouguenec C (2002) Identification of an aggregative adhesion fimbria (AAF) type III-encoding operon in enteroaggregative *Escherichia coli* as a sensitive probe for detecting the AAF encoding operon family. Infect Immun 70:4302–4311

Bielaszewska M, Tarr PI, Karch H, Zhang W, Mathys W (2005) Phenotypic and molecular analysis of tellurite resistance among enterohemorrhagic *Escherichia coli* O157:H7 and sorbitol-fermenting O157:NM clinical isolates. J Clin Microbiol 43:452–454

Boisen N, Struve C, Scheutz F, Krogfelt KA, Nataro JP (2008) New adhesin of enteroaggregative *Escherichia coli* related to the Afa/Dr/AAF family. Infect Immun 76:3281–3292

Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J (2005) ACT: the artemis comparison tool. Bioinformatics 21:3422–3423

Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J (2009) DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25:119–120

Chaudhuri RR, Sebaihia M, Hobman JL, Webber MA, Leyton DL, Goldberg MD, Cunningham AF, Scott-Tucker A, Ferguson PR, Thomas CM, Frankel G, Tang CM, Dudley EG, Roberts IS, Rasko DA, Pallen MJ, Parkhill J, Nataro JP, Thomson NR, Henderson IR (2010) Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042. PLoS ONE 5:e8801

Darling AE, Mau B, Perna NT (2010) progressiveMauve: multiple genome alignment with gene gain, loss, and rearrangement. PLoS ONE 5:e11147

Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679

Fricke WF, McDermott PF, Mammel MK, Zhao S, Johnson TJ, Rasko DA, Fedorka-Cray PJ, Pedroso A, Whichard JM, Leclerc JE, White DG, Cebula TA, Ravel J (2009) Antimicrobial resistance-conferring plasmids with similarity to virulence plasmids from avian pathogenic *Escherichia coli* strains in *Salmonella enterica* serovar Kentucky isolates from poultry. Appl Environ Microbiol 75:5963–5971

Harrington SM, Dudley EG, Nataro JP (2006) Pathogenesis of enteroaggregative *Escherichia coli* infection. FEMS Microbiol Lett 254:12–18

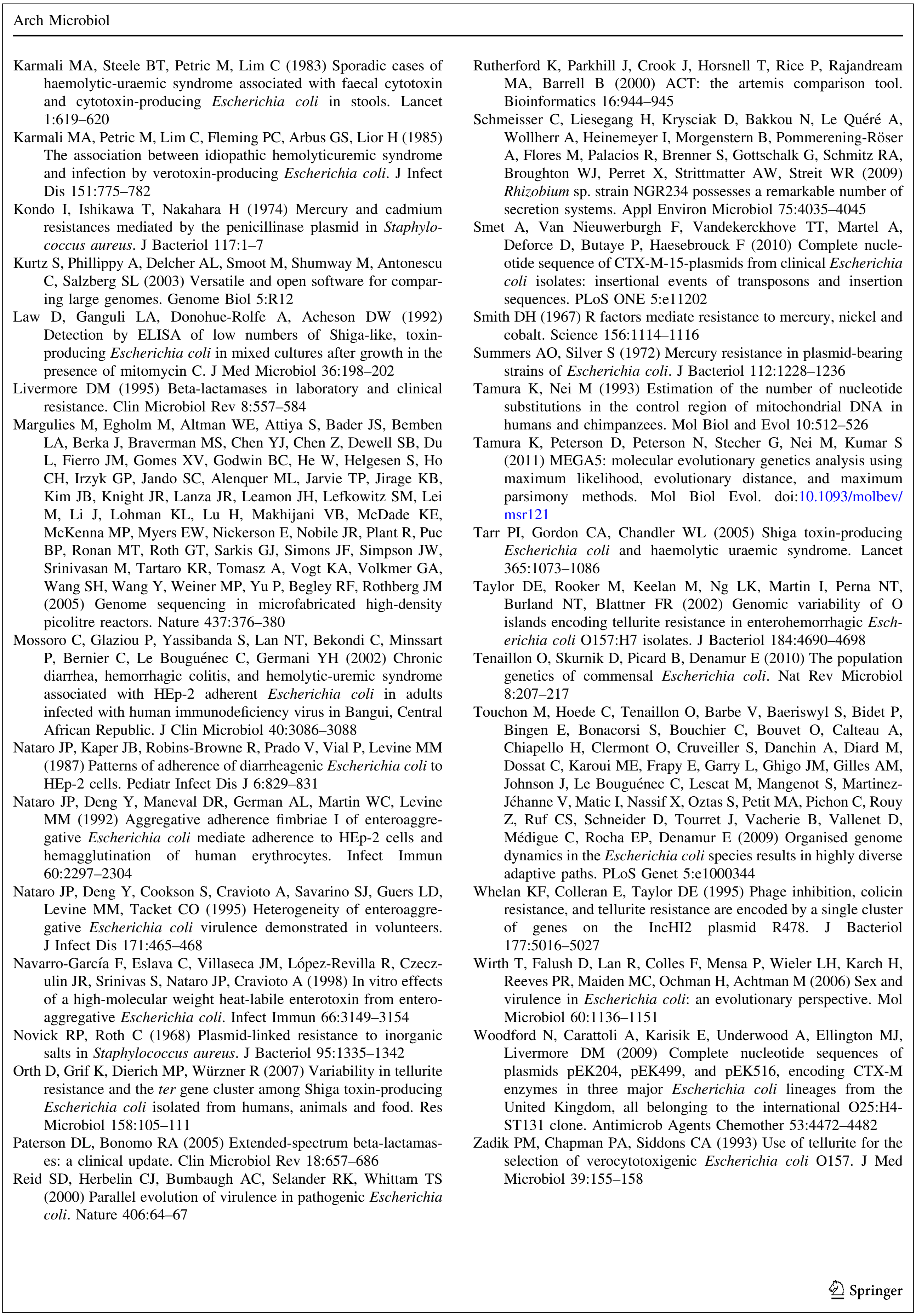
Huang DB, Mohanty A, DuPont HL, Okhuysen PC, Chiang T (2006) A review of an emerging enteric pathogen: enteroaggregative *Escherichia coli*. J Med Microbiol 55:1303–1311

Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, Martin MJ, McGarvey P, Gasteiger E (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics 10:136

Jores J, Rumer L, Wieler LH (2004) Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*. Int J Med Microbiol 294:103–113

Kaper JB, Nataro JP, Mobley HL (2004) Pathogenic *Escherichia coli*. Nat Rev Microbiol 2:123–140

Karch H, Tarr PI, Bielaszewska M (2005) Enterohaemorrhagic *Escherichia coli* in human medicine. Int J Med Microbiol 295:405–418



Karmali MA, Steele BT, Petric M, Lim C (1983) Sporadic cases of haemolytic-uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools. Lancet 1:619–620

Karmali MA, Petric M, Lim C, Fleming PC, Arbus GS, Lior H (1985) The association between idiopathic hemolyticuremic syndrome and infection by verotoxin-producing *Escherichia coli*. J Infect Dis 151:775–782

Kondo I, Ishikawa T, Nakahara H (1974) Mercury and cadmium resistances mediated by the penicillinase plasmid in *Staphylococcus aureus*. J Bacteriol 117:1–7

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2003) Versatile and open software for comparing large genomes. Genome Biol 5:R12

Law D, Ganguli LA, Donohue-Rolfe A, Acheson DW (1992) Detection by ELISA of low numbers of Shiga-like, toxin-producing *Escherichia coli* in mixed cultures after growth in the presence of mitomycin C. J Med Microbiol 36:198–202

Livermore DM (1995) Beta-lactamases in laboratory and clinical resistance. Clin Microbiol Rev 8:557–584

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Mossoro C, Glaziou P, Yassibanda S, Lan NT, Bekondi C, Minssart P, Bernier C, Le Bouguénec C, Germani YH (2002) Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent *Escherichia coli* in adults infected with human immunodeficiency virus in Bangui, Central African Republic. J Clin Microbiol 40:3086–3088

Nataro JP, Kaper JB, Robins-Browne R, Prado V, Vial P, Levine MM (1987) Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells. Pediatr Infect Dis J 6:829–831

Nataro JP, Deng Y, Maneval DR, German AL, Martin WC, Levine MM (1992) Aggregative adherence fimbriae I of enteroaggregative *Escherichia coli* mediate adherence to HEp-2 cells and hemagglutination of human erythrocytes. Infect Immun 60:2297–2304

Nataro JP, Deng Y, Cookson S, Cravioto A, Savarino SJ, Guers LD, Levine MM, Tacket CO (1995) Heterogeneity of enteroaggregative *Escherichia coli* virulence demonstrated in volunteers. J Infect Dis 171:465–468

Navarro-García F, Eslava C, Villaseca JM, López-Revilla R, Czeczulin JR, Srinivas S, Nataro JP, Cravioto A (1998) In vitro effects of a high-molecular weight heat-labile enterotoxin from enteroaggregative *Escherichia coli*. Infect Immun 66:3149–3154

Novick RP, Roth C (1968) Plasmid-linked resistance to inorganic salts in *Staphylococcus aureus*. J Bacteriol 95:1335–1342

Orth D, Grif K, Dierich MP, Würzner R (2007) Variability in tellurite resistance and the *ter* gene cluster among Shiga toxin-producing *Escherichia coli* isolated from humans, animals and food. Res Microbiol 158:105–111

Paterson DL, Bonomo RA (2005) Extended-spectrum beta-lactamases: a clinical update. Clin Microbiol Rev 18:657–686

Reid SD, Herbelin CJ, Bumbaugh AC, Selander RK, Whittam TS (2000) Parallel evolution of virulence in pathogenic *Escherichia coli*. Nature 406:64–67

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) ACT: the artemis comparison tool. Bioinformatics 16:944–945

Schmeisser C, Liesegang H, Krysciak D, Bakkou N, Le Quéré A, Wollherr A, Heinemeyer I, Morgenstern B, Pommerening-Röser A, Flores M, Palacios R, Brenner S, Gottschalk G, Schmitz RA, Broughton WJ, Perret X, Strittmatter AW, Streit WR (2009) *Rhizobium* sp. strain NGR234 possesses a remarkable number of secretion systems. Appl Environ Microbiol 75:4035–4045

Smet A, Van Nieuwerburgh F, Vandekerckhove TT, Martel A, Deforce D, Butaye P, Haesebrouck F (2010) Complete nucleotide sequence of CTX-M-15-plasmids from clinical *Escherichia coli* isolates: insertional events of transposons and insertion sequences. PLoS ONE 5:e11202

Smith DH (1967) R factors mediate resistance to mercury, nickel and cobalt. Science 156:1114–1116

Summers AO, Silver S (1972) Mercury resistance in plasmid-bearing strains of *Escherichia coli*. J Bacteriol 112:1228–1236

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 and Evol 10:512–526

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol. doi:10.1093/molbev/msr121

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

Taylor DE, Rooker M, Keelan M, Ng LK, Martin I, Perna NT, Burland NT, Blattner FR (2002) Genomic variability of O islands encoding tellurite resistance in enterohemorrhagic *Escherichia coli* O157:H7 isolates. J Bacteriol 184:4690–4698

Tenaillon O, Skurnik D, Picard B, Denamur E (2010) The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol 8:207–217

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet 5:e1000344

Whelan KF, Colleran E, Taylor DE (1995) Phage inhibition, colicin resistance, and tellurite resistance are encoded by a single cluster of genes on the IncHI2 plasmid R478. J Bacteriol 177:5016–5027

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60:1136–1151

Woodford N, Carattoli A, Karisik E, Underwood A, Ellington MJ, Livermore DM (2009) Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone. Antimicrob Agents Chemother 53:4472–4482

Zadik PM, Chapman PA, Siddons CA (1993) Use of tellurite for the selection of verocytotoxigenic *Escherichia coli* O157. J Med Microbiol 39:155–158

5.2.1.3 *Supplementary information*

The supplementary figures and one table for Brzuszkiewicz et al. (2011) are presented on pages 115–118. Supplementary Table S2 (predicted proteins of EAHEC orthologous to reference plasmid proteins), Table S3 (putative prophage-encoding genes of *E. coli* GOS1), and Table S4 (putative prophage-encoding genes of *E. coli* GOS2) are too extensive for a reprint, but can be found with the original sources here:
https://link.springer.com/article/10.1007%2Fs00203-011-0725-6

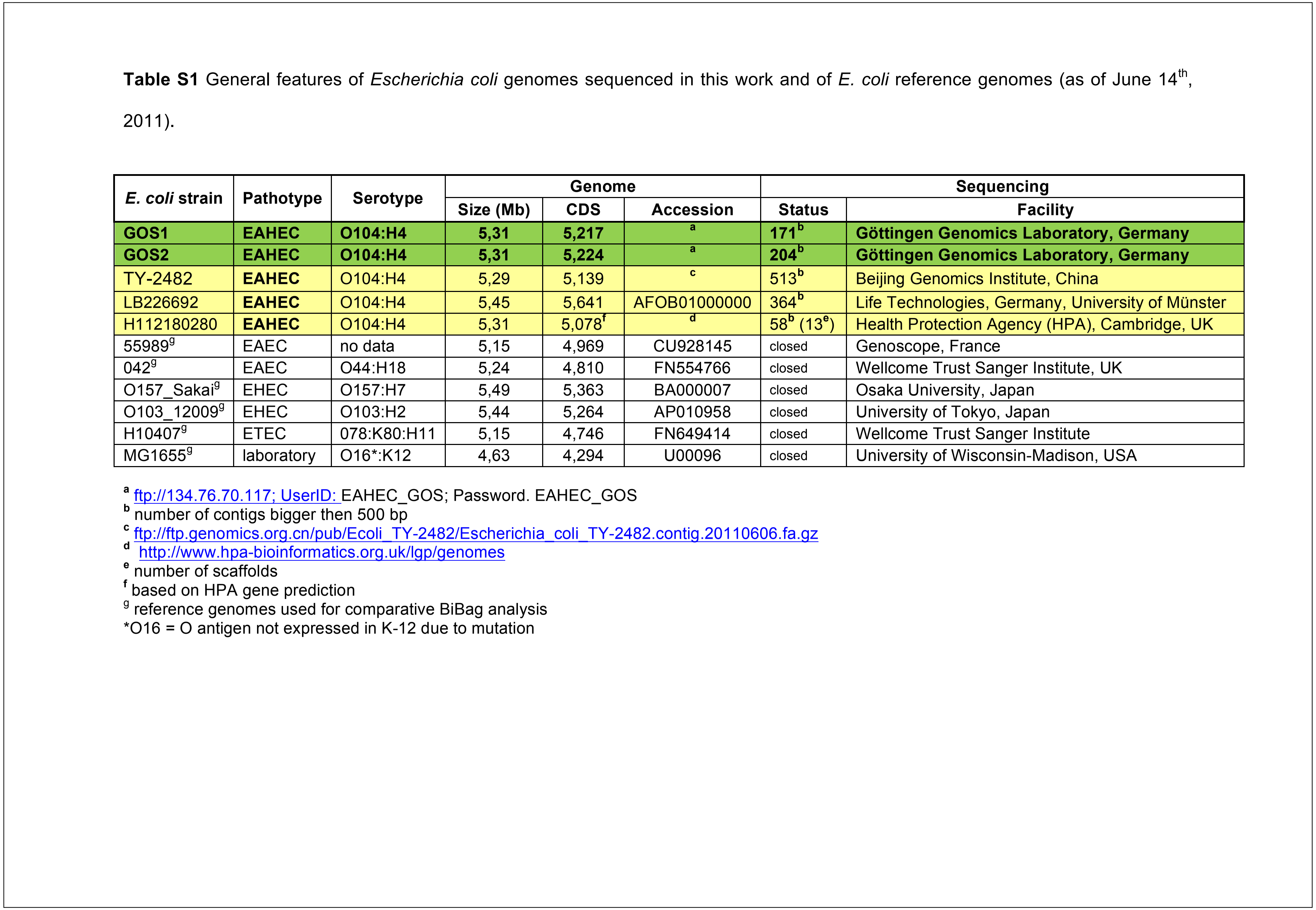


Table S1 General features of *Escherichia coli* genomes sequenced in this work and of *E. coli* reference genomes (as of June 14th, 2011).

E. coli strain	Pathotype	Serotype	Genome			Sequencing	
			Size (Mb)	CDS	Accession	Status	Facility
GOS1	**EAHEC**	**O104:H4**	**5,31**	**5,217**	[a]	**171**[b]	**Göttingen Genomics Laboratory, Germany**
GOS2	**EAHEC**	**O104:H4**	**5,31**	**5,224**	[a]	**204**[b]	**Göttingen Genomics Laboratory, Germany**
TY-2482	**EAHEC**	O104:H4	5,29	5,139	[c]	513[b]	Beijing Genomics Institute, China
LB226692	**EAHEC**	O104:H4	5,45	5,641	AFOB01000000	364[b]	Life Technologies, Germany, University of Münster
H112180280	**EAHEC**	O104:H4	5,31	5,078[f]	[d]	58[b] (13[e])	Health Protection Agency (HPA), Cambridge, UK
55989[g]	EAEC	no data	5,15	4,969	CU928145	closed	Genoscope, France
042[g]	EAEC	O44:H18	5,24	4,810	FN554766	closed	Wellcome Trust Sanger Institute, UK
O157_Sakai[g]	EHEC	O157:H7	5,49	5,363	BA000007	closed	Osaka University, Japan
O103_12009[g]	EHEC	O103:H2	5,44	5,264	AP010958	closed	University of Tokyo, Japan
H10407[g]	ETEC	078:K80:H11	5,15	4,746	FN649414	closed	Wellcome Trust Sanger Institute
MG1655[g]	laboratory	O16*:K12	4,63	4,294	U00096	closed	University of Wisconsin-Madison, USA

[a] ftp://134.76.70.117; UserID: EAHEC_GOS; Password. EAHEC_GOS
[b] number of contigs bigger then 500 bp
[c] ftp://ftp.genomics.org.cn/pub/Ecoli_TY-2482/Escherichia_coli_TY-2482.contig.20110606.fa.gz
[d] http://www.hpa-bioinformatics.org.uk/lgp/genomes
[e] number of scaffolds
[f] based on HPA gene prediction
[g] reference genomes used for comparative BiBag analysis
*O16 = O antigen not expressed in K-12 due to mutation

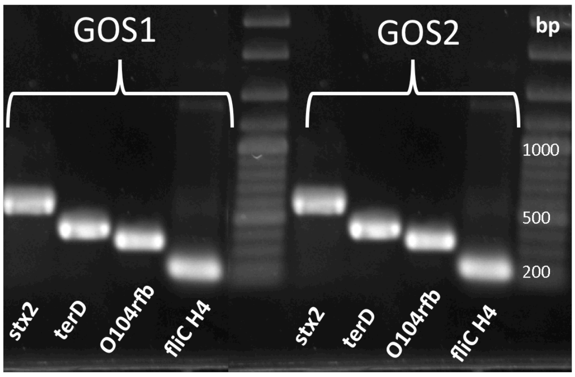


Fig. S1 PCR-based detection of *E. coli* O104:H4 marker genes *stx2, terD, rfbO104*, and *fliC* H4 in the genomes of *E. coli* GOS1 and *E. coli* GOS2. PCR-based detection of the genes was performed according to the PCR typing scheme described by the National Consulting Laboratory on HUS at the University of Münster, Germany (http://www.ehec.org/pdf/ Laborinfo_01062011.pdf). The observed sizes of the PCR products correspond to the sizes of the marker gene PCR products (*stx2*, 584 bp; *terD,* 434 bp; *rfbO104*, 351 bp; and *fliC* H4, 201 bp)

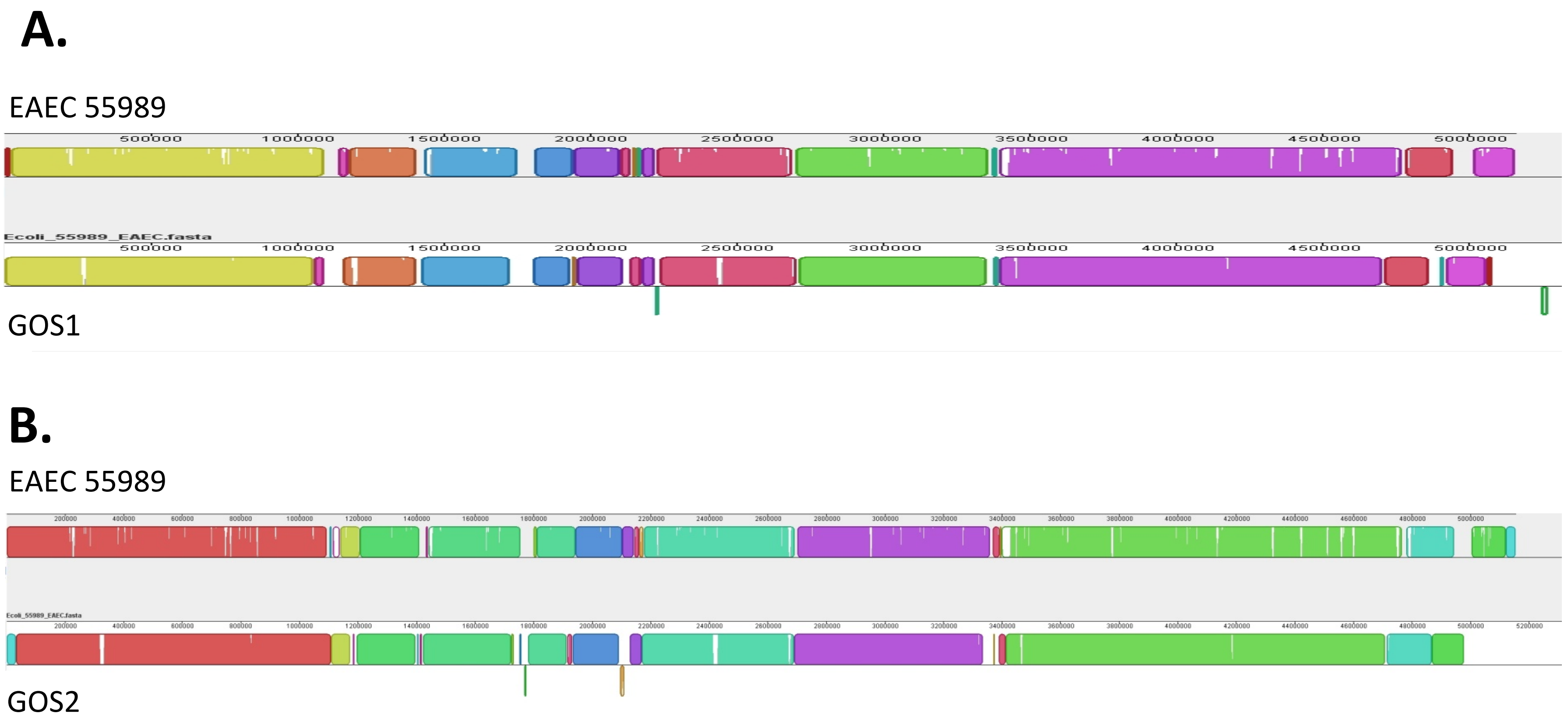


Fig. S2 Contig order of the genomes of *E. coli* GOS1 (A) and GOS2 (B). Contig order was predicted by employing the Mauve Multiple Genome Alignment tool (http://asap.ahabs.wisc.edu/mauve/). Each genome is laid out horizontally and homologous segments are shown as colored blocks that are connected across genomes. Blocks that are shifted downward in any genome represent segments that are inverted relative to the reference genome *E. coli* strain 55989 (top panel).

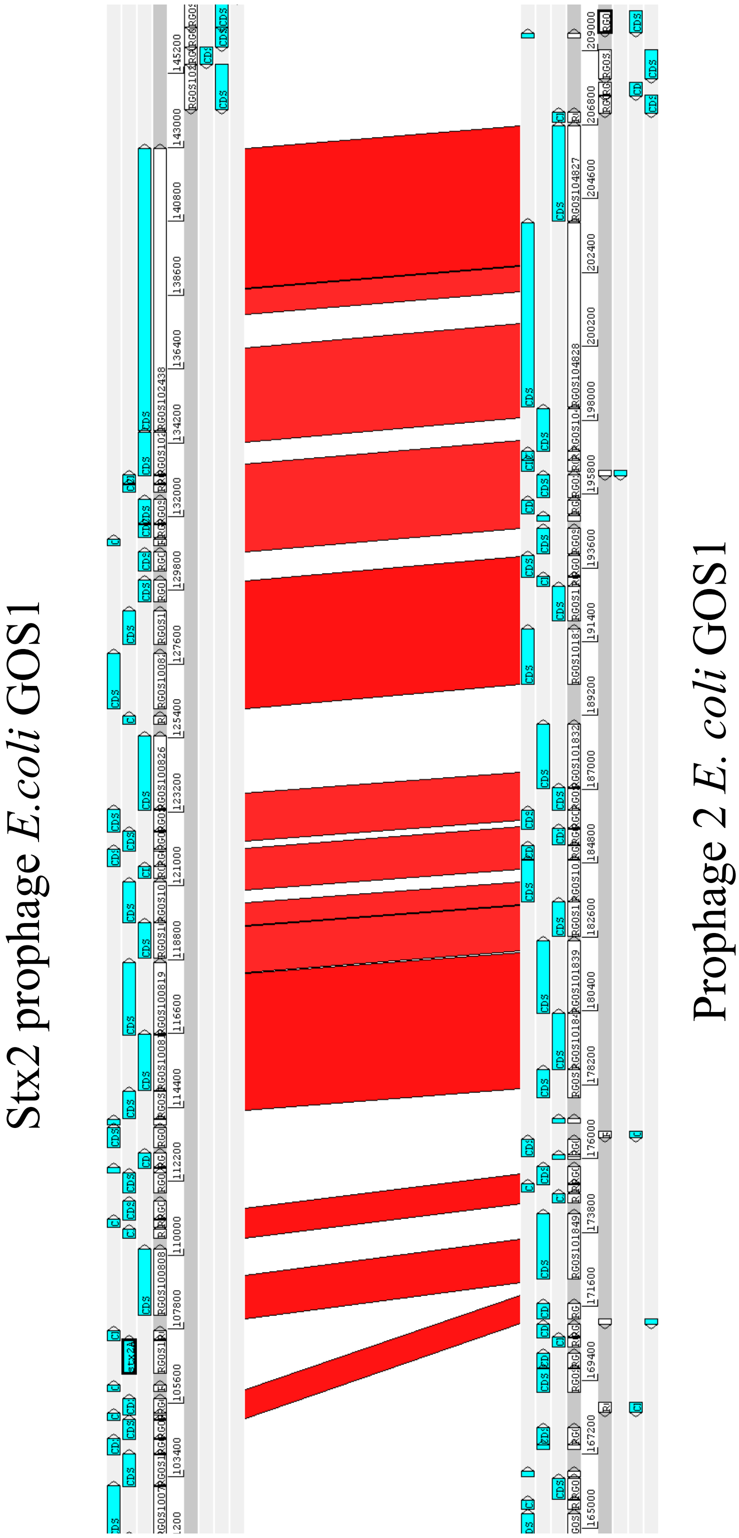


Fig. S3 Comparison of two prophage regions from *Escherichia coli* GOS1 strain. Analyses were done by employing the ACT software tool (Sanger Institute, http://www.sanger.ac.uk)

5.3 ANALYSES OF BOVINE MASTITIS AND COMMENSAL *E. COLI* ISOLATE GENOMES

The following four publications constitute my PhD thesis research on the genomic characterization and comparison of bovine mastitis and fecal commensal *E. coli* isolates (Section 1.5 on page 49).

5.3.1 *The lipopolysaccharide of the mastitis isolate* Escherichia coli *strain 1303 comprises a novel O-antigen and the rare K-12 core type*

Duda KA, Lindner B, Brade H, LEIMBACH A, Brzuszkiewicz E, Dobrindt U, Holst O. 2011. The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type. *Microbiology* 157:1750–1760.
DOI: 10.1099/mic.0.046912-0

5.3.1.1 *Contributions*

Duda et al. (2011) presents the novel genomic sequence and structure of the LPS O-antigen moiety from acute bovine mastitis *E. coli* isolate 1303. Initially, *E. coli* 1303 was falsely typed as an O5 serotype variant. However, we corrected the mistake with new O70 typing information in hindsight (Section 6.1 on page 195). Because LPS is hypothesized to play a deciding role in eliciting bovine mastitis (Section 1.5.2 on page 53) a novel LPS structure is interesting to examine.

The genomic and bioinformatical part of the publication was my contribution, i. e. assembly of the whole genome shotgun sequences of the isolate with homopolymer sequence polishing via PCR amplification and Sanger sequencing of the MAEC 1303 O-antigen region. I analyzed and visualized the genetic structure of the O-antigen region and determined the putative functions of the enclosed open reading frames (ORFs). I wrote the respective parts in the methods section and assisted in writing of the results/discussion section of the publication. Detailed individual author contributions for each part of the paper and each figure/table can be found in Table 14 and Table 15 on pages 232–233, respectively.

5.3.1.2 *Main paper*

Reprinted from Duda et al. (2011) with permission from the Microbiology Society. The publication can be found on pages 120–130 or freely available at:
http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.046912-0

Microbiology (2011), **157**, 1750–1760 DOI 10.1099/mic.0.046912-0

The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type

Katarzyna A. Duda,[1] Buko Lindner,[2] Helmut Brade,[3] Andreas Leimbach,[4,5] Elżbieta Brzuszkiewicz,[5] Ulrich Dobrindt[4,6] and Otto Holst[1]

Correspondence
Katarzyna Duda
kduda@fz-borstel.de

[1]Division of Structural Biochemistry, Research Center Borstel, Leibniz-Center for Medicine and Biosciences, D-23845 Borstel, Germany

[2]Division of Immunochemistry, Research Center Borstel, Leibniz-Center for Medicine and Biosciences, D-23845 Borstel, Germany

[3]Division of Medical and Biochemical Microbiology, Research Center Borstel, Leibniz-Center for Medicine and Biosciences, D-23845 Borstel, Germany

[4]Institute for Molecular Infection Biology, Julius-Maximilians-University of Würzburg, D-97070 Würzburg, Germany

[5]Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University of Göttingen, D-37077 Göttingen, Germany

[6]Institute for Hygiene, University of Münster, D-48149 Münster, Germany

Received 17 November 2010
Revised 23 February 2011
Accepted 1 March 2011

Mastitis represents one of the most significant health problems of dairy herds. The two major causative agents of this disease are *Escherichia coli* and *Staphylococcus aureus*. Of the first, its lipopolysaccharide (LPS) is thought to play a prominent role during infection. Here, we report the O-antigen (OPS, O-specific polysaccharide) structure of the LPS from bovine mastitis isolate *E. coli* 1303. The structure was determined utilizing chemical analyses, mass spectrometry, and 1D and 2D NMR spectroscopy methods. The O-repeating unit was characterized as -[→4)-*β*-D-Qui*p*3NAc-(1→3)-*α*-L-Fuc*p*2OAc-(1→4)-*β*-D-Gal*p*-(1→3)-*α*-D-Gal*p*NAc-(1→]- in which the *O*-acetyl substitution was non-stoichiometric. The nucleotide sequence of the O-antigen gene cluster of *E. coli* 1303 was also determined. This cluster, located between the *gnd* and *galF* genes, contains 13 putative open reading frames, most of which represent unknown nucleotide sequences that have not been described before. The O-antigen of *E. coli* 1303 was shown to substitute O-7 of the terminal LD-heptose of the K-12 core oligosaccharide. Interestingly, the non-OPS-substituted core oligosaccharide represented a truncated version of the K-12 outer core – namely terminal LD-heptose and glucose were missing; however, it possessed a third Kdo residue in the inner core. On the basis of structural and genetic data we show that the mastitis isolate *E. coli* 1303 represents a new serotype and possesses the K-12 core type, which is rather uncommon among human and bovine isolates.

INTRODUCTION

Mastitis is one of the major diseases of cattle, causing high economic losses; bacteria are the main aetiological agents. The outcome of the disease depends on the type of pathogen. Infections induced by *Escherichia coli* often result in an acute mastitis with severe clinical consequences (Petzl *et al.*, 2008). A specific set of virulence-associated genes has not yet been identified for *E. coli* mastitis isolates. Consequently, it has been hypothesized that the cow's genetic predisposition, immune status and lactation stage as well as environmental factors determine the severity of *E. coli* mastitis. Thus, pathogen-associated molecular

Abbreviations: 14 : 0(3-OH), 3-hydroxymyristic acid; COSY: correlation spectroscopy; ESI FT-ICR MS: electrospray ionization Fourier-transformed ion cyclotron resonance mass spectrometry; LD-Hep: L-*glycero*-D-*manno*-heptose; Hex: hexose; HexN: hexosamine; HSQC-DEPT: heteronuclear single-quantum correlation-distortionless enhancement by polarization transfer; Kdo: 3-deoxy-D-*manno*-oct-2-ulosonic acid; NOE: nuclear Overhauser enhancement; OPS: O-specific polysaccharide; P: phosphate; PEtN: 2-aminoethanol phosphate; ROESY: rotating-frame Overhauser-effect spectroscopy; SEC: size-exclusion chromatography; TOCSY: total correlation spectroscopy.

The GenBank accession number for the DNA sequence of the O-antigen gene cluster of *E. coli* strain 1303 described in this paper is FN995094.

A supplementary table is available with the online version of this paper.

patterns (PAMPs), e.g. LPS, could be sufficient to elicit mastitis by *E. coli* (Burvenich *et al.*, 2003). LPS expression contributes to serum resistance and virulence (Raetz & Whitfield, 2002) and may thus offer a selective advantage for *E. coli* during infection of the bovine mammary gland. In many wild-type bacteria LPS consists of the highly antigenic O-specific polysaccharide and the more conserved core oligosaccharide, further divided into outer and inner part and lipid A, the latter of which represents the toxic moiety in toxic LPS (Holst *et al.*, 2009). More than 180 different O-antigens, defining different serogroups, have been described for *E. coli* (Stenutz *et al.*, 2006); however, none has so far been characterized of LPS from a strain causing mastitis. The core region of *E. coli* LPS is represented by five types, R1, R2, R3, R4 and K-12, the chemical structures of which have been published (Jansson *et al.*, 1981; Holst *et al.*, 1991; Haishima *et al.*, 1992; Vinogradov *et al.*, 1999; Müller-Loennies *et al.*, 2002, 2003). The structure of the inner core in case of all these types is very similar and contains the common sequence LD-Hep-(1→7)-[Glc-(1→3)-]-LD-Hep-(1→3)-LD-Hep-(1→5)-[Kdo-(2→4)]-Kdo. It is mainly the structure of the outer core that differentiates the mentioned core types. The unique feature of the K-12 core type is the presence of a fourth Hep residue in the outer core (Holst, 1991). The major core glycoform isolated after complete deacylation of the K-12 LPS possessed the structure shown in Fig. 1 (Müller-Loennies *et al.*, 2003).

The distribution of different *E. coli* core types in the environment is very heterogeneous. Predominantly, the R1 core type is detected in human and cattle populations whereas the K-12 core type is only rarely identified (Heinrichs *et al.*, 1998; Amor *et al.*, 2000; Gibbs *et al.*, 2004). *E. coli* K-12 is commonly used in the laboratory and has ever since its first description in 1944 (Gray & Tatum, 1944) expressed an R-form LPS, which unlike an S-form LPS lacks the O-antigen. Two independent mutations in the *wbbL* O-specific polysaccharide gene cluster were identified in different lineages of *E. coli* K-12. Most strains carry the IS*5* insertion in the last gene of the biosynthetic cluster. It was shown that complementation of the IS*5* mutation leads to the production of an O-antigen in LPS of *E. coli* K-12 which was typed as O16 (Liu & Reeves, 1994). The structure of the O16 antigen was determined (Jann *et al.*, 1994; Stevenson *et al.*, 1994) and it was shown to be attached to O-7 of LD-Hep of the outer core (Feldman *et al.*, 1999).

In this work, the structure of the O-specific polysaccharide (OPS) of the bovine mastitis isolate *E. coli* strain 1303 was elucidated and was shown to be interestingly linked to the K-12 core type. Also, it was proven that the K-12 core type was substituted by the OPS at O-7 of the terminal LD-heptose.

METHODS

Bacterial strain, isolation and degradation of the LPS. *Escherichia coli* 1303, a well-characterized mastitis model strain, was isolated from udder secretions of a cow with clinical mastitis (Petzl *et al.*, 2008). Bacteria were grown in a 10 l fermenter (BIOFLO 110, New Brunswick Scientific) in Luria–Bertani medium, at pH 7.2, 40 % dissolved oxygen and agitation between 300 and 900 r.p.m. The LPS was isolated utilizing the hot phenol/water procedure (Westphal & Jann, 1965), and purified by incubation with DNase and RNase (37 °C, 16 h, with gentle mixing) and proteinase K (56 °C, 6 h with gentle mixing) followed by ultracentrifugation (three times at 105 000 ***g***, 4 °C, 4 h). Subsequently, the LPS (72 mg) was treated with 0.1 M sodium acetate buffer, pH 4.4, for 5 h at 100 °C, and the polysaccharide fraction was separated by size-exclusion chromatography (SEC) on a column of Toyo Pearl HW-40 in 0.05 M pyridinium acetate buffer, pH 4.5 (9 mg). The OPS fraction was further *O*-deacylated utilizing abs. hydrazine (37 °C, 30 min, 6.1 mg; Haishima *et al.*, 1992). Two other fractions were isolated, namely core substituted by an O-antigen (5.6 mg) and the core (4.5 mg). Additionally, another portion of LPS (100 mg) was directly *O*-deacylated and fractionated on Sephacryl 200 eluted with a buffer containing 0.25 % sodium deoxycholate, 0.2 M NaCl, 1 mM EDTA and 10 mM Tris/HCl (pH 9.2). The fraction containing the *O*-deacylated LPS with a short O-antigen (32.6 mg) was used for mass spectrometry analyses.

General and analytical methods. The composition of the isolated fractions was determined by methanolysis (2 M HCl/MeOH, 85 °C, 2 h), followed by acetylation (85 °C, 10 min) and detection by GLC-MS [Hewlett Packard HP 5890 (series II) gas chromatograph equipped with a fused-silica SPB-5 column (Supelco, 30 m × 0.25 mm × 0.25 μm film thickness), FID and MS 5989A mass spectrometer with vacuum gauge controller 59827A]. The temperature programme was 150 °C for 3 min, then 5 °C min^{-1} to 330 °C. Sugars were identified as their alditol acetates after hydrolysis (2 M trifluoroacetic acid, 120 °C, 2 h), reduction ($NaBH_4$, 16 h in the dark) and acetylation (85 °C for 10 min) (Sawardeker *et al.*, 1965) by GLC [HP 5890 (series II) gas chromatograph with FID and a column (30 m × 2.5 mm × 0.25 μm, Agilent Technologies) of polysilican SPD-5]. Helium was used as carrier gas (70 kPa). The temperature programme was 150 °C for 3 min, then

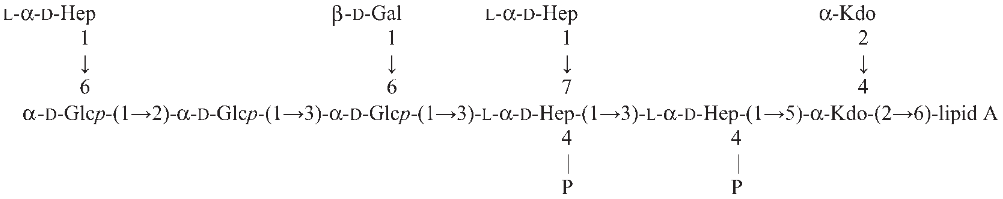

L-α-D-Hep
β-D-Gal
L-α-D-Hep
α-Kdo
α-D-Glc*p*-(1→2)-α-D-Glc*p*-(1→3)-α-D-Glc*p*-(1→3)-L-α-D-Hep-(1→3)-L-α-D-Hep-(1→5)-α-Kdo-(2→6)-lipid A
P
P

Fig. 1. Structure of the major core glycoform isolated after complete deacylation of the K-12 LPS (Müller-Loennies *et al.*, 2003).

3 °C $\min^{-1}$ to 320 °C. The absolute configurations of the sugars were determined as described by Gerwig *et al.* (1979). Methylation was carried out according to Ciucanu & Kerek (1984).

SDS-PAGE and Western blotting. LPS (200 µg) was applied to one large slot (12.5 cm) and separated by SDS-PAGE on a 5 % stacking and 15 % separating gel at a constant voltage of 150 V. The gels were transferred overnight onto PVDF membranes (pore size 0.45 µm, Millipore) by tank blotting (Bio-Rad). Prior to use, the membranes were wetted in methanol for 10 s, after which they were washed in distilled water for at least 5 min. Following transfer, the blots were cut into strips (0.5 cm width) and placed in Mini-incubation trays (Bio-Rad). The following steps were performed at room temperature. After blocking in blotting buffer (50 mM Tris/HCl, 0.2 M NaCl, pH 7.4) supplemented with 10 % non-fat dry milk for 1 h, the antibodies diluted in blocking buffer were added, incubated for 16 h and washed six times (5 min each) in blotting buffer. Alkaline-phosphatase-conjugated goat anti-mouse IgG (heavy and light chain specific, Dianova) was added (diluted 1 : 1000 in blotting buffer) and incubation was continued for another 2 h. After washing as before, 5-bromo-4-chloro-3-indoyl phosphate and *p*-toluidine *p*-nitro blue tetrazolium chloride (Bio-Rad) were added as substrates according to the supplier's instruction. After 15 min the reaction was stopped by the addition of distilled water. The monoclonal antibodies used were FDP-11 (specific for the R1 core type), FDP-3 (reacting with the R2 and K-12 core type) (both antibodies were obtained from F. Di Padova, Novartis Basel), S37-20 (specific for the R3 core type; unpublished data of H. Brade and others), S31-14 (specific for the K-12 core type; Brade *et al.*, 1996) and WN1 222-5 (reacting with all five *E. coli* core types; Di Padova *et al.*, 1993).

Mass spectrometry. Electrospray ionization Fourier-transformed ion cyclotron resonance (ESI FT-ICR) MS was performed in the negative-ion mode using an APEX Qe instrument (Bruker Daltonics) equipped with a 7 T magnet and a dual Apollo ion source. Mass spectra were acquired in broad band modes. The samples (~10 ng μl^{-1}) were dissolved in a 50 : 50 : 0.001 (by vol.) mixture of 2-propanol, water and triethylamine, and were sprayed at a flow rate of 2 µl $\min^{-1}$. Capillary entrance voltage was set to 3.8 kV, and drying gas temperature to 150 °C. Mass spectra were calibrated externally by lipids of known structure, charge deconvoluted, and the given mass numbers were referred to the monoisotopic masses of neutral molecules.

NMR spectroscopy. NMR spectroscopy experiments were carried out after H–^{2}H exchange of the samples utilizing 99.9 % ^2H_2O. All 1D (^{1}H, ^{13}C), and 2D homonuclear (^{1}H, ^{1}H) correlation spectroscopy (COSY), total correlation spectroscopy (TOCSY), and rotating-frame Overhauser-effect spectroscopy (ROESY), as well as heteronuclear (^{1}H, ^{13}C) single-quantum correlation-distortionless enhancement by polarization transfer (HSQC-DEPT) experiments of native OPS, core fraction with short O-antigen (core-short OPS) and non-substituted core fraction were recorded at 300 K with a Bruker DRX Avance 600 MHz spectrometer (operating frequencies 600.31 MHz for ^{1}H NMR, 150.96 MHz for ^{13}C NMR), equipped with a 5 mm QXI multinuclear-inverse probehead with a *z* gradient, and applying standard Bruker software. Spectra of *O*-deacylated OPS were recorded at 300 K with a Bruker DRX Avance 700 MHz spectrometer (operating frequencies 700.75 MHz for ^{1}H NMR, 176.2 MHz for ^{13}C NMR), equipped with a 5 mm CPQCI multinuclear-inverse cryo-probehead with a *z* gradient, and applying standard Bruker software. Chemical shifts were reported relative to an internal standard of acetone (δ_H 2.225, δ_C 31.45). Mixing times of 100 and 250 ms were used in TOCSY and ROESY experiments, respectively.

DNA sequence analysis. Total DNA of *E. coli* 1303 was prepared with the MasterPure DNA Purification kit (Epicentre) according to the manufacturer's instructions. Sequencing was performed with a Genome Sequencer FLX system (Roche Applied Science). The resulting whole-genome shotgun reads were *de novo* assembled with the Roche Newbler assembly software (Margulies *et al.*, 2005).

The program ARTEMIS, version 11 (Rutherford *et al.*, 2000), was used for annotation. Homopolymer stretches leading to frameshifts in the O-antigen gene cluster were resolved by amplifying the region utilizing PCR and subsequent Sanger sequencing. BLAST and PSI-BLAST (Altschul *et al.*, 1997) were used for searching databases including GenBank, COG and the Pfam protein motif database (Bateman *et al.*, 2002). The program TMHMM 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) was used to identify potential transmembrane segments.

The DNA sequence of the O-antigen gene cluster of *E. coli* strain 1303 has been deposited in GenBank under the accession number FN995094.

RESULTS AND DISCUSSION

The structure of the OPS

The LPS was isolated from bacterial cells utilizing hot phenol/water extraction and subsequently purified by enzymic treatment and ultracentrifugation (yield: 1.23 % of bacterial dry weight). Part of it was hydrolysed under mild acidic conditions to give the native OPS, which was purified by SEC (yield: 12.5 % of the LPS) and further *O*-deacylated (*O*-deacylated OPS) by mild hydrazine treatment (37 °C, 30 min, yield: 8.5 % of the LPS).

Compositional analyses of the native OPS fraction revealed the presence of fucose (Fuc), 3-amino-6-deoxyhexose (Qui3N), Gal and GalN. The absolute configurations of Gal and GalN were identified as D and that of Fuc as L. Methylation analysis of the *O*-deacylated OPS identified 1,3,5-tri-*O*-acetyl-6-deoxy-2,4-di-*O*-methylgalactose, 1,4,5-tri-*O*-acetyl-2,3,6-tri-*O*-methylgalactose, 1,4,5-tri-*O*-acetyl-3,6-dideoxy-2-*O*-methyl-3-methylamidoglucose and 1,3,5-tri-*O*-acetyl-2-deoxy-4,6-di-*O*-methyl-2-methylamidogalactose, indicating the OPS to be composed of the four units of 3-substituted L-Fuc, 4-substituted D-Gal, 4-substituted Qui3N and 3-substituted D-GalN. All residues were pyranoses.

These results were further confirmed by high-resolution ESI FT-ICR MS of the native OPS, which possessed groups of molecules representing the core oligosaccharides differing by one O-chain repeating unit. The measured mass difference of 740.285 Da was in excellent agreement with the calculated mass of 740.286 Da of a repeating unit consisting of 1 HexN, 1 deoxy-Hex, 1 Hex, 1 deoxy-HexN and 3 acetyl groups ($C_{30}H_{48}O_{19}N_2$) (mass spectrum not shown). Similar data with one acetyl group less ($C_{28}H_{46}O_{18}N_2$, 698.274 Da) were obtained for the *O*-deacylated OPS.

The structure of the O-repeating unit was established by NMR spectroscopy. The complete assignment of the OPS ^{1}H and ^{13}C resonances (Table 1) was achieved by combining the information obtained from COSY, TOCSY and ROESY, as well as HSQC-DEPT experiments. The ^{1}H NMR spectrum of the native OPS (Fig. 2a) identified five signals in the anomeric region at δ_H 5.24,

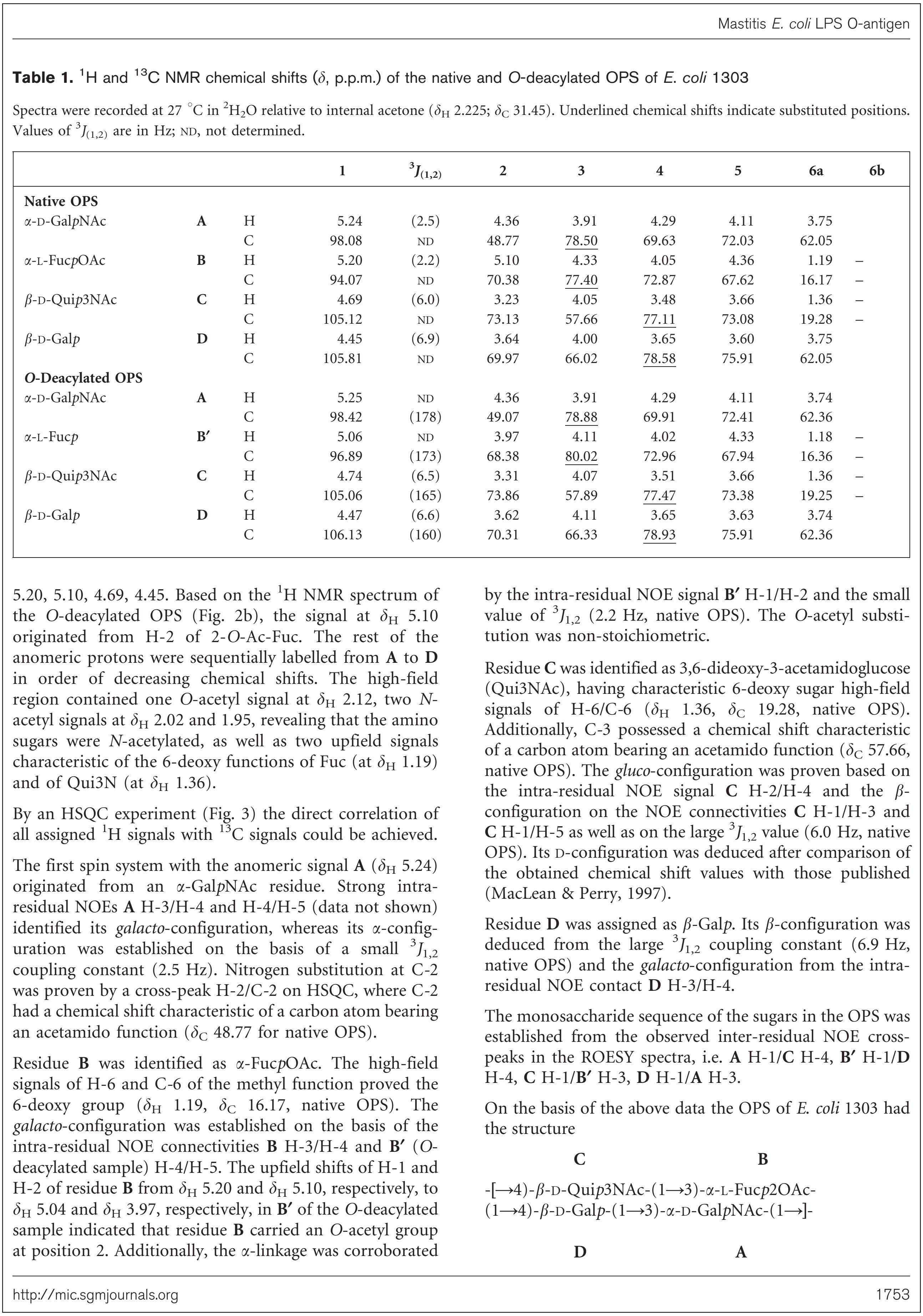


Table 1. ^{1}H and ^{13}C NMR chemical shifts (δ, p.p.m.) of the native and *O*-deacylated OPS of *E. coli* 1303

Spectra were recorded at 27 °C in $^{2}H_2O$ relative to internal acetone (δ_H 2.225; δ_C 31.45). Underlined chemical shifts indicate substituted positions. Values of $^{3}J_{(1,2)}$ are in Hz; ND, not determined.

			1	$^{3}J_{(1,2)}$	2	3	4	5	6a	6b
Native OPS										
α-D-Gal*p*NAc	**A**	H	5.24	(2.5)	4.36	3.91	4.29	4.11	3.75	
		C	98.08	ND	48.77	78.50	69.63	72.03	62.05	
α-L-Fuc*p*OAc	**B**	H	5.20	(2.2)	5.10	4.33	4.05	4.36	1.19	–
		C	94.07	ND	70.38	77.40	72.87	67.62	16.17	–
β-D-Qui*p*3NAc	**C**	H	4.69	(6.0)	3.23	4.05	3.48	3.66	1.36	–
		C	105.12	ND	73.13	57.66	77.11	73.08	19.28	–
β-D-Gal*p*	**D**	H	4.45	(6.9)	3.64	4.00	3.65	3.60	3.75	
		C	105.81	ND	69.97	66.02	78.58	75.91	62.05	
***O*-Deacylated OPS**										
α-D-Gal*p*NAc	**A**	H	5.25	ND	4.36	3.91	4.29	4.11	3.74	
		C	98.42	(178)	49.07	78.88	69.91	72.41	62.36	
α-L-Fuc*p*	**B′**	H	5.06	ND	3.97	4.11	4.02	4.33	1.18	–
		C	96.89	(173)	68.38	80.02	72.96	67.94	16.36	–
β-D-Qui*p*3NAc	**C**	H	4.74	(6.5)	3.31	4.07	3.51	3.66	1.36	–
		C	105.06	(165)	73.86	57.89	77.47	73.38	19.25	–
β-D-Gal*p*	**D**	H	4.47	(6.6)	3.62	4.11	3.65	3.63	3.74	
		C	106.13	(160)	70.31	66.33	78.93	75.91	62.36	

5.20, 5.10, 4.69, 4.45. Based on the ^{1}H NMR spectrum of the *O*-deacylated OPS (Fig. 2b), the signal at δ_H 5.10 originated from H-2 of 2-*O*-Ac-Fuc. The rest of the anomeric protons were sequentially labelled from **A** to **D** in order of decreasing chemical shifts. The high-field region contained one *O*-acetyl signal at δ_H 2.12, two *N*-acetyl signals at δ_H 2.02 and 1.95, revealing that the amino sugars were *N*-acetylated, as well as two upfield signals characteristic of the 6-deoxy functions of Fuc (at δ_H 1.19) and of Qui3N (at δ_H 1.36).

By an HSQC experiment (Fig. 3) the direct correlation of all assigned ^{1}H signals with ^{13}C signals could be achieved.

The first spin system with the anomeric signal **A** (δ_H 5.24) originated from an α-Gal*p*NAc residue. Strong intra-residual NOEs **A** H-3/H-4 and H-4/H-5 (data not shown) identified its *galacto*-configuration, whereas its α-configuration was established on the basis of a small $^{3}J_{1,2}$ coupling constant (2.5 Hz). Nitrogen substitution at C-2 was proven by a cross-peak H-2/C-2 on HSQC, where C-2 had a chemical shift characteristic of a carbon atom bearing an acetamido function (δ_C 48.77 for native OPS).

Residue **B** was identified as α-Fuc*p*OAc. The high-field signals of H-6 and C-6 of the methyl function proved the 6-deoxy group (δ_H 1.19, δ_C 16.17, native OPS). The *galacto*-configuration was established on the basis of the intra-residual NOE connectivities **B** H-3/H-4 and **B′** (*O*-deacylated sample) H-4/H-5. The upfield shifts of H-1 and H-2 of residue **B** from δ_H 5.20 and δ_H 5.10, respectively, to δ_H 5.04 and δ_H 3.97, respectively, in **B′** of the *O*-deacylated sample indicated that residue **B** carried an *O*-acetyl group at position 2. Additionally, the α-linkage was corroborated by the intra-residual NOE signal **B′** H-1/H-2 and the small value of $^{3}J_{1,2}$ (2.2 Hz, native OPS). The *O*-acetyl substitution was non-stoichiometric.

Residue **C** was identified as 3,6-dideoxy-3-acetamidoglucose (Qui3NAc), having characteristic 6-deoxy sugar high-field signals of H-6/C-6 (δ_H 1.36, δ_C 19.28, native OPS). Additionally, C-3 possessed a chemical shift characteristic of a carbon atom bearing an acetamido function (δ_C 57.66, native OPS). The *gluco*-configuration was proven based on the intra-residual NOE signal **C** H-2/H-4 and the β-configuration on the NOE connectivities **C** H-1/H-3 and **C** H-1/H-5 as well as on the large $^{3}J_{1,2}$ value (6.0 Hz, native OPS). Its D-configuration was deduced after comparison of the obtained chemical shift values with those published (MacLean & Perry, 1997).

Residue **D** was assigned as β-Gal*p*. Its β-configuration was deduced from the large $^{3}J_{1,2}$ coupling constant (6.9 Hz, native OPS) and the *galacto*-configuration from the intra-residual NOE contact **D** H-3/H-4.

The monosaccharide sequence of the sugars in the OPS was established from the observed inter-residual NOE cross-peaks in the ROESY spectra, i.e. **A** H-1/**C** H-4, **B′** H-1/**D** H-4, **C** H-1/**B′** H-3, **D** H-1/**A** H-3.

On the basis of the above data the OPS of *E. coli* 1303 had the structure

C **B**

-[→4)-β-D-Qui*p*3NAc-(1→3)-α-L-Fuc*p*2OAc-
(1→4)-β-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-

D **A**

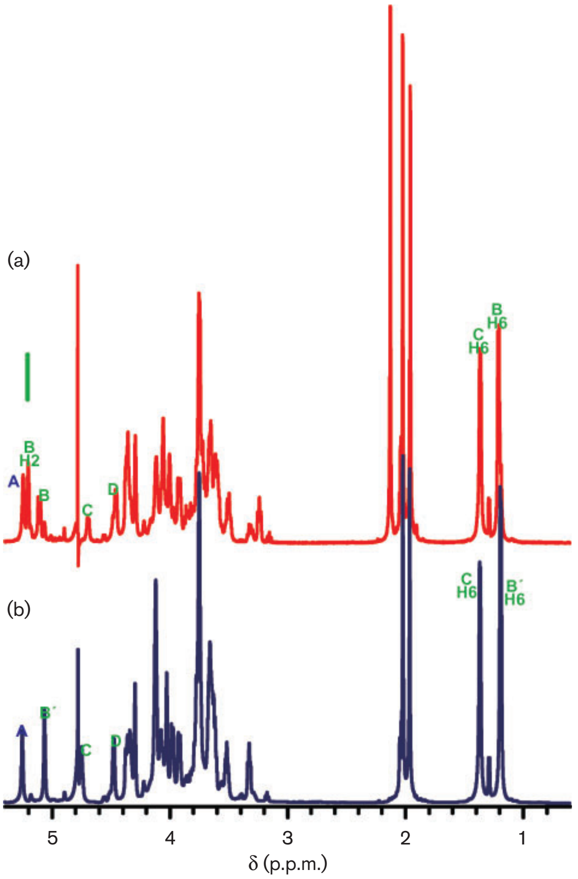


Fig. 2. ^{1}H NMR spectrum of the native OPS (600 MHz, 300K, ^2H_2O) (a) and *O*-deacylated OPS (700 MHz, 300 K, ^2H_2O) (b) of *E. coli* 1303.

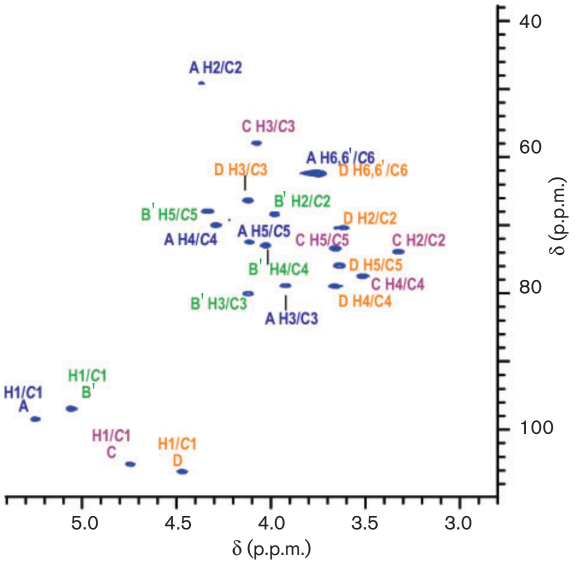


Fig. 3. Heteronuclear 2D ^{13}C–^{1}H chemical shift correlation of the anomeric and ring region resonances of the *O*-deacylated OPS of *E. coli* 1303. The spectrum was recorded at 700 MHz and 300 K.

The structure of the OPS of the mastitis isolate *E. coli* 1303 showed great similarities with, but was not identical to, those present in the LPS of two subtypes of the *E. coli* O5 serotype known to date, namely O5ab (MacLean & Perry, 1997) and O5ac (strain 180/C3) (Urbina *et al.*, 2005). In OPS of *E. coli* 1303 α-L-Fuc*p*OAc was present instead of *β*-D-Rib*f* as in OPS of O5ab and O5ac, and additionally *β*-D-Qui*p*3NAc was substituted at position 4 and not at position 2 as in the O5ac strain.

E. coli O5ab:

-[→4)-*β*-D-Qui*p*3NAc-(1→3)-*β*-D-Rib*f*-(1→4)-*β*-D-Gal*p*-(1→3)-α-D-Gal*p*NAc-(1→]-

E. coli O5ac:

-[→2)-*β*-D-Qui*p*3NAc-(1→3)-*β*-D-Rib*f*-(1→4)-*β*-D-Gal*p*-(1→3)-α-D-Gal*p*NAc(1→]-

Sequence analysis of the O-antigen gene cluster

The O-antigen determinant located between *galF* and *gnd* on the *E. coli* strain 1303 chromosome, as also reported for other O-antigen clusters (Reeves & Wang, 2002), was sequenced and the genetic structure of this 13 095 bp DNA region was analysed in detail. The O-antigen gene cluster has an overall G+C content of 36.9 mol% (Fig. 4a). As also described for other *E. coli* O antigen gene clusters, all predicted ORFs, with the exception of one transposase-encoding ORF, have a lower G+C content than the average *E. coli* genome, suggesting that they may have been acquired by horizontal transfer from other species (Reeves & Wang, 2002).

Thirteen putative open reading frames (ORFs) with the same transcriptional direction were identified as shown in Fig. 4(b). The majority of the putative ORFs represented unknown nucleotide sequences that have not been described before. Whereas the *rmlB* gene (ORF 1) and an IS*4*-family transposase-encoding gene (ORF 12) have been described in other *E. coli* isolates as well, only *rmlA* (ORF 2) and ORF 3 exhibited similarity (72 % and 74 % identity) to fragments of the glucose-1-phosphate thymidylyltransferase-encoding gene Abu_1817 of *Arcobacter butzleri* RM4018 (accession no. CP000361) and *fdtA* of *Salmonella enterica* subsp. *enterica* serovar Pomona strain NML 07-0213 (accession no. EU805803), respectively.

The gene products of ORF 1 (*rmlB*) and ORF 2 (*rmlA*) showed 92 % and 82 % identity to other known RmlB and RmlA homologues, respectively (Table 2; see also Supplementary Table S1, available with the online version of this paper). The *rmlA* and *rmlB* genes have been well characterized in *E. coli*. RmlA converts D-glucose 1-phosphate to dTDP-D-glucose, which is then converted by RmlB to dTDP-4-dehydro-6-deoxy-D-glucose. The latter

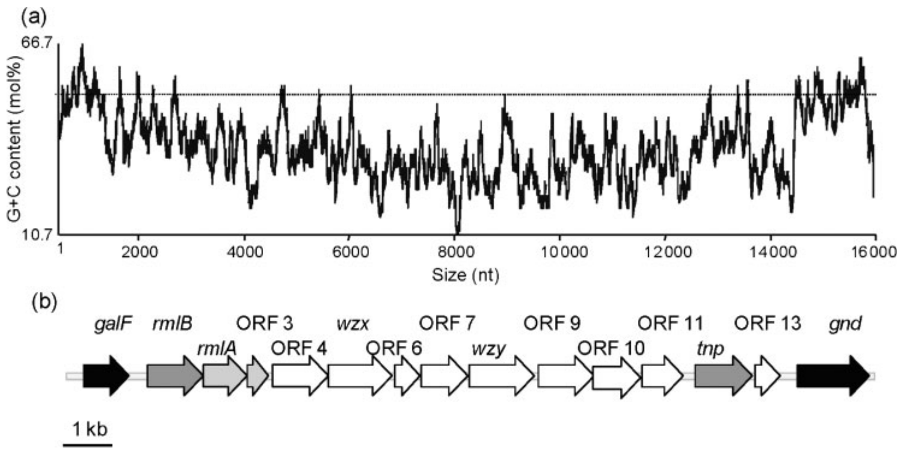


Fig. 4. Genetic structure of the antigen-encoding determinant of the mastitis isolate *E. coli* 1303. (a) G+C content of the O-antigen gene cluster relative to the average G+C content of the *E. coli* core chromosome (50.8 %, horizontal line). (b) The positions and transcriptional directions of identified putative ORFs are indicated by arrows. The *galF* and *gnd* genes flanking the O-antigen gene cluster are indicated in black. ORFs with high similarity to known bacterial DNA sequences are indicated in grey.

compound is a common intermediate of many different sugars (Graninger *et al.*, 2002; Samuel & Reeves, 2003). Similar to this OPS and the O5 oligosaccharide variants, the *E. coli* O91 oligosaccharide contains, among other sugar residues, a modified D-Qui*p*3N. For the corresponding *wb**$_{\mathrm{O91}}$ gene cluster the *wbsB* gene has been predicted to code for an isomerase catalysing the conversion of dTDP-4-dehydro-6-deoxy-D-glucose into dTDP-3-dehydro-6-deoxy-D-glucose (Perelle *et al.*, 2002). Also in *Thermoanaerobacterium thermosaccharolyticum* dTDP-α-D-Qui*p*3NAc biosynthesis has been shown to include an isomerase which catalyses the formation of dTDP-3-dehydro-6-deoxy-D-glucose from the RmlB product dTDP-4-dehydro-6-deoxy-D-glucose. In the case of the O-antigen determinant of *E. coli* 1303, the ORF 3-encoded protein contained a C-terminal isomerase domain of WxcM-like proteins (PF05523) and was 69 % and 46 % identical to the dTDP-4-oxo-6-deoxy-D-glucose-3,4-oxoisomerase WbsB (accession no. AAK60451) of *E. coli* O91 and the QdtA isomerase (accession no. AAR85518) of *T. thermosaccharolyticum* E207-71, respec-
tively. Therefore, ORF 3 (*qdtA*) may encode an isomerase responsible for the formation of dTDP-3-dehydro-6-deoxy-D-glucose. In *T. thermosaccharolyticum,* 3-acetamido-3,

Table 2. Characteristics of the ORFs located in the O-antigen gene cluster of *E. coli* 1303

An extended version of this table is available as Supplementary Table S1 with the online version of this paper.

Putative ORF no.	Designation	Length (bp)	G+C content (mol%)	No. of aa	Putative function of protein
1	*rmlB* (GI:315461754)	1086	43.4	361	dTDP-glucose 4,6-dehydratase
2	*rmlA* (GI:315461755)	870	37.8	289	Glucose-1-phosphate thymidylyltransferase
3	*qdtA* (GI:315461756)	408	36.3	135	dTDP-4-oxo-6-deoxy-D-glucose-3,4-oxoisomerase
4	*qdtB* (GI:315461757)	1104	36.8	367	Transaminase
5	*wzx* (GI:315461758)	1251	34.9	416	O-antigen flippase
6	*wbnC* (GI:315461759)	513	30.01	170	Acetyltransferase
7	ORF 7 (GI:315461760)	921	30.8	306	Glycosyltransferase
8	*wzy* (GI:315461761)	1281	30.1	426	O-antigen polymerase
9	ORF 9 (GI:315461762)	1083	30.8	360	Glycosyltransferase
10	*wcaG* (GI:315461763)	924	30.8	307	Nucleoside-diphosphate-sugar epimerase
11	ORF 11 (GI:315461764)	810	34.2	269	Glycosyltransferase
12	*tnp* (GI:315461765)	1119	41.1	372	Transposase
13	*qdtC* (GI:315461766)	483	32.7	160	Transacetylase

6-dideoxy-α-D-glucose formation involves subsequent reaction steps catalysed by a transaminase and a transacetylase (Pföstl *et al.*, 2008). The corresponding enzymes of the tested O-antigen gene cluster may be encoded by ORF 4 and ORF 13. The ORF 4 (*qdtB*)-encoded gene product belonged to the DegT/DnrJ/EryC1/StrS aminotransferase family and showed 49 % identity to the transaminase QdtB (accession no. AAR85519) of *T. thermosaccharolyticum* E207-71. ORF 13 (*qdtC*) coded for a putative transacetylase with a WcxM-like, left-handed parallel *β*-helix (LbH) N-terminal domain responsible for the transacetylation function. This protein exhibited 76 % identity to the transacetylase QdtC of *Salmonella choleraesuis* (accession no. D7PFB5) and 44 % identity to the QdtC transacetylase of *T. thermosaccharolyticum* E207-71 (accession no. AAR85517).

The OPS of *E. coli* 1303 also contained α-L-Fuc*p*OAc. Fucose biosynthesis involves a three-step pathway for converting GDP-D-mannose to GDP-L-fucose (Ginsburg, 1961). First, the GDP-mannose dehydratase Gmd converts GDP-mannose into GDP-4-keto-6-deoxymannose. This is followed by epimerase and reductase reaction steps to give GDP-L-fucose. In *E. coli*, the latter two steps are catalysed by a single bifunctional enzyme, the GDP-fucose synthase Fcl (Samuel & Reeves, 2003). The *gmd* and *fcl* genes are usually adjacently located within the colanic acid gene cluster of *E. coli*, as has been shown for *E. coli* MG1655 (Andrianopoulos *et al.*, 1998). In the *wb**$_{1303}$ gene cluster, the protein encoded by ORF 10 exhibited 42 % and 22 % identity to the putative UDP-glucose-4-epimerase EDWATA_01329 (accession no. D4F3L9) of *Edwardsiella tarda* ATCC 23685 and the GDP-fucose synthase Fcl (accession no. NP_416556) of *Escherichia coli* MG1655. The presence of a conserved domain of the NAD-dependent epimerase/dehydratase family (PF01370), which also included the conserved WcaG nucleoside-diphosphate-sugar epimerase domain (COG0451), further supports the idea that this protein may be involved in fucose biosynthesis. The protein encoded by ORF 6 (*wbnC*) belonged to the trimeric LpxA-like enzymes superfamily of acetyltransferases from a wide range of bacteria and is 39 % identical to the *O*-acetyltransferase WbnC (Q9RP59) of *E. coli* O113. Thus, WbnC is likely to be responsible for *O*-acetylation of Fuc in the tested O-polysaccharide, although further experimental determination is needed. The O-antigen flippase (Wzx) and O-antigen polymerase (Wzy) are hydrophobic membrane proteins involved in O-antigen processing. In *E. coli* strain 1303, 12 transmembrane helices were predicted for the deduced amino acid sequence of ORF 5 (*wzx*), which shared 42 % identity with the *E. coli* O114 Wzx protein (Q697E1). ORF 8 (*wzy*) showed no marked similarity to other *wzy* genes, but as the deduced amino acid sequence exhibited some homology to the putative O-antigen polymerases it was considered as the putative *wzy* gene. Furthermore, 11 transmembrane helices have been predicted from the deduced amino acid sequence. This is also the case for the putative Wzy proteins of an *E. coli* O4 strain (accession no. AAC43898) and of *E. coli* K-12 (accession no. AAB88404). Generally, the number of transmembrane helices of Wzy proteins of other *E. coli* serogroups is variable, ranging from 8 to 12. A large number of transmembrane segments and a large periplasmic loop are typical topological characteristics of Wzy proteins (Daniels *et al.*, 1998).

Four additional ORFs involved in O-antigen biosynthesis have been predicted (ORFs 7, 9, 11 and 12). Their precise function will have to be studied experimentally. ORF 7 coded for a member of the glycosyltransferase family 2 which is 41 % and 34 % identical to the glycosyltransferase ESA_01184 (accession no. A9Y3E9) of *Enterobacter sakazakii* (Mullane *et al.*, 2008) or putative glycosyltransferase WbtE (accession no. Q6QNC3) of *Escherichia coli* O103. The ORF 9-encoded protein belonged to the group 1 glycosyltransferases and showed 30 % identity to the glycosyltransferase WclF of *E. coli* O155 (accession no. AAV74551) and 27 % identity to WbuB (accession no. AAT28929), a protein encoded in the *E. coli* O26 O-antigen gene cluster (D'Souza *et al.*, 2002). The fucosyltransferase WbuB has been proposed to be a transferase for the linkage α-L-Fuc*p*NAc-(1→3)-α-D-Glc*p*NAc (D'Souza *et al.*, 2002). This linkage is absent in the *E. coli* 1303 O-antigen, but as an α-L-Fuc*p*2OAc-(1→4)-*β*-D-Gal linkage is present, ORF 9 may encode a Fuc*p*OAc transferase. The ORF 11 gene product belonged to the glycosyltransferase family 2 and shares 60 % or 48 % amino acid identity with glucosyltransferase WbeD (accession no. B8R1W7) of *Salmonella enterica* serovar Pomona and *E. coli* O107 (accession no. B8QSK1), respectively. In addition, the ORF 11-encoded protein showed 49 % identity to WbwC of *E. coli* O104, which is a galactosyltransferase responsible for the *β*-D-Gal-(1→3)-*β*-D-GalNAc linkage (Wang *et al.*, 2009). Accordingly, ORF 11 encoded a galactosyltransferase which may link the Gal*p* and Gal*p*NAc moieties present in the 1303 OPS.

ORF 12 encoded a protein unrelated to O-antigen biosynthesis. Together with its sequence context, ORF 12 was identified as a remnant of an H-repeat-type transposable element (positions 1863–3744). H-repeats may be involved in horizontal gene transfer and in the generation of polymorphisms in O-antigen gene clusters (Xiang *et al.*, 1994).

In *E. coli*, the genes for the synthesis of nucleotide precursors of common sugars, e.g. Glc*p*NAc, Glc*p* and Gal*p*, are usually located outside the O-antigen gene cluster (Samuel & Reeves, 2003). UDP-GalNAc is synthesized from UDP-GlcNAc by the UDP-GlcNAc-4-epimerase encoded by the *gne* gene, which can be part of the O-antigen determinant located between *galF* and *gnd*. Alternatively, *gne* can be immediately upstream of *galF* (Wang *et al.*, 2005). In *E. coli* strain 1303, the *gne* gene is absent from the O-antigen gene cluster and is likely to be upstream of *galF*. O-repeating unit synthesis in enterobacteria is often initiated by transferring GlcNAc 1-phosphate or GalNAc 1-phosphate to an undecaprenol phosphate carrier. The corresponding enzyme WecA is encoded outside the O-antigen gene cluster as well as other O-antigen processing proteins, such as the chain-length determinant Wzz (Samuel & Reeves, 2003).

Identification of the core oligosaccharide type and the linkage site of the OPS

Western blot analysis of LPS from *E. coli* 1303 was performed with monoclonal antibodies specific for the different *E. coli* core types; this revealed that strain 1303 carried the K-12 core type in its LPS (Fig. 5). This result was rather unexpected, since (i) this *E. coli* core type had been detected earlier in only 4 % of faecal human and bovine isolates (Gibbs *et al.*, 2004), (ii) *E. coli* K-12 strains that are widely used in laboratories produce an R-form LPS lacking OPS repeating units (Feldman *et al.*, 1999), and (iii) an *E. coli* K-12 strain in which O-antigen assembly was restored exhibited serotype O16 (Liu & Reeves, 1994).

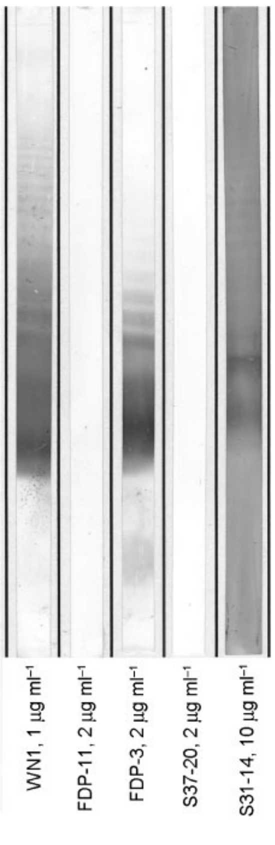


Fig. 5. Western blot of LPS from *E. coli* 1303 strain (amounts given on the figure) with monoclonal antibodies specific for different *E. coli* core types: WN1 222-5, binds to all five *E. coli* core types (Di Padova *et al.*, 1993); FDP-11, specific for R1 core type; FDP-3, specific for R2 and K-12 core types; S37-20, specific for R3 core type; and S31-14, specific for K-12 core type (Brade *et al.*, 1996).

The ESI MS spectrum of *O*-deacylated LPS (Fig. 6) comprised four groups of molecules showing heterogeneity originating from non-stoichiometric substitutions with PEtN, sodium and potassium adducts (not labelled). The first complex group of molecular ions around 2957.93 u referred to the *O*-deacylated lipid A + core whereas the three other groups around 3790.26, 4488.55 and 5186.81 u represented the *O*-deacylated lipid A, the core + one, two and three repeating units, respectively, in which each repeating unit consisted of HexN, deoxy-Hex, Hex, deoxy-HexN and two *N*-acetyl groups, with a total mass of 698.27 u. The most prominent molecular peak at 2957.93 u represented a molecule composed of 4 P, 1 PEtN, 2 HexN, 2 14 : 0(3-OH), 3 Kdo, 3 Hep and 3 Hex corresponding to *O*-deacylated lipid A plus a truncated K-12 core region with an additional third Kdo residue, as compared to the published data (Holst *et al.*, 1991; Müller-Loennies *et al.*, 2003).

The component belonging to the molecular peak at 2737.86 u was composed of 4 P, 1 PEtN, 2 HexN, 2 14 : 0(3-OH), 3 Hep, 3 Hex and 2 Kdo, and was the core species to which OPS was attached after previous addition of 1 Hex moiety and 1 Hep moiety. The molecular ion corresponding to the complete core (Müller-Loennies *et al.*, 2003) was not observed; however, the molecule consisting of the complete core and one *O*-deacylated O-repeating unit (HexN, deoxy-Hex, Hex, deoxy-HexN, two *N*-acetyl groups, calculated mass 698.27 u) was found at 3790.26 u. Thus, the mass at 2957.93 u originated from the core oligosaccharide that was not substituted by OPS, and that at 3790.26 u from a core oligosaccharide substituted by one O-antigen repeating unit. Interestingly, the non-substituted core oligosaccharide differed in structure from those substituted with the OPS, i.e. it represented a truncated version with an additional, third Kdo residue. Moreover, the substituted core corresponded to the most prominent glycoform, namely glycoform 1 of the K-12 core type (Müller-Loennies *et al.*, 2003), and the non-substituted oligosaccharide represented a novel glycoform of the K-12 structure, i.e. a truncated core lacking any rhamnose residues. Previously the occurrence of three Kdo moieties in the truncated K-12 core was associated with simultaneous presence of one rhamnose (Rha) residue (Müller-Loennies *et al.*, 2003).

The NMR study of the fraction core-short OPS isolated from the LPS after acetate buffer hydrolysis and separation on Toyo Pearl HW-40 identified the biological O-repeating unit of the OPS from *E. coli* 1303 and the linkage site to the core region. The spin systems of two OPS repeating units, one terminal (with a terminal β-D-Qui*p*3NAc residue) and one connected to the core (where the 3-substituted D-Gal*p*NAc was shown to be β-configured, and not α-configured as in the other O-repeating units) were identified. Due to the high heterogeneity of the sample, the complete spin systems of only the three first sugars of the

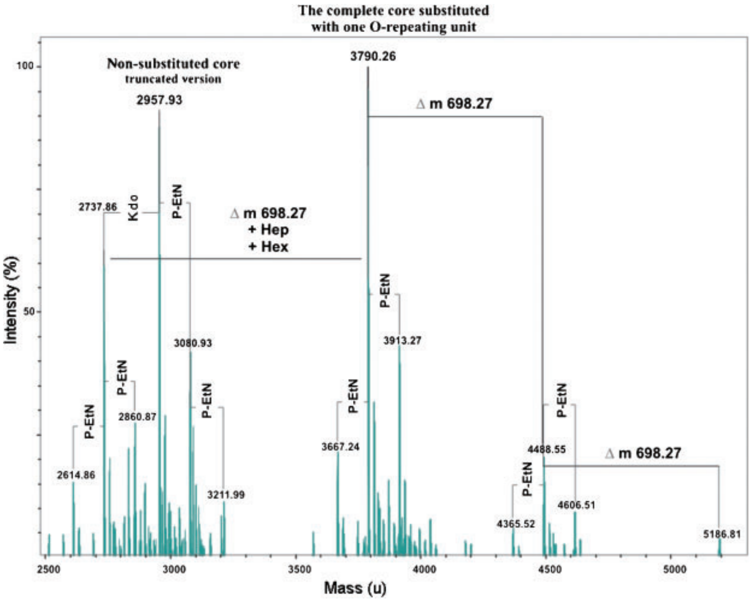


Fig. 6. Charge-deconvoluted ESI FT-ICR MS spectrum of *O*-deacylated LPS from *E. coli* 1303 recorded in the negative-ion mode. The molecular peak at 2957.93 u corresponds to a molecule composed of 2 P, 2 HexN, 2 14 : 0(3-OH) (*O*-deacylated lipid A), 2 P, 1 PEtN, 3 Kdo, 3 Hep, 3 Hex (core region). The component with molecular mass 3790.26 u lacked 1 Kdo residue and possessed in addition 1 Hep, 1 Hex and the first O-repeating unit, as compared to the non-substituted core. P, phosphate; HexN, hexosamine; 14 : 0(3-OH), 3-hydroxymyristic acid; PEtN, 2-aminoethanol phosphate; Kdo, 3-deoxy-D-*manno*-oct-2-ulosonic acid; Hep, L-*glycero*-D-*manno*-heptose; Hex, hexose. Δ m 698.27 u corresponded to the mass of one *O*-deacylated O-repeating unit.

outer core region (7-substituted L-α-D-Hep*p* and 6- and 2-substituted α-D-Glc*p*) could be resolved (data not shown), as compared with published data (Müller-Loennies *et al.*, 2003). Thus, the disaccharide *β*-D-Gal*p*NAc-(1→7)-L-α-D-Hep*p* was proven, which identified the site of attachment of the OPS at core oligosaccharide:

β-D-Qui3NAc-(1→3)-*α*-L-FucOAc-(1→4)-*β*-D-Gal-(1→3)-*α*-D-GalNAc-(1→4)-*β*-D-Qui3NAc-(1→3)-*α*-L-FucOAc-(1→4)-*β*-D-Gal-(1→3)-***β*-D-GalNAc-(1→7)-*α*-LD-Hep**

This is consistent with the data obtained for *E. coli* K-12/O16 MFF1, a mutant strain expressing LPS consisting

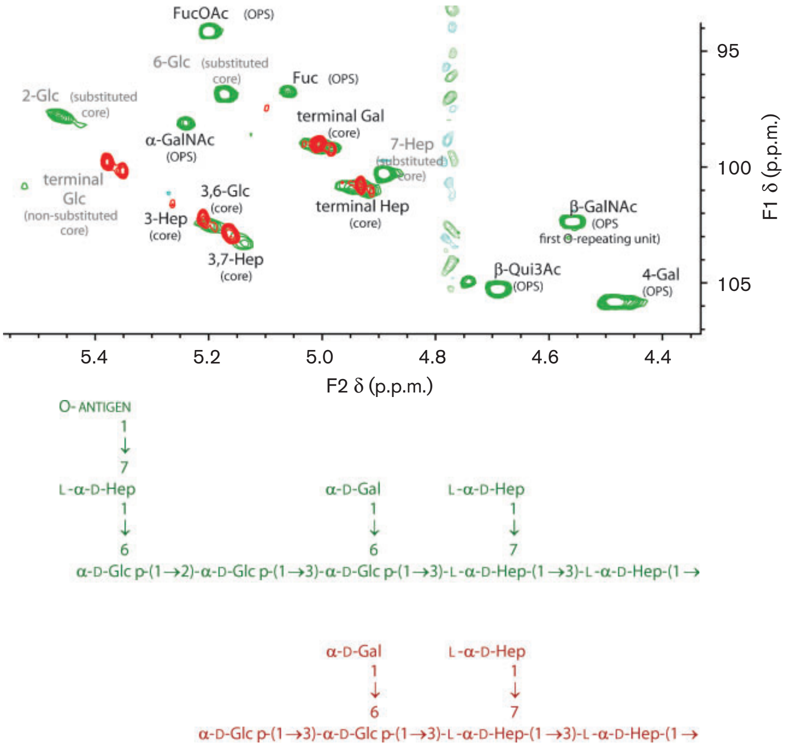
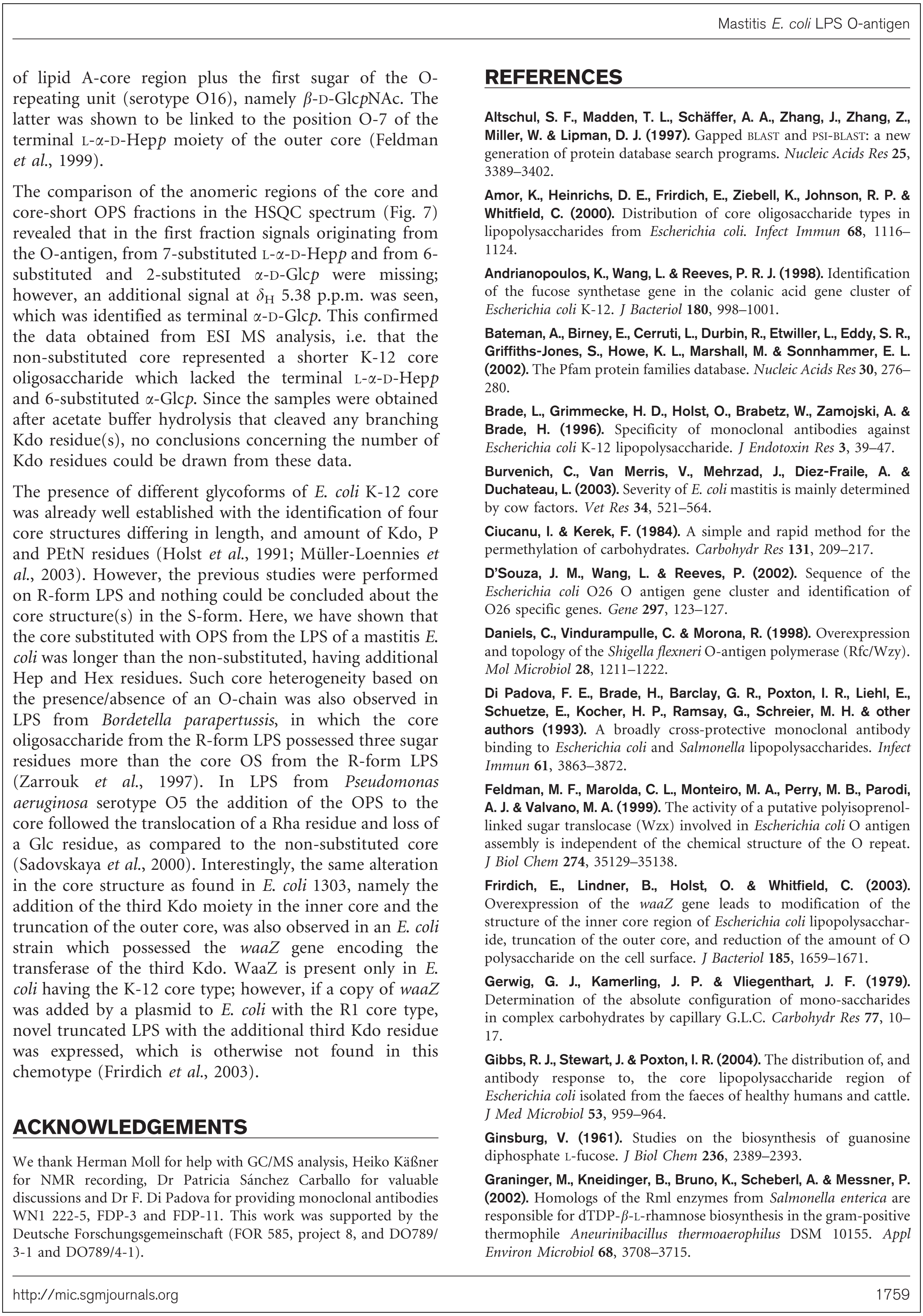


Fig. 7. Overlay of the anomeric regions of HSQC-DEPT spectra of the non-substituted core (red) and core-short OPS fraction (green) of *E. coli* 1303. The spectra were recorded at 700 MHz and 300 K. The signals differing between the two fractions are marked in grey. The structures of the non-substituted, truncated core (red) and substituted (green) are given below, drawn according to Müller-Loennies *et al.* (2003).



of lipid A-core region plus the first sugar of the O-repeating unit (serotype O16), namely *β*-D-Glc*p*NAc. The latter was shown to be linked to the position O-7 of the terminal L-α-D-Hep*p* moiety of the outer core (Feldman *et al.*, 1999).

The comparison of the anomeric regions of the core and core-short OPS fractions in the HSQC spectrum (Fig. 7) revealed that in the first fraction signals originating from the O-antigen, from 7-substituted L-α-D-Hep*p* and from 6-substituted and 2-substituted α-D-Glc*p* were missing; however, an additional signal at δ_H 5.38 p.p.m. was seen, which was identified as terminal α-D-Glc*p*. This confirmed the data obtained from ESI MS analysis, i.e. that the non-substituted core represented a shorter K-12 core oligosaccharide which lacked the terminal L-α-D-Hep*p* and 6-substituted α-Glc*p*. Since the samples were obtained after acetate buffer hydrolysis that cleaved any branching Kdo residue(s), no conclusions concerning the number of Kdo residues could be drawn from these data.

The presence of different glycoforms of *E. coli* K-12 core was already well established with the identification of four core structures differing in length, and amount of Kdo, P and PEtN residues (Holst *et al.*, 1991; Müller-Loennies *et al.*, 2003). However, the previous studies were performed on R-form LPS and nothing could be concluded about the core structure(s) in the S-form. Here, we have shown that the core substituted with OPS from the LPS of a mastitis *E. coli* was longer than the non-substituted, having additional Hep and Hex residues. Such core heterogeneity based on the presence/absence of an O-chain was also observed in LPS from *Bordetella parapertussis*, in which the core oligosaccharide from the R-form LPS possessed three sugar residues more than the core OS from the R-form LPS (Zarrouk *et al.*, 1997). In LPS from *Pseudomonas aeruginosa* serotype O5 the addition of the OPS to the core followed the translocation of a Rha residue and loss of a Glc residue, as compared to the non-substituted core (Sadovskaya *et al.*, 2000). Interestingly, the same alteration in the core structure as found in *E. coli* 1303, namely the addition of the third Kdo moiety in the inner core and the truncation of the outer core, was also observed in an *E. coli* strain which possessed the *waaZ* gene encoding the transferase of the third Kdo. WaaZ is present only in *E. coli* having the K-12 core type; however, if a copy of *waaZ* was added by a plasmid to *E. coli* with the R1 core type, novel truncated LPS with the additional third Kdo residue was expressed, which is otherwise not found in this chemotype (Frirdich *et al.*, 2003).

ACKNOWLEDGEMENTS

We thank Herman Moll for help with GC/MS analysis, Heiko Käßner for NMR recording, Dr Patricia Sánchez Carballo for valuable discussions and Dr F. Di Padova for providing monoclonal antibodies WN1 222-5, FDP-3 and FDP-11. This work was supported by the Deutsche Forschungsgemeinschaft (FOR 585, project 8, and DO789/3-1 and DO789/4-1).

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Amor, K., Heinrichs, D. E., Frirdich, E., Ziebell, K., Johnson, R. P. & Whitfield, C. (2000). Distribution of core oligosaccharide types in lipopolysaccharides from *Escherichia coli*. *Infect Immun* **68**, 1116–1124.

Andrianopoulos, K., Wang, L. & Reeves, P. R. J. (1998). Identification of the fucose synthetase gene in the colanic acid gene cluster of *Escherichia coli* K-12. *J Bacteriol* **180**, 998–1001.

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. & Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* **30**, 276–280.

Brade, L., Grimmecke, H. D., Holst, O., Brabetz, W., Zamojski, A. & Brade, H. (1996). Specificity of monoclonal antibodies against *Escherichia coli* K-12 lipopolysaccharide. *J Endotoxin Res* **3**, 39–47.

Burvenich, C., Van Merris, V., Mehrzad, J., Diez-Fraile, A. & Duchateau, L. (2003). Severity of *E. coli* mastitis is mainly determined by cow factors. *Vet Res* **34**, 521–564.

Ciucanu, I. & Kerek, F. (1984). A simple and rapid method for the permethylation of carbohydrates. *Carbohydr Res* **131**, 209–217.

D'Souza, J. M., Wang, L. & Reeves, P. (2002). Sequence of the *Escherichia coli* O26 O antigen gene cluster and identification of O26 specific genes. *Gene* **297**, 123–127.

Daniels, C., Vindurampulle, C. & Morona, R. (1998). Overexpression and topology of the *Shigella flexneri* O-antigen polymerase (Rfc/Wzy). *Mol Microbiol* **28**, 1211–1222.

Di Padova, F. E., Brade, H., Barclay, G. R., Poxton, I. R., Liehl, E., Schuetze, E., Kocher, H. P., Ramsay, G., Schreier, M. H. & other authors (1993). A broadly cross-protective monoclonal antibody binding to *Escherichia coli* and *Salmonella* lipopolysaccharides. *Infect Immun* **61**, 3863–3872.

Feldman, M. F., Marolda, C. L., Monteiro, M. A., Perry, M. B., Parodi, A. J. & Valvano, M. A. (1999). The activity of a putative polyisoprenol-linked sugar translocase (Wzx) involved in *Escherichia coli* O antigen assembly is independent of the chemical structure of the O repeat. *J Biol Chem* **274**, 35129–35138.

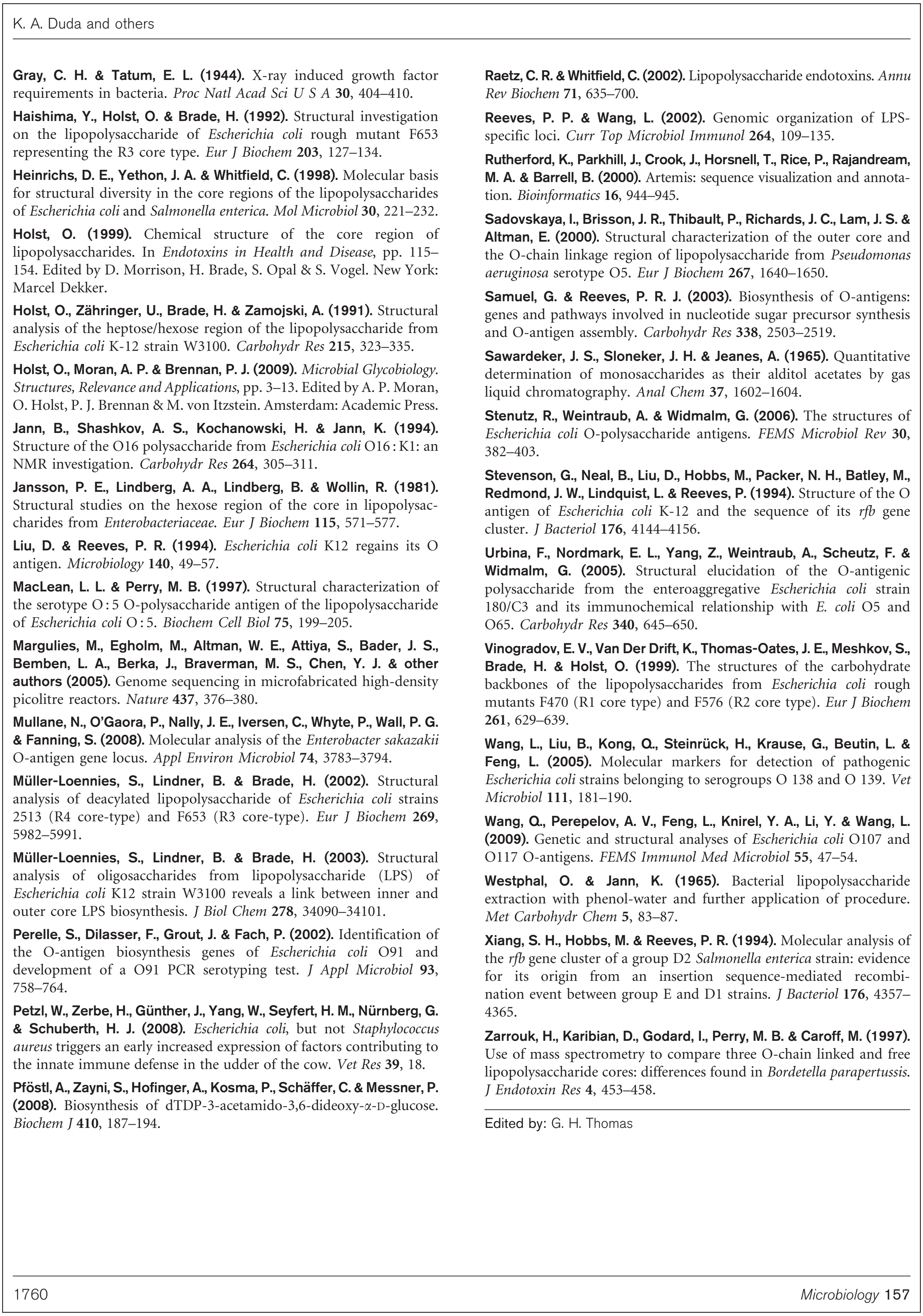
Frirdich, E., Lindner, B., Holst, O. & Whitfield, C. (2003). Overexpression of the *waaZ* gene leads to modification of the structure of the inner core region of *Escherichia coli* lipopolysaccharide, truncation of the outer core, and reduction of the amount of O polysaccharide on the cell surface. *J Bacteriol* **185**, 1659–1671.

Gerwig, G. J., Kamerling, J. P. & Vliegenthart, J. F. (1979). Determination of the absolute configuration of mono-saccharides in complex carbohydrates by capillary G.L.C. *Carbohydr Res* **77**, 10–17.

Gibbs, R. J., Stewart, J. & Poxton, I. R. (2004). The distribution of, and antibody response to, the core lipopolysaccharide region of *Escherichia coli* isolated from the faeces of healthy humans and cattle. *J Med Microbiol* **53**, 959–964.

Ginsburg, V. (1961). Studies on the biosynthesis of guanosine diphosphate L-fucose. *J Biol Chem* **236**, 2389–2393.

Graninger, M., Kneidinger, B., Bruno, K., Scheberl, A. & Messner, P. (2002). Homologs of the Rml enzymes from *Salmonella enterica* are responsible for dTDP-*β*-L-rhamnose biosynthesis in the gram-positive thermophile *Aneurinibacillus thermoaerophilus* DSM 10155. *Appl Environ Microbiol* **68**, 3708–3715.



Gray, C. H. & Tatum, E. L. (1944). X-ray induced growth factor requirements in bacteria. *Proc Natl Acad Sci U S A* **30**, 404–410.

Haishima, Y., Holst, O. & Brade, H. (1992). Structural investigation on the lipopolysaccharide of *Escherichia coli* rough mutant F653 representing the R3 core type. *Eur J Biochem* **203**, 127–134.

Heinrichs, D. E., Yethon, J. A. & Whitfield, C. (1998). Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*. *Mol Microbiol* **30**, 221–232.

Holst, O. (1999). Chemical structure of the core region of lipopolysaccharides. In *Endotoxins in Health and Disease*, pp. 115–154. Edited by D. Morrison, H. Brade, S. Opal & S. Vogel. New York: Marcel Dekker.

Holst, O., Zähringer, U., Brade, H. & Zamojski, A. (1991). Structural analysis of the heptose/hexose region of the lipopolysaccharide from *Escherichia coli* K-12 strain W3100. *Carbohydr Res* **215**, 323–335.

Holst, O., Moran, A. P. & Brennan, P. J. (2009). *Microbial Glycobiology. Structures, Relevance and Applications*, pp. 3–13. Edited by A. P. Moran, O. Holst, P. J. Brennan & M. von Itzstein. Amsterdam: Academic Press.

Jann, B., Shashkov, A. S., Kochanowski, H. & Jann, K. (1994). Structure of the O16 polysaccharide from *Escherichia coli* O16 : K1: an NMR investigation. *Carbohydr Res* **264**, 305–311.

Jansson, P. E., Lindberg, A. A., Lindberg, B. & Wollin, R. (1981). Structural studies on the hexose region of the core in lipopolysaccharides from *Enterobacteriaceae*. *Eur J Biochem* **115**, 571–577.

Liu, D. & Reeves, P. R. (1994). *Escherichia coli* K12 regains its O antigen. *Microbiology* **140**, 49–57.

MacLean, L. L. & Perry, M. B. (1997). Structural characterization of the serotype O : 5 O-polysaccharide antigen of the lipopolysaccharide of *Escherichia coli* O : 5. *Biochem Cell Biol* **75**, 199–205.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J. & other authors (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.

Mullane, N., O'Gaora, P., Nally, J. E., Iversen, C., Whyte, P., Wall, P. G. & Fanning, S. (2008). Molecular analysis of the *Enterobacter sakazakii* O-antigen gene locus. *Appl Environ Microbiol* **74**, 3783–3794.

Müller-Loennies, S., Lindner, B. & Brade, H. (2002). Structural analysis of deacylated lipopolysaccharide of *Escherichia coli* strains 2513 (R4 core-type) and F653 (R3 core-type). *Eur J Biochem* **269**, 5982–5991.

Müller-Loennies, S., Lindner, B. & Brade, H. (2003). Structural analysis of oligosaccharides from lipopolysaccharide (LPS) of *Escherichia coli* K12 strain W3100 reveals a link between inner and outer core LPS biosynthesis. *J Biol Chem* **278**, 34090–34101.

Perelle, S., Dilasser, F., Grout, J. & Fach, P. (2002). Identification of the O-antigen biosynthesis genes of *Escherichia coli* O91 and development of a O91 PCR serotyping test. *J Appl Microbiol* **93**, 758–764.

Petzl, W., Zerbe, H., Günther, J., Yang, W., Seyfert, H. M., Nürnberg, G. & Schuberth, H. J. (2008). *Escherichia coli*, but not *Staphylococcus aureus* triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow. *Vet Res* **39**, 18.

Pföstl, A., Zayni, S., Hofinger, A., Kosma, P., Schäffer, C. & Messner, P. (2008). Biosynthesis of dTDP-3-acetamido-3,6-dideoxy-α-D-glucose. *Biochem J* **410**, 187–194.

Raetz, C. R. & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annu Rev Biochem* **71**, 635–700.

Reeves, P. P. & Wang, L. (2002). Genomic organization of LPS-specific loci. *Curr Top Microbiol Immunol* **264**, 109–135.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**, 944–945.

Sadovskaya, I., Brisson, J. R., Thibault, P., Richards, J. C., Lam, J. S. & Altman, E. (2000). Structural characterization of the outer core and the O-chain linkage region of lipopolysaccharide from *Pseudomonas aeruginosa* serotype O5. *Eur J Biochem* **267**, 1640–1650.

Samuel, G. & Reeves, P. R. J. (2003). Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr Res* **338**, 2503–2519.

Sawardeker, J. S., Sloneker, J. H. & Jeanes, A. (1965). Quantitative determination of monosaccharides as their alditol acetates by gas liquid chromatography. *Anal Chem* **37**, 1602–1604.

Stenutz, R., Weintraub, A. & Widmalm, G. (2006). The structures of *Escherichia coli* O-polysaccharide antigens. *FEMS Microbiol Rev* **30**, 382–403.

Stevenson, G., Neal, B., Liu, D., Hobbs, M., Packer, N. H., Batley, M., Redmond, J. W., Lindquist, L. & Reeves, P. (1994). Structure of the O antigen of *Escherichia coli* K-12 and the sequence of its *rfb* gene cluster. *J Bacteriol* **176**, 4144–4156.

Urbina, F., Nordmark, E. L., Yang, Z., Weintraub, A., Scheutz, F. & Widmalm, G. (2005). Structural elucidation of the O-antigenic polysaccharide from the enteroaggregative *Escherichia coli* strain 180/C3 and its immunochemical relationship with *E. coli* O5 and O65. *Carbohydr Res* **340**, 645–650.

Vinogradov, E. V., Van Der Drift, K., Thomas-Oates, J. E., Meshkov, S., Brade, H. & Holst, O. (1999). The structures of the carbohydrate backbones of the lipopolysaccharides from *Escherichia coli* rough mutants F470 (R1 core type) and F576 (R2 core type). *Eur J Biochem* **261**, 629–639.

Wang, L., Liu, B., Kong, Q., Steinrück, H., Krause, G., Beutin, L. & Feng, L. (2005). Molecular markers for detection of pathogenic *Escherichia coli* strains belonging to serogroups O 138 and O 139. *Vet Microbiol* **111**, 181–190.

Wang, Q., Perepelov, A. V., Feng, L., Knirel, Y. A., Li, Y. & Wang, L. (2009). Genetic and structural analyses of *Escherichia coli* O107 and O117 O-antigens. *FEMS Immunol Med Microbiol* **55**, 47–54.

Westphal, O. & Jann, K. (1965). Bacterial lipopolysaccharide extraction with phenol-water and further application of procedure. *Met Carbohydr Chem* **5**, 83–87.

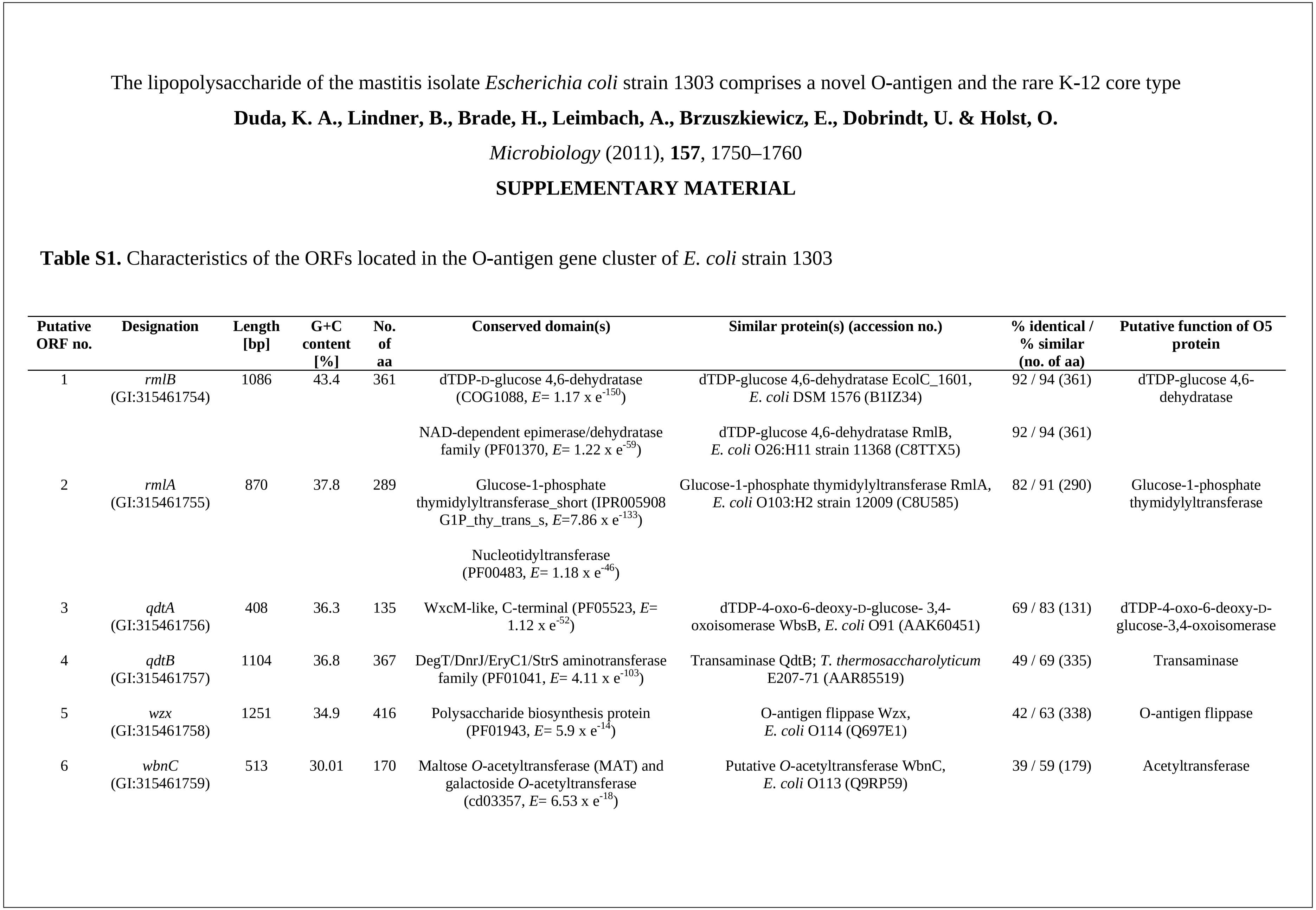
Xiang, S. H., Hobbs, M. & Reeves, P. R. (1994). Molecular analysis of the *rfb* gene cluster of a group D2 *Salmonella enterica* strain: evidence for its origin from an insertion sequence-mediated recombination event between group E and D1 strains. *J Bacteriol* **176**, 4357–4365.

Zarrouk, H., Karibian, D., Godard, I., Perry, M. B. & Caroff, M. (1997). Use of mass spectrometry to compare three O-chain linked and free lipopolysaccharide cores: differences found in *Bordetella parapertussis*. *J Endotoxin Res* **4**, 453–458.

Edited by: G. H. Thomas

5.3.1.3 *Supplementary information*

The supplementary Table S1 for Duda et al. (2011) is presented on pages 132–133. The original source can be found here: `http://mic.microbiologyresearch.org/content/journal/micro/10.1099/mic.0.046912-0`



The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type

Duda, K. A., Lindner, B., Brade, H., Leimbach, A., Brzuszkiewicz, E., Dobrindt, U. & Holst, O.

Microbiology (2011), **157**, 1750–1760

SUPPLEMENTARY MATERIAL

Table S1. Characteristics of the ORFs located in the O-antigen gene cluster of *E. coli* strain 1303

Putative ORF no.	Designation	Length [bp]	G+C content [%]	No. of aa	Conserved domain(s)	Similar protein(s) (accession no.)	% identical / % similar (no. of aa)	Putative function of O5 protein
1	*rmlB* (GI:315461754)	1086	43.4	361	dTDP-D-glucose 4,6-dehydratase (COG1088, *E*= 1.17 x e^{-150})	dTDP-glucose 4,6-dehydratase EcolC_1601, *E. coli* DSM 1576 (B1IZ34)	92 / 94 (361)	dTDP-glucose 4,6-dehydratase
					NAD-dependent epimerase/dehydratase family (PF01370, *E*= 1.22 x e^{-59})	dTDP-glucose 4,6-dehydratase RmlB, *E. coli* O26:H11 strain 11368 (C8TTX5)	92 / 94 (361)	
2	*rmlA* (GI:315461755)	870	37.8	289	Glucose-1-phosphate thymidylyltransferase_short (IPR005908 G1P_thy_trans_s, *E*=7.86 x e^{-133}) Nucleotidyltransferase (PF00483, *E*= 1.18 x e^{-46})	Glucose-1-phosphate thymidylyltransferase RmlA, *E. coli* O103:H2 strain 12009 (C8U585)	82 / 91 (290)	Glucose-1-phosphate thymidylyltransferase
3	*qdtA* (GI:315461756)	408	36.3	135	WxcM-like, C-terminal (PF05523, *E*= 1.12 x e^{-52})	dTDP-4-oxo-6-deoxy-D-glucose- 3,4-oxoisomerase WbsB, *E. coli* O91 (AAK60451)	69 / 83 (131)	dTDP-4-oxo-6-deoxy-D-glucose-3,4-oxoisomerase
4	*qdtB* (GI:315461757)	1104	36.8	367	DegT/DnrJ/EryC1/StrS aminotransferase family (PF01041, *E*= 4.11 x e^{-103})	Transaminase QdtB; *T. thermosaccharolyticum* E207-71 (AAR85519)	49 / 69 (335)	Transaminase
5	*wzx* (GI:315461758)	1251	34.9	416	Polysaccharide biosynthesis protein (PF01943, *E*= 5.9 x e^{-14})	O-antigen flippase Wzx, *E. coli* O114 (Q697E1)	42 / 63 (338)	O-antigen flippase
6	*wbnC* (GI:315461759)	513	30.01	170	Maltose *O*-acetyltransferase (MAT) and galactoside *O*-acetyltransferase (cd03357, *E*= 6.53 x e^{-18})	Putative *O*-acetyltransferase WbnC, *E. coli* O113 (Q9RP59)	39 / 59 (179)	Acetyltransferase

7	ORF 7 (GI:315461760)	921	30.8	306	Glycosyltransferase family 2 (PF00535, *E*= 1.12 x e^{-22})	Glycosyltransferase, *Enterobacter sakazakii* ESA_01184 (A9Y3E9)	41 / 61 (314)	Glycosyltransferase
8	*wzy* (GI:315461761)	1281	30.1	426	-	Putative O antigen polymerase, *Salmonella enterica* serovar Dakar (D1FXG0)	23/42 (431)	O-antigen polymerase
9	ORF 9 (GI:315461762)	1083	30.8	360	Glycosyltransferase_GTB_type super-family (cl10013, *E*= 2.09e^{-39}) Glycosyltransferases group 1 (PF00534, *E*= 1.03 x e^{-20})	Glycosyltransferase, group 1 WclF, *E. coli* O155 (AAV74551)	30 / 47 (291)	Glycosyltransferase
10	*wcaG* (GI:315461763)	924	30.8	307	WcaG nucleoside-diphosphate-sugar epimerase (COG0451, *E*=2.87 x e^{-33})	Putative UDP-glucose 4-epimerase EDWATA_01329, *Edwardsiella tarda* ATCC 23685 (D4F3L9)	42 / 61 (316)	Nucleoside-diphosphate-sugar epimerase
					NAD-dependent epimerase/dehydratase family (PF01370, *E*= 5.93 x e^{-29})	Bifunctional GDP-fucose synthetase: GDP-4-dehydro-6-deoxy-D-mannose epimerase/ GDP-4-dehydro-6-L-deoxygalactose reductase, *E. coli* MG1655 (NP_416556)	22 / 42 (325)	
11	ORF 11 (GI:315461764)	810	34.2	269	Glycosyltransferase family 2 (PF00535, *E*= 5.43 x e^{-18})	Glucosyltransferase WbeD, *Salmonella enterica* subsp. *enterica* serovar Pomona (B8R1W7)	60 / 75 (274)	Glycosyltransferase
12	*tnp* (GI:315461765)	1119	41.1	372	Transposase (COG5433, *E*= 1.07 x e^{-22}) Transposase DDE domain (PF01609, *E*= 5.68 x e^{-08})	Putative transposase YhhI, *E. coli* O8 strain IAI1 (B7M4P4)	91 / 93 (378)	Transposase
13	*qdtC* (GI:315461766)	483	32.7	160	WcxM-like, left-handed parallel beta-helix (LbH) N-terminal domain (cd03358, *E*= 2.80 x e^{-46}) Maltose *O*-acetyltransferase (MAT) and galactoside *O*-acetyltransferase (cd03357, *E*= 3.44 x e^{-22})	Transacetylase QdtC, *Salmonella choleraesuis* (D7PFB5)	76 / 88 (151)	Transacetylase

5.3.2 *Complete genome sequences of* Escherichia coli *strains 1303 and ECC-1470 isolated from bovine mastitis*

LEIMBACH A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R, Dobrindt U. 2015. Complete genome sequences of *Escherichia coli* strains 1303 and ECC-1470 isolated from bovine mastitis. *Genome Announc.* 3:e00182-15.
DOI: 10.1128/genomeA.00182-15

5.3.2.1 *Contributions*

Leimbach et al. (2015) presents the finished genomic sequences (Chain et al., 2009) of *E. coli* O70:H32 strain 1303, isolated from an acute case of mastitis, and *E. coli* Ont:Hnt ECC-1470, isolated from a persistent mammary infection. *Closing* and *finishing* bacterial genomic sequences has the advantage to resolve large structural variations, correct repeat construction, and does not miss gene calls in contrast to *draft* genomes (Section 1.1.3 on page 11) (Koren and Phillippy, 2015). Both genomes were accurately manually annotated to maximize the use in the research field. This is a data publication accompanying the submission to the public International Nucleotide Sequence Database Collaboration (INSDC) genomic databases, mostly describing the detailed methodology. Only minimal data analysis and functional inference is included. For comparative genomics with other *E. coli* bovine isolates, detailed analysis of VFs carried by these two genomes, and the implications for bovine mastitis pathogenesis see Leimbach et al. (2017) (Section 5.3.4 on page 140).

I designed the study and conducted all the research of the publication, with the exception of HTS, including sequence read QC and trimming, assembly and assembly QC, gap closure, sequence polishing, automatic annotation with manual curation, and submission to the INSDC public repositories. 690 PCRs and 1114 Sanger sequencing reactions were performed with 588 primers to close and polish the genome of *E. coli* 1303. The genome of *E. coli* ECC-1470 was finished with 257 PCRs and 668 Sanger sequencing reactions using 468 primers. Furthermore, I contributed scripts for the bioinformatical tasks in this publication (as described in Chapter 4 on page 73). I wrote all parts of the manuscript. Detailed individual author contributions for each part of the paper can be found in Table 16 on page 233.

5.3.2.2 *Main paper*

This open access publication can be found on pages 135–136 or freely available and to reuse (licensed under a Creative Commons Attribution 3.0 Unported License (CC BY 3.0) (cc) BY) at:
`http://genomea.asm.org/content/3/2/e00182-15`

genomeA
Journals.ASM.org

Complete Genome Sequences of *Escherichia coli* Strains 1303 and ECC-1470 Isolated from Bovine Mastitis

Andreas Leimbach,[a,b,c] Anja Poehlein,[b] Anika Witten,[d] Flemming Scheutz,[e] Ynte Schukken,[f,g] Rolf Daniel,[b] Ulrich Dobrindt[a,c]

Institute of Hygiene, University of Münster, Münster, Germany[a]; Department of Genomic and Applied Microbiology, Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University of Göttingen, Göttingen, Germany[b]; Institute for Molecular Infection Biology, Julius-Maximilians-University of Würzburg, Würzburg, Germany[c]; Institute for Human Genetics, University of Münster, Münster, Germany[d]; Statens Serum Institut, Copenhagen, Denmark[e]; Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, New York, USA[f]; GD Animal Health, Deventer, The Netherlands[g]

***Escherichia coli* is the leading causative agent of acute bovine mastitis. Here, we report the complete genome sequence of *E. coli* O70:H32 strain 1303, isolated from an acute case of bovine mastitis, and *E. coli* Ont:Hnt strain ECC-1470, isolated from a persistent infection.**

Received 9 February 2015 **Accepted** 13 February 2015 **Published** 26 March 2015

Citation Leimbach A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R, Dobrindt U. 2015. Complete genome sequences of *Escherichia coli* strains 1303 and ECC-1470 isolated from bovine mastitis. Genome Announc 3(2):e00182-15. doi:10.1128/genomeA.00182-15.

Address correspondence to Ulrich Dobrindt, dobrindt@uni-muenster.de.

The outcome and severity of *E. coli* intramammary infections were previously mainly associated with cow factors reacting to pathogen-associated molecular patterns rather than the genomic makeup of the infecting strain (1). Nevertheless, certain *E. coli* strains consistently cause an acute severe onset and others a mild chronic outcome (2, 3). Currently only the draft genome sequence of mastitis-associated *E. coli* O32:H37 strain P4 has been published (4).

E. coli 1303 was isolated from udder secretions of a cow with clinical mastitis (5) and *E. coli* ECC-1470 from a chronically infected cow (6). Both genomes were sequenced via whole-genome sequencing with the 454 FLX genome sequencer with GS20 chemistry (Roche Life Science, Mannheim, Germany) to a 27.8-fold or 13.4-fold coverage, respectively. Strain ECC-1470 was also sequenced with a 6-kb insert paired-end (PE) 454 sequencing library. Additionally, Nextera XT chemistry (Illumina, San Diego, CA, USA) for library preparation and a 101-bp PE sequencing run was used to sequence both strains on an Illumina HiScan SQ sequencer.

The 454 reads were *de novo* assembled with Newbler (v2.0.00.20 for 1303 and v2.3 for ECC-1470; Roche). The 454 and Illumina reads were *de novo* assembled using MIRA v3.4.0.1 (7). The hybrid assembly was combined with the initial Newbler assembly within the Gap4 software (v4.11.2) of the Staden package (8). Gaps were closed by primer walking via PCR and Sanger sequencing.

E. coli 1303 possesses a 4,948,797-bp and strain ECC-1470 a 4,803,751-bp chromosome. Each strain harbors an F-plasmid designated p1303_109 (108,501 bp) or pECC-1470_100 (100,061 bp), respectively. Additionally, strain 1303 contains a bacteriophage P1-like plasmid p1303_95 (94,959 bp) and a small cryptic plasmid p1303_5 (4,671 bp).

Annotation was done with Prokka v1.9 (9) and *E. coli* K-12 MG1655 (NC_000913.3) as reference. Annotations were manually curated by employing the Swiss-Prot, TrEMBL (10), IMG/ER (11), and Ecocyc databases (12). Open reading frame (ORF) finding was verified with YACOP v1 (13) and the reference strain MG1655's annotation using ACT v12.1.1 (14) for manual curation with Artemis v15.1.1 (15) and tbl2tab v0.1 (https://github.com/aleimba/bac-genomics-scripts/tree/master/tbl2tab). A total of 4,734 coding DNA sequences (CDS) were identified in *E. coli* 1303 with 22 rRNAs and 91 tRNAs (via tRNAscan-SE v1.3.1 [16]). The *E. coli* ECC-1470 genome includes 4,506 CDS with 22 rRNAs and 90 tRNAs.

By assigning multilocus sequence types (STs) using ecoli_mlst v0.3 (https://github.com/aleimba/bac-genomics-scripts/tree/master/ecoli_mlst) strains 1303 and ECC-1470 were allocated to phylogroups A (ST10) and B1 (ST847), respectively (17).

The most prominent virulence factors in both strains are the enterobactin siderophore, the group 4-capsule, and the *E. coli* type III secretion system 2. The genes *flu* (Ag43), *astA* (enteroaggregative *E. coli* heat-stable enterotoxin 1), *iss* (increased serum survival), an AMR-SSuT genomic island (antimicrobial resistance to streptomycin, sulfonamide, and tetracycline), and the second flagellar cluster, Flag-2, are only present in *E. coli* 1303. Putative virulence factors that are only present in strain ECC-1470 are two type VI secretion systems, the long polar fimbriae, Pix fimbriae, and the alternative flagellin Flk.

Nucleotide sequence accession numbers. The genome sequences have been deposited at DDBJ/ENA/GenBank under the accession numbers CP009166 to CP009169 (strain 1303) and CP010344 and CP010345 (strain ECC-1470).

ACKNOWLEDGMENTS

We thank I. Zude (Münster) for help with manual annotation and K. Gollnow (Göttingen) and B. Plaschke (Würzburg) for technical assistance.

This work was supported by the Deutsche Forschungsgemeinschaft (DO789/3-1 and DO789/4-1) and the Open Access Publishing Fund of the University of Münster.

REFERENCES

1. **Burvenich C, Van Merris V, Mehrzad J, Diez-Fraile A, Duchateau L.** 2003. Severity of *E. coli* mastitis is mainly determined by cow factors. Vet Res **34:**521–564. http://dx.doi.org/10.1051/vetres:2003023.
2. **Shpigel NY, Elazar S, Rosenshine I.** 2008. Mammary pathogenic *Escherichia coli*. Curr Opin Microbiol **11:**60–65. http://dx.doi.org/10.1016/j.mib.2008.01.004.
3. **Almeida RA, Dogan B, Klaessing S, Schukken YH, Oliver SP.** 2011. Intracellular fate of strains of *Escherichia coli* isolated from dairy cows with acute or chronic mastitis. Vet Res Commun **35:**89–101. http://dx.doi.org/10.1007/s11259-010-9455-5.
4. **Blum S, Sela N, Heller ED, Sela S, Leitner G.** 2012. Genome analysis of bovine-mastitis-associated *Escherichia coli* O32:H37 strain P4. J Bacteriol **194:**3732. http://dx.doi.org/10.1128/JB.00535-12.
5. **Petzl W, Zerbe H, Günther J, Yang W, Seyfert HM, Nürnberg G, Schuberth HJ.** 2008. *Escherichia coli*, but not *Staphylococcus aureus* triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow. Vet Res **39:**18. http://dx.doi.org/10.1051/vetres:2007057.
6. **Dogan B, Klaessig S, Rishniw M, Almeida RA, Oliver SP, Simpson K, Schukken YH.** 2006. Adherent and invasive *Escherichia coli* are associated with persistent bovine mastitis. Vet Microbiol **116:**270–282. http://dx.doi.org/10.1016/j.vetmic.2006.04.023.
7. **Chevreux B.** 2005. MIRA: an automated genome and EST assembler. Ph.D. thesis. The Ruprecht-Karls-University, Heidelberg, Germany.
8. **Staden R, Beal KF, Bonfield JK.** 2000. The Staden package, 1998. Methods Mol Biol **132:**115–130.
9. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153.
10. **UniProt Consortium.** 2014. Activities at the universal protein resource (UniProt). Nucleic Acids Res **42:**D191–D198. http://dx.doi.org/10.1093/nar/gkt1140.
11. **Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC.** 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics **25:**2271–2278. http://dx.doi.org/10.1093/bioinformatics/btp393.
12. **Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Schröder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD.** 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res **41:**D605–D612. http://dx.doi.org/10.1093/nar/gks1027.
13. **Tech M, Merkl R.** 2003. YACOP: enhanced gene prediction obtained by a combination of existing methods. In Silico Biol **3:**441–451.
14. **Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J.** 2005. ACT: the Artemis comparison tool. Bioinformatics **21:**3422–3423. http://dx.doi.org/10.1093/bioinformatics/bti553.
15. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.
16. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.
17. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol **60:**1136–1151. http://dx.doi.org/10.1111/j.1365-2958.2006.05172.x.

5.3.3 *Whole-genome draft sequences of six commensal fecal and six mastitis-associated* Escherichia coli *strains of bovine origin*

LEIMBACH A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, Zerbe H, Daniel R, Dobrindt U. 2016. Whole-genome draft sequences of six commensal fecal and six mastitis-associated *Escherichia coli* strains of bovine origin. *Genome Announc.* 4:e00753-16. DOI: 10.1128/genomeA.00753-16

5.3.3.1 *Contributions*

Leimbach et al. (2016) presents the "high-quality draft" (Chain et al., 2009) genomic sequences of six bovine mastitis and six bovine fecal commensal *E. coli* isolates. Annotation was adapted to the manually curated annotations of MAEC 1303 and ECC-1470 (Section 5.3.2 on page 134). As with Leimbach et al. (2015) in Section 5.3.2 on page 134, this paper is a data publication accompanying the submission of the twelve genomes to the INSDC databases. These genomes were analyzed in detail and used for phylogenetic and gene content comparisons in relation to the pathotype background (bovine mastitis or commensal *E. coli* isolates) in Leimbach et al. (2017) (Section 5.3.4 on page 140).

For Leimbach et al. (2016) I designed the study and conducted all the research. In detail, I performed all data collection, analysis and curation, as well as all bioinformatical tasks. HTS was performed by Anika Witten. Additionally, I contributed scripts to perform several of the bioinformatical activities (as described in Chapter 4 on page 73). I drafted the whole manuscript. Detailed individual author contributions for each part of the paper and Table 1 of the publication can be found in Table 17 and Table 18 on page 234, respectively.

5.3.3.2 *Main paper*

This open access publication can be found on pages 138–139 or freely available and to reuse (licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0) cc BY) at:
`http://genomea.asm.org/content/4/4/e00753-16`

AMERICAN SOCIETY FOR MICROBIOLOGY genomeAnnouncements

Whole-Genome Draft Sequences of Six Commensal Fecal and Six Mastitis-Associated *Escherichia coli* Strains of Bovine Origin

Andreas Leimbach,[a,b,c] Anja Poehlein,[b] Anika Witten,[d] Olga Wellnitz,[e] Nahum Shpigel,[f] Wolfram Petzl,[g] Holm Zerbe,[g] Rolf Daniel,[b] Ulrich Dobrindt[a,c]

Institute for Hygiene, University of Münster, Münster, Germany[a]; Department of Genomic and Applied Microbiology, Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University of Göttingen, Göttingen, Germany[b]; Institute for Molecular Infection Biology, Julius-Maximilians-University of Würzburg, Würzburg, Germany[c]; Institute for Human Genetics, University of Münster, Münster, Germany[d]; Veterinary Physiology, Vetsuisse Faculty University of Bern, Bern, Switzerland[e]; Koret School of Veterinary Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel[f]; Clinic for Ruminants, Ludwig-Maximilians-University of Munich, Oberschleißheim, Germany[g]

The bovine gastrointestinal tract is a natural reservoir for commensal and pathogenic *Escherichia coli* strains with the ability to cause mastitis. Here, we report the whole-genome sequences of six *E. coli* isolates from acute mastitis cases and six *E. coli* isolates from the feces of udder-healthy cows.

Received 6 June 2016 **Accepted** 10 June 2016 **Published** 28 July 2016

Citation Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, Zerbe H, Daniel R, Dobrindt U. 2016. Whole-genome draft sequences of six commensal fecal and six mastitis-associated *Escherichia coli* strains of bovine origin. Genome Announc 4(4):e00753-16. doi:10.1128/genomeA.00753-16.

Address correspondence to Ulrich Dobrindt, dobrindt@uni-muenster.de.

Although bovine intramammary infections with *Escherichia coli* mostly lead to an acute onset of mastitis, they can also result in a persistent infection of the udder with alternating subclinical or clinical periods (1). Additionally, no common virulence factor subset of mastitis-causing *E. coli* strains has been identified in previous studies (2).

To investigate the genomic potential of *E. coli* isolated from bovine mastitis, several draft genomes (3–5), as well as two complete genomes (6), have been published thus far. However, only two recent genomic *E. coli* mastitis studies included one commensal bovine isolate (7, 8). Because cows are a natural reservoir not only for pathogenic but also for commensal *E. coli* of high phylogenetic and genotypic diversity (2), we present here the draft genomes of six *E. coli* strains isolated from serous udder exudate of mastitis-afflicted cows and six *E. coli* strains isolated from the feces of udder-healthy cows (Table 1).

All genomes were sequenced with an Illumina HiScan SQ sequencer with Nextera XT chemistry (Illumina, San Diego, CA, USA) for library preparation and a 101-bp paired-end sequencing run. Raw reads were quality controlled with FastQC version 0.11.2 (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Low-quality reads and adapter contaminations were trimmed with Cutadapt version 1.6 (9). All reads were randomly subsampled to an approximate 70-fold coverage for each strain with seqtk version 1.0-r32 (https://github.com/lh3/seqtk). Subsequently, the reads were *de novo* assembled with SPAdes version 3.1.1 (10). Assembly statistics were evaluated with QUAST version 2.3 (11), resulting in 59 to 290 contigs >500 bp and genome sizes ranging from 4,765,494 to 5,459,392 bp (Table 1).

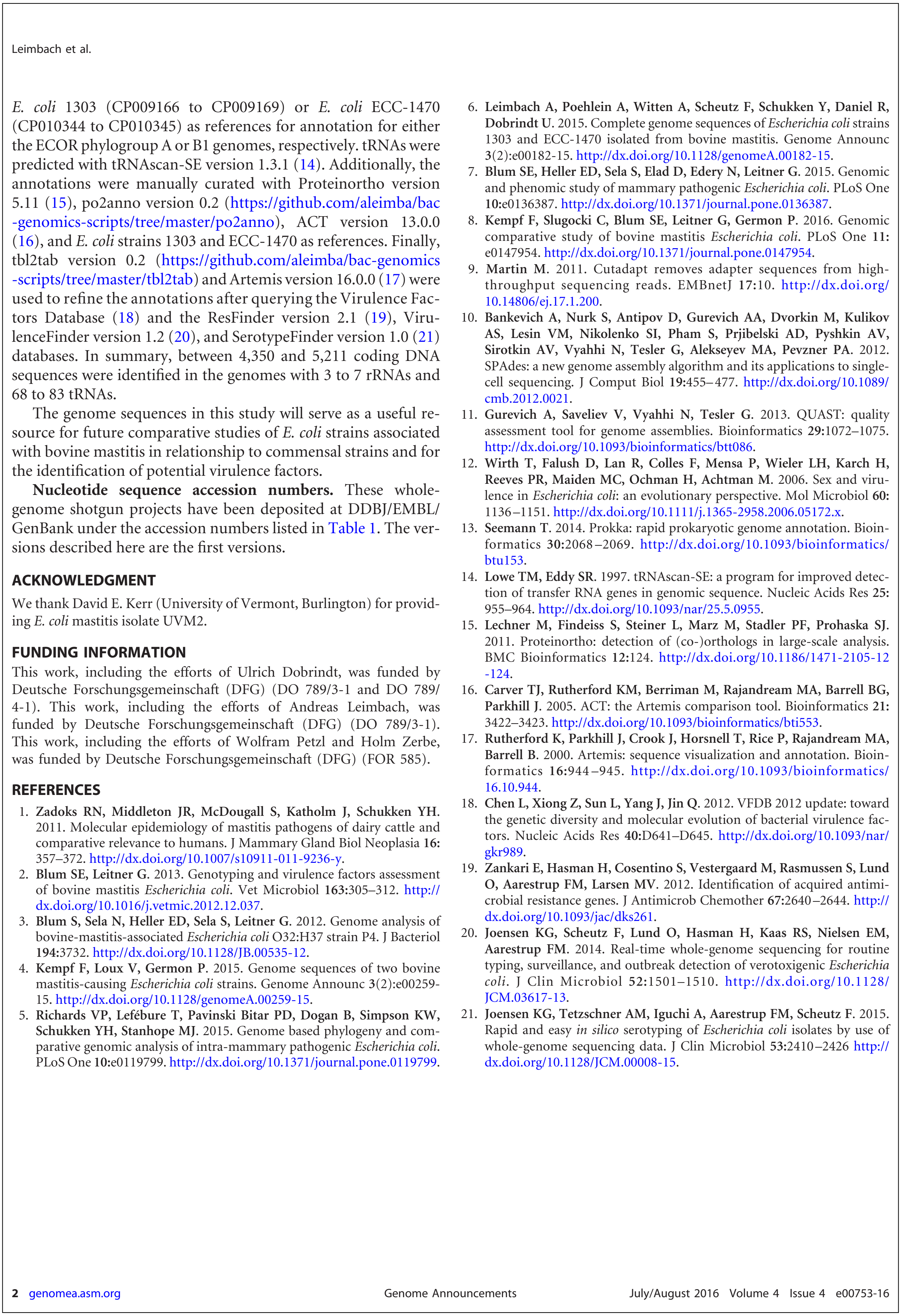
The strains were classified evenly into phylogroups A or B1, regardless of isolation source, through the assignment of sequence types (ST) with e. coli_mlst version 0.3 (https://github.com/aleimba/bac-genomics-scripts/tree/master/ecoli_mlst) (12). The most prominent sequence type is ST10, but most of the strains were not closely phylogenetically related.

All genomes were annotated with Prokka version 1.9 (13) with

TABLE 1 Genome features and assembly metrics of the 12 *E. coli* whole-genome sequences

Strain	ECOR phylogroup (ST)	Source of isolation	Genome size (bp)	No. of contigs >500 bp	N_{50} (bp)	No. of CDSs[a]	Accession no.
131/07	A (ST10)	Udder acute mastitis	5,459,392	270	79,414	5,123	JXUH00000000
2772a	B1 (ST156)	Udder acute mastitis	4,949,901	93	163,837	4,621	LCVG00000000
3234/A	A (ST10)	Udder acute mastitis	5,482,981	290	95,923	5,211	LCVH00000000
MPEC4839	A (ST10)	Udder acute mastitis	4,866,885	124	133,521	4,502	JYHP00000000
MPEC4969	B1 (ST1125)	Udder acute mastitis	4,833,611	130	103,834	4,468	JYHQ00000000
RiKo 2299/09	B1 (ST448)	Healthy cow feces	4,954,750	129	114,991	4,587	JYKB00000000
RiKo 2305/09	B1 (ST410)	Healthy cow feces	4,806,931	123	129,952	4,429	JYPB00000000
RiKo 2308/09	A (ST167)	Healthy cow feces	5,112,873	186	83,735	4,685	LCVI00000000
RiKo 2331/09	B1 (ST1614)	Healthy cow feces	4,765,494	59	224,192	4,350	LCVJ00000000
RiKo 2340/09	A (ST167)	Healthy cow feces	5,024,854	204	82,522	4,568	LAGW00000000
RiKo 2351/09	B1 (ST88)	Healthy cow feces	5,297,190	252	102,610	4,931	LAUC00000000
UVM2	A (ST1091)	Udder acute mastitis	4,926,170	149	86,033	4,614	LAUD00000000

[a] CDS, coding sequence.



E. coli 1303 (CP009166 to CP009169) or *E. coli* ECC-1470 (CP010344 to CP010345) as references for annotation for either the ECOR phylogroup A or B1 genomes, respectively. tRNAs were predicted with tRNAscan-SE version 1.3.1 (14). Additionally, the annotations were manually curated with Proteinortho version 5.11 (15), po2anno version 0.2 (https://github.com/aleimba/bac-genomics-scripts/tree/master/po2anno), ACT version 13.0.0 (16), and *E. coli* strains 1303 and ECC-1470 as references. Finally, tbl2tab version 0.2 (https://github.com/aleimba/bac-genomics-scripts/tree/master/tbl2tab) and Artemis version 16.0.0 (17) were used to refine the annotations after querying the Virulence Factors Database (18) and the ResFinder version 2.1 (19), VirulenceFinder version 1.2 (20), and SerotypeFinder version 1.0 (21) databases. In summary, between 4,350 and 5,211 coding DNA sequences were identified in the genomes with 3 to 7 rRNAs and 68 to 83 tRNAs.

The genome sequences in this study will serve as a useful resource for future comparative studies of *E. coli* strains associated with bovine mastitis in relationship to commensal strains and for the identification of potential virulence factors.

Nucleotide sequence accession numbers. These whole-genome shotgun projects have been deposited at DDBJ/EMBL/GenBank under the accession numbers listed in Table 1. The versions described here are the first versions.

ACKNOWLEDGMENT

We thank David E. Kerr (University of Vermont, Burlington) for providing *E. coli* mastitis isolate UVM2.

FUNDING INFORMATION

This work, including the efforts of Ulrich Dobrindt, was funded by Deutsche Forschungsgemeinschaft (DFG) (DO 789/3-1 and DO 789/4-1). This work, including the efforts of Andreas Leimbach, was funded by Deutsche Forschungsgemeinschaft (DFG) (DO 789/3-1). This work, including the efforts of Wolfram Petzl and Holm Zerbe, was funded by Deutsche Forschungsgemeinschaft (DFG) (FOR 585).

REFERENCES

1. **Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH.** 2011. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia **16:** 357–372. http://dx.doi.org/10.1007/s10911-011-9236-y.
2. **Blum SE, Leitner G.** 2013. Genotyping and virulence factors assessment of bovine mastitis *Escherichia coli*. Vet Microbiol **163:**305–312. http://dx.doi.org/10.1016/j.vetmic.2012.12.037.
3. **Blum S, Sela N, Heller ED, Sela S, Leitner G.** 2012. Genome analysis of bovine-mastitis-associated *Escherichia coli* O32:H37 strain P4. J Bacteriol **194:**3732. http://dx.doi.org/10.1128/JB.00535-12.
4. **Kempf F, Loux V, Germon P.** 2015. Genome sequences of two bovine mastitis-causing *Escherichia coli* strains. Genome Announc **3**(2):e00259-15. http://dx.doi.org/10.1128/genomeA.00259-15.
5. **Richards VP, Lefébure T, Pavinski Bitar PD, Dogan B, Simpson KW, Schukken YH, Stanhope MJ.** 2015. Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic *Escherichia coli*. PLoS One **10:**e0119799. http://dx.doi.org/10.1371/journal.pone.0119799.
6. **Leimbach A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R, Dobrindt U.** 2015. Complete genome sequences of *Escherichia coli* strains 1303 and ECC-1470 isolated from bovine mastitis. Genome Announc 3(2):e00182-15. http://dx.doi.org/10.1128/genomeA.00182-15.
7. **Blum SE, Heller ED, Sela S, Elad D, Edery N, Leitner G.** 2015. Genomic and phenomic study of mammary pathogenic *Escherichia coli*. PLoS One **10:**e0136387. http://dx.doi.org/10.1371/journal.pone.0136387.
8. **Kempf F, Slugocki C, Blum SE, Leitner G, Germon P.** 2016. Genomic comparative study of bovine mastitis *Escherichia coli*. PLoS One **11:** e0147954. http://dx.doi.org/10.1371/journal.pone.0147954.
9. **Martin M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetJ **17:**10. http://dx.doi.org/10.14806/ej.17.1.200.
10. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA.** 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol **19:**455–477. http://dx.doi.org/10.1089/cmb.2012.0021.
11. **Gurevich A, Saveliev V, Vyahhi N, Tesler G.** 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics **29:**1072–1075. http://dx.doi.org/10.1093/bioinformatics/btt086.
12. **Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MC, Ochman H, Achtman M.** 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol **60:** 1136–1151. http://dx.doi.org/10.1111/j.1365-2958.2006.05172.x.
13. **Seemann T.** 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics **30:**2068–2069. http://dx.doi.org/10.1093/bioinformatics/btu153.
14. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:** 955–964. http://dx.doi.org/10.1093/nar/25.5.0955.
15. **Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ.** 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics **12:**124. http://dx.doi.org/10.1186/1471-2105-12-124.
16. **Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J.** 2005. ACT: the Artemis comparison tool. Bioinformatics **21:** 3422–3423. http://dx.doi.org/10.1093/bioinformatics/bti553.
17. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.
18. **Chen L, Xiong Z, Sun L, Yang J, Jin Q.** 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res **40:**D641–D645. http://dx.doi.org/10.1093/nar/gkr989.
19. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV.** 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother **67:**2640–2644. http://dx.doi.org/10.1093/jac/dks261.
20. **Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM.** 2014. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol **52:**1501–1510. http://dx.doi.org/10.1128/JCM.03617-13.
21. **Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F.** 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. J Clin Microbiol **53:**2410–2426 http://dx.doi.org/10.1128/JCM.00008-15.

5.3.4 *No evidence for a bovine mastitis* Escherichia coli *pathotype*

> LEIMBACH A, Poehlein A, Vollmers J, Görlich D, Daniel R, Dobrindt U. 2017. No evidence for a bovine mastitis *Escherichia coli* pathotype. *BMC Genomics* 18:359.
> DOI: 10.1186/s12864-017-3739-x

This publication is also published on the bioRχiv preprint server in the "Genomics" subject area. The initial version of the manuscript *submitted* to BMC Genomics (`http://biorxiv.org/content/early/2016/12/23/096479`) and the *accepted* version of the manuscript (`http://biorxiv.org/content/early/2017/04/21/096479`) can be found at bioRχiv under DOI: 10.1101/096479.

5.3.4.1 *Contributions*

With the fourteen genomes from the two previous Genome Announcement publications, Leimbach et al. (2015) (Section 5.3.2 on page 134) and Leimbach et al. (2016) (Section 5.3.3 on page 137), plus additional eleven publicly available bovine *E. coli* reference genomes it was possible to investigate the phylogeny and gene content of mastitis and commensal isolates in sufficient detail. By comparing these results to the overall diversity of bovine *E. coli* isolates, Leimbach et al. (2017) provides evidence that no specific MPEC pathotype exists.

I designed the study, performed all parts of the research, prepared all figures, and drafted the manuscript. Scripts written for the study and used for the bioinformatical tasks are described in Chapter 4 on page 73. Anja Poehlein supported data curation for the INSDC databases, John Vollmers supported the gene content tree calculation, and Dennis Görlich implemented Fisher's exact test for statistical analyses in R. Rolf Daniel and Ulrich Dobrindt provided resources. Moreover, Ulrich Dobrindt acquired the funding, participated in project supervision, and was instrumental in reviewing the initial draft manuscript. The individual author contributions for each part of the paper and each dataset/figure/table can be found in Table 19 and Table 20 on page 235, respectively.

5.3.4.2 *Main paper*

This open access publication can be found on pages 141–162 or freely available and to reuse (licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0) cc BY) at:
`https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3739-x`

Leimbach *et al. BMC Genomics* (2017) 18:359
DOI 10.1186/s12864-017-3739-x

BMC Genomics

RESEARCH ARTICLE Open Access

No evidence for a bovine mastitis *Escherichia coli* pathotype

Andreas Leimbach[1,2,3*], Anja Poehlein[2], John Vollmers[4], Dennis Görlich[5], Rolf Daniel[2] and Ulrich Dobrindt[1,3*]

Abstract

Background: *Escherichia coli* bovine mastitis is a disease of significant economic importance in the dairy industry. Molecular characterization of mastitis-associated *E. coli* (MAEC) did not result in the identification of common traits. Nevertheless, a mammary pathogenic *E. coli* (MPEC) pathotype has been proposed suggesting virulence traits that differentiate MAEC from commensal *E. coli*. The present study was designed to investigate the MPEC pathotype hypothesis by comparing the genomes of MAEC and commensal bovine *E. coli*.

Results: We sequenced the genomes of eight *E. coli* isolated from bovine mastitis cases and six fecal commensal isolates from udder-healthy cows. We analyzed the phylogenetic history of bovine *E. coli* genomes by supplementing this strain panel with eleven bovine-associated *E. coli* from public databases. The majority of the isolates originate from phylogroups A and B1, but neither MAEC nor commensal strains could be unambiguously distinguished by phylogenetic lineage. The gene content of both MAEC and commensal strains is highly diverse and dominated by their phylogenetic background. Although individual strains carry some typical *E. coli* virulence-associated genes, no traits important for pathogenicity could be specifically attributed to MAEC. Instead, both commensal strains and MAEC have very few gene families enriched in either pathotype. Only the aerobactin siderophore gene cluster was enriched in commensal *E. coli* within our strain panel.

Conclusions: This is the first characterization of a phylogenetically diverse strain panel including several MAEC and commensal isolates. With our comparative genomics approach we could not confirm previous studies that argue for a positive selection of specific traits enabling MAEC to elicit bovine mastitis. Instead, MAEC are facultative and opportunistic pathogens recruited from the highly diverse bovine gastrointestinal microbiota. Virulence-associated genes implicated in mastitis are a by-product of commensalism with the primary function to enhance fitness in the bovine gastrointestinal tract. Therefore, we put the definition of the MPEC pathotype into question and suggest to designate corresponding isolates as MAEC.

Keywords: *E. coli*, Pathotype, Bovine mastitis, Commensals, Comparative genomics, Phylogeny, Virulence, Genomic diversity

Background

Bovine mastitis is a common disease in dairy cows with a global economic impact [1]. Mastitis is an inflammation of the cow udder mostly triggered by the invasion of pathogenic bacteria, leading to reduced milk production and quality. *Escherichia coli* is a major causative agent involved in acute bovine mastitis with a usually fast recovery rate. However, in extreme cases *E. coli* mastitis can lead to severe systemic clinical symptoms like sepsis concurrent with fever [2, 3]. Occasionally, an infection with *E. coli* results in a subclinical and persistent pathology [4, 5]. Traditionally, *E. coli* associated with intramammary infections are considered to be environmental opportunistic pathogens [6]. Thus, the outcome and severity of *E. coli* mastitis was mainly attributed to environmental factors and the innate immune response of the cow reacting to pathogen-associated molecular patterns (PAMPs) (most prominently lipopolysaccharide, LPS) rather than the virulence potential of the invading strain [7]. Intramammary infusion of purified LPS induces udder inflammation symptoms similar, yet not identical, to

* Correspondence: aleimba@gmx.de; dobrindt@uni-muenster.de
[1]Institute of Hygiene, University of Münster, Mendelstrasse 7, 48149 Münster, Germany
Full list of author information is available at the end of the article

E. coli invasion [7, 8]. The bovine gastrointestinal tract is a natural reservoir for commensal and pathogenic *E. coli* of high phylogenetic and genotypic diversity with the putative ability to cause mastitis [9]. Nevertheless, it was proposed that various genotypes of *E. coli* with specific phenotypes are better suited to elicit mastitis than others [3, 10, 11].

E. coli is a highly diverse species with commensal as well as pathogenic strains, which can colonize and persist in humans, animals, as well as abiotic environments [12, 13]. The population history of *E. coli* is largely clonal and can be structured into six major phylogenetic groups: A, B1, B2, D1, D2, and E [12, 14, 15], some publications also designate phylogroups D1 and D2 as D and F, respectively. These phylogroups have a different prevalence in various human and animal populations, but no host-restricted strains could be identified [12]. Pathogenic *E. coli* isolates are classified in different pathotypes according to the site of infection, clinical manifestation of the disease, and virulence factor (VF) repertoire. The group of intestinal pathogenic *E. coli* (IPEC) includes diarrheagenic pathotypes, which are obligate pathogens. The most prominent extraintestinal pathogenic *E. coli* (ExPEC) pathotypes are uropathogenic *E. coli* (UPEC), newborn meningitis-associated *E. coli* (MNEC), and avian pathogenic *E. coli* (APEC) [16–18]. In contrast to IPEC, which are traditionally considered to have a conserved VF repertoire, ExPEC are derived from different phylogenetic lineages and have variable VF content. Various combinations of VFs can lead to the same extraintestinal disease outcome, which solely defines an ExPEC pathotype [15, 16, 18]. However, many of these virulence-associated factors are also present in commensal strains and can be considered fitness factors (FFs), that enable or facilitate initial colonization and the establishment of an infection. These FFs have primarily evolved for gastrointestinal colonization as well as persistence, and the ability to cause extraintestinal disease is a coincidental by-product of commensalism. As a consequence, ExPEC are facultative pathogens that are recruited from the normal intestinal microbiota [12, 18, 19].

The broad spectrum of *E. coli* lifestyles and phenotypes is a result of the underlying genomic plasticity of *E. coli* strains [18]. Only up to 60% of each genome is shared by all isolates, the so-called core genome [20]. The remaining flexible genome is highly variable in individual strains. It includes genes for specific habitat adaptations or environmental conditions, and is the basis for the phenotypic diversity of *E. coli* [15, 18]. The flexible genome consists largely of mobile genetic elements (MGEs), including plasmids, genomic islands (GIs), and phages, which facilitate horizontal gene transfer (HGT) and are the driving forces for microbial diversity, evolution, and adaptation potential [21].

Despite the proposal of a mammary pathogenic *E. coli* (MPEC) pathotype [3] and extensive research, no common genetic traits or VFs have been identified for *E. coli* mastitis isolates, so far [11, 22–24]. Recently, several publications analyzed *E. coli* genomes from intramammary infections, thereby expanding the method spectrum by comparative genomics approaches [25–28]. All of these studies identified various MPEC genome regions and genes with different specificity criteria and significance, many of which are not considered to be classical VFs (or even encode for unknown hypothetical functions), but also genes coding for a type VI secretion system (T6SS), LPS biosynthesis, biofilm association, metabolic functions, and the ferric iron(III)-dicitrate (Fec) uptake system. However, the studies could mostly not agree upon a common set of putative VFs, except for the Fec siderophore system. Also, these studies suffer from small genome sample size constraints, lack of phylogenetic diversity, and/or did not include commensal bovine *E. coli* comparator strains of suitable phylogenetic and genotypic diversity. So far, depending on the study, no or only one bovine commensal *E. coli* isolate has been included in these corresponding analyses [25–28].

We wanted to advance upon the previous studies by analyzing a strain panel of phylogenetic and genomic diversity comparable to *E. coli* from the bovine habitat, especially by including fecal commensal isolates from udder-healthy cows. This enables our main goal, to characterize genetic traits which define mastitis-associated *E. coli* (MAEC) in comparison to non-pathogenic commensals, while keeping track of their phylogenetic background. Putative VFs important for bovine mastitis pathogenesis should be present in the majority of mastitis isolates, regardless of phylogroup, and mostly absent in commensals. We collected a large *E. coli* VF panel from different pathotypes for detailed candidate gene and gene composition analyses. By sequencing two MAEC genomes to closure, we made it possible to analyze MGEs and evaluate their role in HGT as well as virulence of MAEC and commensal isolates. Finally, several studies suggested that mastitis virulence might have evolved in separate *E. coli* lineages and phylogroups in parallel [10, 11, 26, 27], which might involve different virulence traits and strategies. Thus, we investigated the distribution of three putative phylogroup A MPEC-specific regions from Goldstone et al. [26] within the phylogroups of our strain panel for pathotype association.

Results

Bovine-associated *E. coli* are phylogenetically highly diverse and dominated by phylogroups A and B1

We compiled a strain panel of eight MAEC and six fecal commensal strains and supplemented it with the genomes from eleven reference strains from public databases (Table 1). The reference strains are composed of eight MAEC, two fecal commensal strains, and one milk commensal strain. Serotypes were predicted *in silico* (Table 1), but could not be determined unambiguously for several draft genomes. Nevertheless, none of the analyzed strains displayed identical serotypes (except non-typable MAEC strains 131/07 and 3234/A). Thus, a correlation between certain serotypes and MAEC was not detected.

The detected serotypes already indicated a high phylogenetic diversity in the strain panel. In order to obtain a more detailed view of the phylogenetic relationship of the strains, we calculated a core genome phylogeny based on a multiple whole genome nucleotide alignment (WGA) with 39 reference *E. coli* strains, four *Shigella* spp., and one *Escherichia fergusonii* strain as an outgroup. The filtered core genome WGA had a final alignment length of 2,272,130 bp, which is approximately 44% of the average *E. coli* genome size in the phylogeny (5,122,252 bp, Additional file 1 Table S1). The resulting *E. coli* population structure resolved the phylogenetic lineages described for *E. coli*, A, B1, E, D1, D2, and B2, with high bootstrap support values (Fig. 1) and is in consensus with earlier studies [12, 14]. The 25 *E. coli* genomes of the bovine-associated strain panel were mostly associated with phylogroups A and B1 (13 and 10, respectively; Table 1). Most of the MAEC (11/16, 69%) belong to phylogroup A and the majority of commensal strains to group B1 (6/9, 67%). MAEC D6-113.11 and commensal AA86 are the exception to the rule by being associated with phylogroups E and B2, respectively. All phylogenetic group affiliations of the included reference strains were in accordance to their source publications (Table 1).

To enhance backwards compatibility, we determined the sequence types (ST) for all strains analyzed in the WGA phylogeny according to the Achtman *E. coli* multilocus sequence typing (MLST) scheme (Additional file 2: Table S2) [13]. The calculated minimum spanning tree (MST) supports the phylogenetic history depicted in the

Table 1 Characteristics of the bovine-associated *E. coli* strain panel

Strain	Pathotype	Phylogroup (ST, CC)	Serotype	No. of CDS	Contigs	Reference
1303	MAEC	A (10, 10)	O70:H32	4734	finished	This study, [45]
131/07	MAEC	A (10, 10)	Ont:H39	5123	270	This study, [68]
2772a	MAEC	B1 (156, 156)	O174:H28	4621	93	This study, [68]
3234/A	MAEC	A (10, 10)	Ont:H39	5211	290	This study, [68]
AA86	fecal commensal	B2 (91, 1876)	O39:H4	4627	5	[46]
D6-113.11	MAEC	E (4175, 4175)	O80:H45	4750	89	[27]
D6-117.07	MAEC	A (10, 10)	O45:H11	4477	51	[27]
D6-117.29	MAEC	A (10, 10)	O28ac/O42:H37	4732	980	Direct submission
ECA-727	MAEC	A (10, 10)	O99:H9	4779	539	[28]
ECA-O157	MAEC	A (398, 398)	O29:H27	4434	1173	[28]
ECC-1470	MAEC	B1 (847, 847)	Ont:H2	4506	finished	This study, [45]
ECC-Z	MAEC	A (10, 10)	O74:H39	4600	24	[28]
MPEC4839	MAEC	A (10, 10)	O105:H32	4502	124	This study, [68]
MPEC4969	MAEC	B1 (1125, 161)	O139:H19	4468	130	This study, [68]
O157:H43 T22	milk commensal	B1 (155, 58)	O157:H43	4792	64	[48]
O32:H37 P4	MAEC	A (10, 10)	O32:H37	4581	72	[25]
P4-NR	MAEC	B1 (602, 446)	O15:H21/H54	4569	107	Direct submission
RiKo 2299/09	fecal commensal	B1 (448, 448)	O8/O160:H8	4587	129	This study, [68]
RiKo 2305/09	fecal commensal	B1 (410, 88)	O8:H21	4429	123	This study, [68]
RiKo 2308/09	fecal commensal	A (167, 10)	O9a/O89:H9	4685	186	This study, [68]
RiKo 2331/09	fecal commensal	B1 (1614, NA)	Ont:H23	4350	59	This study, [68]
RiKo 2340/09	fecal commensal	A (167, 10)	O89:H9	4568	204	This study, [68]
RiKo 2351/09	fecal commensal	B1 (88, 88)	O21:H4	4931	252	This study, [68]
UVM2	MAEC	A (1091, 10)	O53:H10	4614	149	This study, [68]
W26	fecal commensal	B1 (1081, 533)	O45:H14	4865	165	[107]

Strains sequenced in this study are highlighted in bold. Finished sequencing standard for complete genomes according to Chain et al. [44]. *CDS* coding sequences, *ST* sequence type, *CC* clonal complex

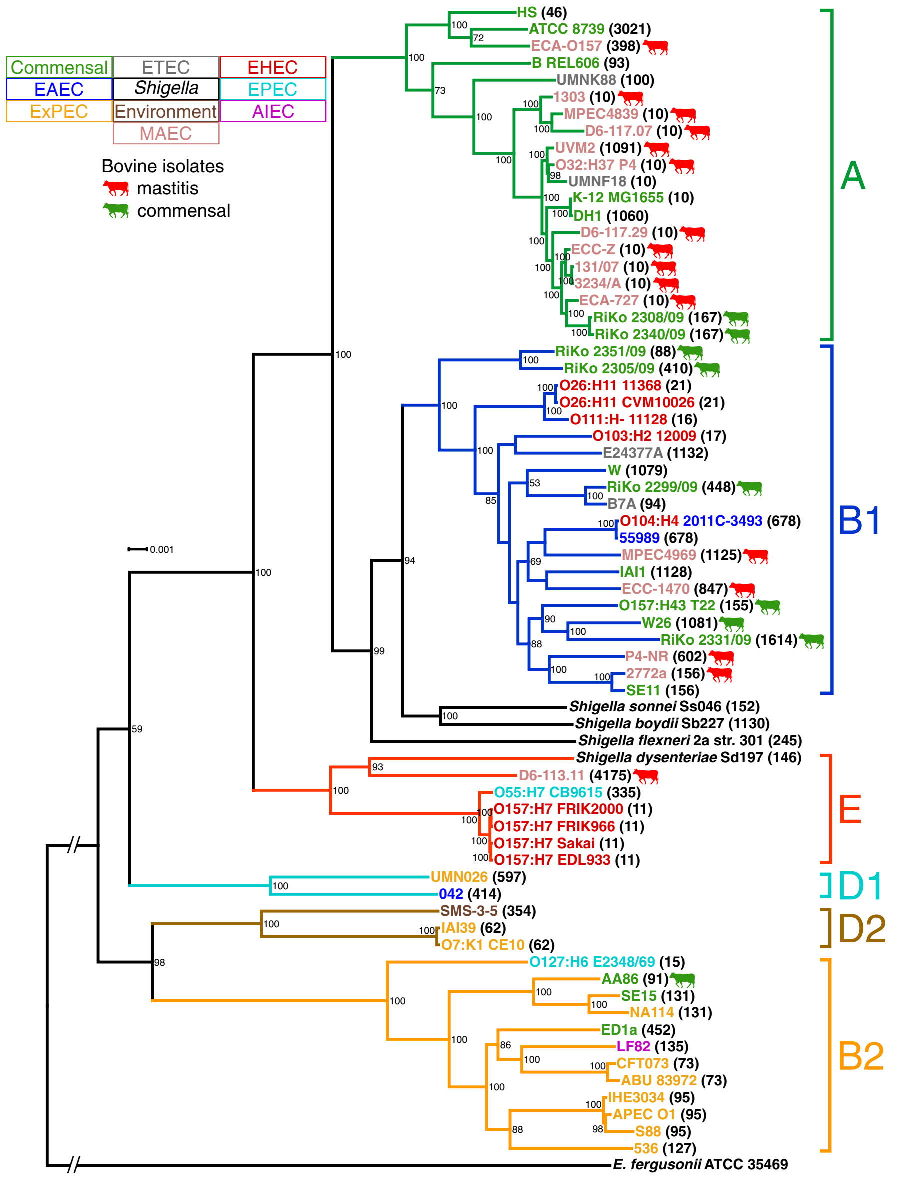


Fig. 1 Whole genome alignment phylogeny of bovine-associated and reference *E. coli* strains. The phylogeny is based on a whole core genome alignment of 2,272,130 bp. The best scoring maximum likelihood (ML) tree was inferred with RAxML's GTRGAMMA model with 1000 bootstrap resamplings. The tree was visualized with Dendroscope and bootstrap values below 50 removed. *E. fergusonii* serves as an outgroup and the corresponding branch is not to scale. Bovine-associated *E. coli* are indicated by colored cows, and both *E. coli* pathotypes and phylogroups are designated with a color code. ST numbers from the MLST analysis for each strain are given in parentheses. *E. coli* isolated from cows are widely distributed in the phylogroups and both commensal and MAEC strains are interspersed in the phylogenetic groups with a polyphyletic history

WGA phylogram and confirms the diversity of bovine-associated *E. coli* (Additional file 3: Figure S1). ST10, with nine occurrences, is the most common ST in the 25 *E. coli* genomes from the bovine-associated strain panel. In fact, all bovine-associated *E. coli* of phylogroup A are members of clonal complex 10 (CC10), except for *E. coli* ECA-O157 (ST398, CC398). Nevertheless, the majority of the 25 *E. coli* genomes have different STs, corroborating their phylogenetic diversity.

Gene content correlates with phylogenetic lineages of bovine-associated *E. coli*

Despite the phylogenetic diversity of the bovine-associated *E. coli*, we were interested to see if functional convergence of bovine MAEC or commensals exists. There might be a defining subset of genes or VFs for MAEC from different phylogenetic backgrounds that would point to a putative MPEC pathotype. For this purpose we determined the similarity of the genomes based on the presence/absence of all orthologous groups (OG) calculated for the strain panel. Such an analysis, visualized as a so-called gene content tree, has the advantage of considering the core as well as the flexible genome, in contrast to the WGA core genome phylogeny (in which the flexible genome is intentionally filtered out in order to maximize the robustness of the inferred phylogenetic history). Thus, this method can be used to detect functional similarities based on similar gene content. We clustered all strains based on gene content by calculating the best scoring maximum likelihood (ML) tree of the binary matrix representing the presence and absence of OGs (Additional file 4: Dataset S1). The topology of the resulting gene content tree mirrors the phylogenetic lineages of the WGA phylogeny with high analogy (Fig. 2). All bifurcations that define phylogroups in the gene content tree have high bootstrap values. For comparison purposes we visualized the high similarity between WGA genealogy and gene content tree in a tanglegram (Additional file 3: Figure S2A and B). This diagram shows that not only the phylogroups are conserved, but also the phylogenetic relationships between individual *E. coli* isolates within the phylogroups. However, some minor differences in the bifurcations between phylogeny and gene content clustering were detected. The two biggest differences concern the placement of phylogroups B2/E and MAEC ECA-O157. In contrast to the WGA-based phylogeny, which clusters phylogroups B2 and E outside the A/B1 sister taxa, the gene content dendrogram places these phylogroups closer to B1 than A (Fig. 2 and Additional file 3: Figure S2B). This appears to be due to a more similar gene content, as phylogroups B2/E have a higher recombination frequency with phylogroup B1 than with A [29, 30]. Strain ECA-O157 represents an outlier branch in comparison to all other included *E. coli* genomes based on gene content (Fig. 2). As this strain is the only strain in phylogroup A that does not belong to the closely related CC10 cluster, this explains its gene content divergence to the other A strains in the gene content tree, which is also apparent in the WGA core genome phylogeny. However, the outlier-clustering of ECA-O157 might also be a result of the high fragmentation of the draft genome (nearly 1000 contigs > 500 bp, Additional file 5: Table S3) and the resulting uncertain accuracy of CDS (coding DNA sequence) predictions on which OG analyses are dependent.

In conclusion, no functional convergence of bovine MAEC or commensals could be detected and the phylogenetic diversity of the strains is also apparent in a highly diverse gene content.

MAEC possess no virulence-attributed orthologs in comparison to commensal strains

Since no large scale gain or loss of bovine MAEC- or commensal-associated genes could be detected in the gene content tree, we looked into the distribution of OGs in more detail, in order to search for genotypic traits enriched in bovine mastitis or commensal isolates.

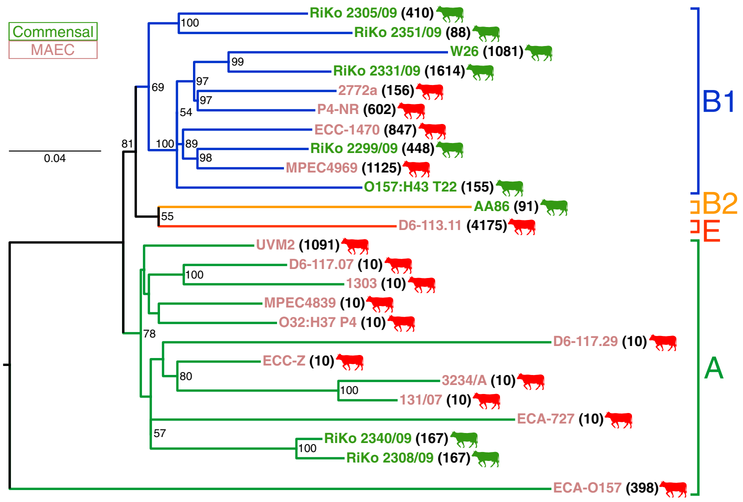


Fig. 2 Gene content clustering tree of the bovine-associated *E. coli*. The gene content best scoring ML dendrogram is based upon the presence or absence of orthologous groups (OGs) with 1000 resamplings for bootstrap support values. The tree was visualized midpoint rooted with FigTree and bootstrap values below 50 removed. The distance between the genomes is proportional to the OGs present or absent. The tree topology of the gene content tree follows closely the core genome WGA phylogeny. There is no functional convergence between MAEC or commensal strains, rather a highly diverse gene content

From our point of view, only the comparison of a larger set of MAEC genome sequences with that of bovine commensals is suitable to address this question. If any VFs/FFs existed, that play an important role in the pathogenesis of MAEC, we would expect a wide distribution of the encoding genes among MAEC strains compared to commensals.

The pan-genome of the 25 bovine-associated *E. coli* strains amounted to 116,535 CDS and a total of 13,481 OGs using BLASTP+ with 70% identity and coverage cutoffs. Because of the open nature of the *E. coli* pan-genome [31], all genomes included OGs, which were absent in every other compared strain (so-called singletons; Additional file 6: Dataset S2). The largest numbers of singletons were detected in the highly fragmented genomes of strains D6-117.29 ($n = 455$), ECA-O157 ($n = 865$), and ECA-727 ($n = 615$), a likely consequence of the high number of contig ends and uncertain open reading frame (ORF) predictions (Additional file 5: Table S3). Also, large numbers of singletons in genomes AA86 ($n = 422$) and D6-113.11 ($n = 361$) are to be expected, as these are the only compared genomes of their respective phylogroups, B2 and E. The majority of singletons encode typical proteins of the flexible genome, like hypothetical proteins, proteins associated with MGEs (transposases, phages), restriction modification systems, O-antigen biosynthesis, CRISPR, conjugal transfer systems, and sugar transport/utilization operons. Although several of these genes and gene functions have previously been identified as MAEC-associated in small strain panels [25, 27], they most likely play no role in mastitis because of their presence in commensals and/or low prevalence in MAEC.

To determine OGs which are characteristic of mastitis-associated or bovine commensal isolates, we screened the 13,481 OGs of the bovine-associated *E. coli* pan-genome for OGs which are significantly ($p < 0.05$) associated with one of these two groups of strains, using Fisher's exact test. 240 OGs displayed a significant association with a pathotype. However, none of these OGs remained significantly associated with either mastitis or commensal isolates when a Bonferroni correction was applied (Fig. 3a, Additional file 3: Figure S3A). Furthermore, none of these OGs were exclusively present in all mastitis, but absent from all commensal isolates tested and vice versa. In order to identify OGs with a wide distribution in one pathotype in comparison to the other, we looked for OGs present in at least 70% of the genomes of one pathotype and maximally in 30% of the other. This resulted in 36 "MAEC-" and 48 "commensal-enriched" OGs, most of which displayed a significant association (Fig. 3a and Additional file 7: Dataset S3).

Because phylogeny was shown to exhibit a strong effect on the gene content of *E. coli* isolates and shared ancestry might overshadow functional relatedness, we tested the 13,481 OGs additionally for a significant association with the phylogroups A or B1. This resulted in 410 significantly associated OGs. After Bonferroni correction, six OGs remained significantly associated to phylogroup A, whereas 14 OGs remained significantly phylogroup B1-associated (Additional file 3: Figure S3B). We used the same inclusion and exclusion cutoffs to identify OGs that were enriched in genomes of the four phylogroups (A, B1, B2, and E; Fig. 3b and Additional file 8: Dataset S4). This analysis resulted in many phylogroup-

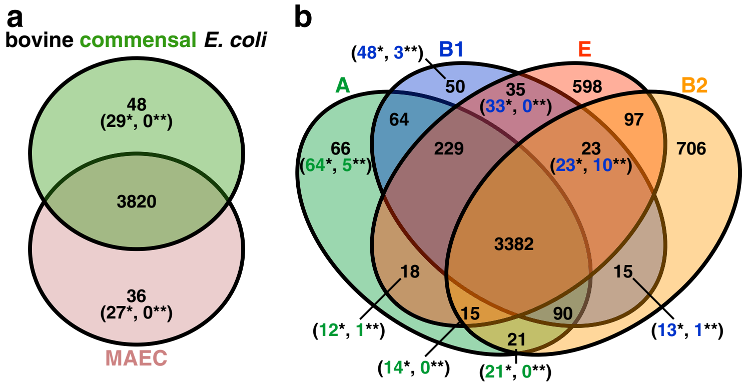


Fig. 3 Venn diagrams for gene family enrichment in pathotypes or phylogroups. **a** Enrichment of OGs in pathotypes (MAEC or commensal) was determined statistically (numbers in parentheses; Fisher's exact test, $p < 0.05$) after applying 70% inclusion and 30% exclusion group cutoffs (numbers without parentheses). Numbers with a *single asterisk* correspond to OGs with a statistically significant association while *two asterisks* indicate remaining significant associations after a Bonferroni correction. **b** Enrichment of OGs in phylogenetic groups (A, B1, B2, or E) was determined based on 70% inclusion and 30% exclusion group cutoffs. Statistic testing for OG association (Fisher's exact test, $p < 0.05$) was performed only for the multi-genome phylogroups A versus B1. Only very few OGs could be detected as pathotype enriched. Instead, OG distribution is strongly affected by phylogenetic background

enriched OGs, supporting the impact of phylogeny on gene content and the similarity between the gene content tree and WGA phylogeny. An "all-strain" soft core genome (as defined by Kaas et al. [20]) with this 70% inclusion cutoff included 3842 OGs, which is about 82% of the average number of CDS in the genomes (Additional file 9: Table S4).

Commensal-enriched orthologous groups are associated with fitness factors

Of the 29 significant commensal-enriched OGs, eight OGs are not simultaneously enriched in a phylogroup (Additional file 7: Dataset S3). These include the aerobactin siderophore biosynthesis operon (*iucABCD*) with the siderophore receptor-encoding (*iutA*) and the associated putative transport protein ShiF-encoding genes (locus tags in strain RiKo 2340/09 RIKO2340_186c00010 to RIKO2340_186c00060) as well as two OGs coding for IS element (insertion sequence)-related proteins (paralogs RIKO2340_128c00050/RIKO2340_203c00010 and RIKO2 340_203c00020). 20 of the 21 remaining commensal-enriched OGs are significantly associated with phylogroup B1 and include fimbrial genes, genes of the galactitol (*gatZCR*) phosphotransferase systems (PTS) as well as genes for sucrose catabolism (*cscKA*), a putative ABC transporter, the ChpB-ChpS toxin-antitoxin system, and a lipoprotein. Their role for bovine commensalism remains unclear, especially because of their additional association with phylogroup B1.

MAEC-enriched orthologous groups are mostly associated with mobile genetic elements

Thirteen of the 27 significantly mastitis-associated OGs are significantly enriched in phylogroup A, two are present in the soft core genome, and two are absent in the phylogroup B2 (commensal AA86). The remaining ten mastitis-associated OGs, which do not display phylogenetic or core enrichment, do not include any coherent gene cluster (Additional file 7: Dataset S3). However, eight of them are located in close proximity to each other in the genome of strain 1303 (*rzpQ* EC1303_c16730, *ydfR* EC1303_c16750, *quuQ_1* EC1303_c16790 (paralog to *quuQ*_2 in 1303 prophage 4), *relE* EC1303_c16830, *relB* EC1303_c16840, *flxA* EC1303_c16860, putative integrase EC1303_c16890, and hypothetical protein EC1303_c16900). All of these proteins belong to a prophage without noticeable features (see *E. coli* 1303 prophage 2 below). Additionally, three genes included in the soft core and enriched in phylogroups A/B1/E, *cspI*, *cspB* ("cold shock proteins", EC1303_c16710 and EC1303_c16770, respectively), and *recE* ("exonuclease VIII, 5′ - > 3′ specific dsDNA exonuclease", EC1303_c16970) also lie within the same prophage region. Because the *E. coli* 1303 prophage 2 genome does not contain genes related to metabolic or virulence functions, the role of the respective encoded gene products in mastitis cannot be determined. The last two OGs without phylogroup or core enrichment, *ylbG* (E1470_c05180) and *ybbC* (EC1303_c04920), encode for a putative DNA-binding transcriptional regulator and a putative immunity protein, respectively, and are associated with an *rhs* element. The 13 OGs that are also significantly enriched in phylogroup A encode for a transcriptional regulator (*rmhR* EC1303_c24270), an alpha amylase (EC13107_63c00240), a toxin/antitoxin system (*yafNO*, EC1303_c02750 and EC1303_c02760), a lipoprotein (*ybfP* EC1303_c06580), a phsohpodiesterase (*yaeI* EC1303_c01600), a malonyl CoA-acyl carrier protein transacylase (*ymdE* EC1303_c10470), a transposase (*insL*1_2 gene EC1303_c28750), and hypothetical proteins. According to the sequence contexts in these strains, the genes cannot be unambiguously localized in prophage regions or typical pathogenicity islands. Additionally, *eprI* (EC1303_c29770) encodes a type III secretion apparatus inner ring protein and is associated with a pathogenicity island (PAI). Finally, two genes contained in MAEC 1303 prophage 1, encoding for an exonuclease (EC1303_c12230) and an envelope protein (EC1303_c12530), are also associated with phylogroups A/B1/E.

In summary, the putative mastitis-eliciting function of any of the genes within the significantly MAEC-associated OGs is unclear. A truly meaningful correlation between OGs and pathotypes (mastitis vs. commensal) could not be observed. Instead, several OGs are significantly associated with phylogroups A or B1. No traditional *E. coli* VFs have been found among MAEC-enriched OGs.

Genomic islands and prophages in MAEC 1303 and ECC-1470 contain only few well-known virulence-associated genes

Both finished *E. coli* 1303 and ECC-1470 genomes include several putative pathogenicity, resistance, and metabolic islands, as well as prophages (Additional file 10: Dataset S5 and Additional file 11: Dataset S6). GIs could only be detected in the chromosomes of the closed genomes, but not on the respective plasmids. However, on the F plasmid present in *E. coli* 1303, p1303_109, a smaller 17-kb transposable element was identified. Mastitis isolate 1303 additionally harbors an episomal circularized P1 bacteriophage [32], designated p1303_95.

Generally, the genome of mastitis isolate 1303 includes twelve GIs ranging in size from 11 to 88 kb and encoding from 11 to 81 CDSs (Additional file 10: Dataset S5). One large composite GI (GI4) combines pathogenicity- and resistance-related genes. It partly contains the biofilm-associated polysaccharide synthesis *pga* locus.

The resistance-related genes of GI4 are located on the AMR-SSuT (antimicrobial multidrug resistance to streptomycin, sulfonamide, and tetracycline) island, which is prevalent in *E. coli* from the bovine habitat [33, 34]. The encoded resistance genes are *strAB*, *sul2*, and *tetDCBR*. A comparison of the corresponding genomic region of *E. coli* 1303 with two publicly available AMR-SSuT island sequences is shown in Additional file 3: Figure S4. Transposon Tn*10*, also present on the resistance plasmid R100, is an integral part of the AMR-SSuT island and comprises the *tetDCBR* genes. This highlights the composite nature of the AMR-SSuT island and of GI4 in general. The resistance markers of AMR-SSuT are prevalent, as seven strains of the panel contain some or all of the genes (D6-117.29, ECA-727, RiKo 2305/09, RiKo 2308/09, RiKo 2340/09, RiKo 2351/09, and W26).

The twelve GIs harbored by mastitis isolate ECC-1470 vary in size between 8 to 58 kb and code for 9 to 61 CDSs (Additional file 11: Dataset S6). *E. coli* ECC-1470 (Ont:H2) encodes for a flagellin of serogroup H2 and an uncharacterized small alternative flagellin, FlkA, encoded on GI10. The neighbouring *flkB* gene encodes for a FliC repressor. This small alternative flagellin islet can elicit unilateral H-antigen phase variation [35, 36]. The MAEC strain P4-NR (O15:H21/H54), which usually expresses a serotype H21 flagellin, also harbours a similar alternative flagellin system determinant consisting of the serotype H54 flagellin gene *flmA54* and the associated *fliC* repressor-encoding gene *fljA54*. GI12 of ECC-1470 is a large PAI containing a fimbrial operon of the P adhesin family (*pixGFJDCHAB*, *pixD* is a pseudogene), a phosphoglycerate transport operon (*pgtABCP*), the putative MAEC-associated Fec transport operon (*fecEDCBARI*), the 9-O-acetyl-*N*-acetylneuraminic acid utilization operon (*nanSMC*), and the type 1 fimbriae operon (*fimBEAICDFGH*). This PAI is a composite island with the 5′-end similar to PAI V from UPEC strain 536 with the *pix* and *pgt* loci, also present in human commensal *E. coli* A0 34/86 [37, 38], and the 3′-end similar to GI12 of MAEC 1303 with the *nan* and *fim* gene clusters. *E. coli* ECC-1470 GI4 codes for a lactose/cellobiose PTS system (*bcgAHIFER*, *bcgI* is a pseudogene).

Four prophages were predicted in the genome of MAEC 1303 ranging from 29 to 48 kb encoding for 44 to 59 CDSs (Additional file 10: Dataset S5). These prophage genomes do not comprise many virulence-associated genes, and mostly code for functions required for maintenance and mobilization. The only exception is *bor*, a gene of phage lambda widely conserved in *E. coli* and encoded by strain 1303 chromosomal prophage 1. The outer membrane lipoprotein Bor is homologous to Iss (increased serum survival) and involved in serum resistance of ExPEC [39, 40]. The lack of putative *E. coli* VFs encoded by prophages is also true for the five predicted prophage genomes of MAEC ECC-1470 (Additional file 11: Dataset S6). Two outer membrane proteins (OMPs) are encoded by ECC-1470 prophage 1, the porin NmpC and the omptin OmpT.

In summary, the MGEs of MAEC strains 1303 and ECC-1470 do not carry many known virulence-associated genes, which may entail an advantage to mastitis pathogens. To illustrate the resulting mosaic-like structure of *E. coli*, we created circular genome diagrams for all MAEC 1303 and ECC-1470 replicons indicating the core and the flexible genome by labeling the predicted GIs and prophages (Additional file 3: Figure S5A and B). Importantly, the prevalence and dissemination of the MGEs were not correlated with the pathotypes.

Virulence or fitness factors present in bovine commensal *E. coli* or MAEC

To examine the distribution of virulence-associated factors in more detail we searched for well-known *E. coli* VFs encoded by the bovine-associated *E. coli* genomes (Additional file 12: Table S5) [41]. Only about half of the 1069 gene products involved in the biosynthesis and function of 200 *E. coli* virulence and fitness-associated factors yielded BLASTP+ hits in the 25 bovine-associated *E. coli* genomes. Virulence-associated proteins of the VF panel present in (556) and absent from (513) these *E. coli* genomes are listed in Additional file 12: Table S5. Results of the BLASTP+ hits for the virulence-associated proteins are listed in Additional file 13: Table S6. Many classical IPEC VFs [42] were not present in the bovine-associated strains. Interestingly, all major virulence factors of EHEC are missing. Furthermore, several VFs associated with ExPEC [17] were absent, such as several typical serine protease autotransporters of *Enterobacteriaceae* (SPATE) like Sat and Pic (type V secretion systems, T5SS), S fimbriae, salmochelin siderophore, colicin V, and colibactin. The fecal isolate RiKo 2351/09 of phylogroup B1 yielded the most virulence-associated protein hits (297), whereas MAEC ECA-O157 of phylogroup A the fewest (162). There were 241 virulence-associated protein hits on average in the strains included in this study. We could not detect a correlation between the number of virulence-associated genes and pathotype as both, commensal strains and MAEC, exhibited comparable average virulence-associated genes hits (250 and 237, respectively). The average number of virulence-associated genes was in the same range in the *E. coli* genomes of the different phylogroups (phylogroup A: 233, B1: 254, B2: 227, and E: 235).

We converted the BLASTP+ VF hits for each strain into a presence/absence binary matrix (Additional file 14: Dataset S7) to enable grouping of the compared strains according to their VF content (Additional file 3: Figure S2C). Most of the genomes belonging to the same phylogroup clustered together. However, the phylogenetic relationships of the strains from the gold standard WGA phylogeny are not all retained. Consequently, the association of the strains with phylogroups in the VF content tree is not as well conserved as in the overall gene content tree, as shown by a tanglegram with the WGA phylogeny (Additional file 3: Figure S2D). The presence and absence of the VFs in the different strains were visualized in a heatmap in which the respective genome columns are ordered according to the clustering results (Fig. 4). This heatmap is replicated with the corresponding virulence-associated gene names in Additional file 3: Figure S6. Analogous to the all-strain soft core genome we determined an “all-strain” soft core VF set. In consideration that many fragmented draft genomes are included in the strain panel, we once more applied a 70% inclusion cutoff. As a result, virulence-associated genes were included if they were present in at least 18 of the 25 bovine *E. coli* genomes analyzed. The resulting 182 virulence-associated genes (Additional file 15: Table S7) included determinants generally considered to be widely present in *E. coli* isolates, like the Flag-1 flagella system, the operons encoding type 1 fimbriae, and the *E. coli* common pilus (ECP). But also curli fimbriae, the lipoprotein NlpI, outer membrane protein OmpA, and several iron transport systems (ferrous low pH (*efe*/*ycd*), enterobactin (*ent*, *fes*, and *fep*), ferrous (*feo*), and ferrichrome (*fhu*)) are included. Additionally, several T2SS genes, 16 of the 32 *E. coli* type three secretion system 2 (ETT2) genes, and two genes from the ECC-1470 T6SS/1, *impA* and a gene coding for a Hcp T6SS effector-like protein (E1470_c02180), are enclosed.

In conclusion, the VF variety observed is in accordance with the high diversity of bovine-associated *E. coli*.

Specific virulence or fitness genes cannot be unambiguously detected for MAEC or commensal bovine isolates

According to Fisher’s exact test performed with the 556 VF-related genes detected in our strain panel, 30 were significantly associated with mastitis or commensal isolates (Additional file 16: Dataset S8). However, with a Bonferroni correction for multiple comparisons we could not detect a significant association (Additional file 3: Figure S3C), also no VF was exclusively present in MAEC or commensal isolates. Nine virulence genes were significantly associated with mastitis genomes. Although overrepresented in mastitis isolates, the *fecRIABCDE* genes as well as the type 1 fimbriae minor subunit-encoding gene *fimG* are also present in at least 50% of the commensal genomes analyzed. The only MAEC-enriched virulence-associated OG that fulfills the 70%/30% inclusion/exclusion cutoffs, was the ETT2 *eprI* gene. Nevertheless, this gene is also enriched in phylogroups A/E (Fig. 5a, Table 2, and Additional file 17: Dataset S9). Additionally, *eprI* and the *fecBCDE* genes were also tested significantly enriched in phylogroup A strains in comparison to B1 strains; *eprI* even with a Bonferroni correction. Overall, 58 VF-related genes were significantly associated with phylogroup A or B1, and of these six with phylogroup A and 12 with phylogroup B1 after a Bonferroni correction (Additional file 3: Figure S3D and Additional file 17: Dataset S9). 21 virulence-associated genes were associated with commensal strains. Seven of them, including EC042_1639 and *ydeT* (coding for parts of Yde fimbriae), *gspFHI* and *yghJ* (coding for components of the T2SS-2 system), as well as the T3SS effector-encoding *espX1* gene display at the same time a significant enrichment in phylogroup B1 strains. Of these, the Yde fimbrial genes are significantly phylogroup B1-associated also with a Bonferroni correction and fulfill the 70%/30% inclusion/exclusion cutoff for phylogroups B1/B2/E (Table 2). The residual fourteen VF genes significantly associated with commensal isolates were not phylogroup-enriched. These virulence-associated genes are involved in biosynthesis and transport of the aerobactin siderophore (*iucABCD*, *iutA*, *shiF*), F17 fimbriae biogenesis (f17d-C, pVir_8, pVir_9) or code for an enterotoxin (*senB*) and colicin-related functions (*cjrABC*, *imm*). The presence of the aerobactin genes were also within the 70%/30% inclusion/exclusion cutoffs and not simultaneously enriched in a phylogroup (Table 2 and Additional file 16: Dataset S8). Altogether, most of the significantly phylogroup-associated VFs were also included with the 70%/30% cutoffs (Fig. 5b and Additional file 17: Dataset S9).

Because T3SS-related genes were present in MAEC and commensals, we wanted to analyze the ETT2 determinant in more detail in our strain panel. In addition to ETT2, we also examined the large ECC-1470 T6SS/1 and Flag-2 gene regions. All three putative virulence regions show a high amount of mutational isoforms and/or absence in the strain panel (Fig. 4), warranting a detailed analysis. For this purpose, the gene composition of such regions was depicted for all bovine-associated *E. coli* from the strain panel (Fig. 6 and Additional file 3: Figure S7A and B). In the case of strain D6-117.29 the ETT2 and T6SS regions could probably not be fully manually assembled, because of the high fragmentation of the genome.

The ETT2 gene cluster shows high genetic flexibility and many deletions and insertions (Fig. 6). Nevertheless, small features still reveal a phylogenetic relationship of similar pseudogene composition. For example *eprJI*, *orgB*, and *epaO* are mostly pseudogenes in B2 strains,

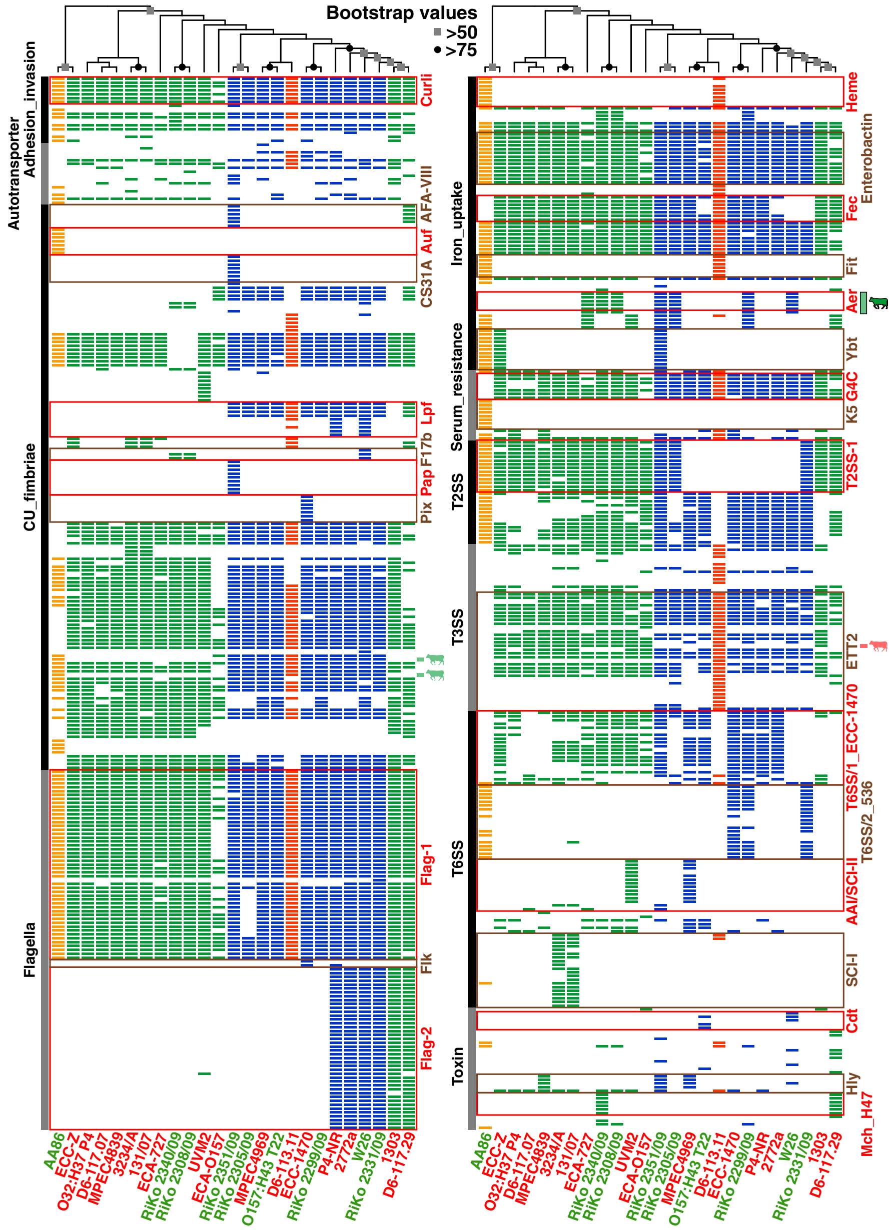


Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Heatmap indicating presence or absence of virulence factors. Each row of the binary matrix indicates the presence or absence of a virulence-associated gene (a BLASTP+ hit). VF classes are indicated at the side in *black* and *grey*. Strain names are color-coded for MAEC (*red*) or commensal (*green*) pathotype affiliation, columns for strain phylogroup affiliation (*green*: A, *blue*: B1, *orange*: B2, *red*: E). The clustering dendrogram attached to the heatmaps is based upon the whole binary dataset (not for each heatmap separately) of a best scoring ML tree with 1000 bootstrap resamplings (a more detailed representation of the cladogram can be found in Additional file 3: Figure S2C). Bootstrap support values are arbitrarily indicated at the bifurcations of the cladogram. Statistically significant pathotype-enriched VF genes are indicated for MAEC and commensal isolates by cows in *red* or *green*, respectively. Only the aerobactin biosynthesis cluster (Aer) plus transport protein ShiF is significantly commensal-enriched and not associated with a phylogroup (indicated by a *black*-rimmed and opaque *green* cow). All other pathotype-enriched virulence-associated genes also have a significant phylogroup association. The genes of well-known and important *E. coli* VFs are highlighted in alternating *red* and *brown* squares: Curli = curli fibres, AFA-VIII = aggregative adherence fimbriae AFA-VIII, Auf = fimbrial adhesin, CS31A = CS31A capsule-like antigen (K88 family adhesin), Lpf = long polar fimbriae, F17b = F17b fimbriae, Pap = P/Pap pilus, Pix = Pix fimbriae, Flag-1 = *E. coli* peritrichous flagella 1 gene cluster, Flk = alternative *flk* flagellin islet, Flag-2 = *E. coli* lateral flagella 2 gene cluster, Heme = *chu* heme transport system, Enterobactin = enterobactin biosynthesis/transport gene cluster, Fec = ferric iron(III)-dicitrate uptake system, Fit = ferrichrome iron transport system, Aer = aerobactin biosynthesis cluster with *iutA* receptor, Ybt = yersiniabactin iron transport system, G4C = group 4 capsule, K5 = K5 capsule, T2SS-1 = *gsp* general secretion pathway 1, ETT2 = *E. coli* type three secretion system 2, T6SS/1_ECC-1470 = MAEC ECC-1470 subtype i1 T6SS/1, T6SS/2_536 = UPEC 536 subtype i2 T6SS 2, AAI/SCI-II = EAEC 042 subtype i4b T6SS 3, SCI-I = EAEC 042 subtype i2 T6SS 2, Cdt = cytolethal distending toxins, Hly = alpha-hemolysin, Mch_H47 = microcin H47. Clustering of the strains according to virulence-associated gene presence/absence also follows mostly the phylogenetic history of the strains, no clustering of pathotypes was detected. Both MAEC and commensal isolates are distinguished by the lack of classical pathogenic *E. coli* VFs. The same heatmap, but including gene names/locus tags, can be found in Additional file 3: Figure S6

but the genes seem to be functional in all phylogroup A and E strains. No comparable pattern was found in relation to the pathotypes. Almost all genomes lack a fragment present in the putatively intact ETT2 region of phylogroup D1 EAEC strain 042 (*eivJICAEGF*), which is located between two small direct repeats and thus often deleted [43]. Only the ETT2 gene cluster in phylogroup E isolate D6-113.11 has an identical structure as 042 (phylogroup E is most closely related to phylogroup D1).

The Flag-2 region is basically present or entirely absent in the strain panel. No intermediate attrition isoforms are observable (Additional file 3: Figure S7A). A large deletion is apparent in O157:H43 strain T22. This deletion encompasses the whole Flag-2 region and respective flanking backbone genes. Thus, *E. coli* O157:H43 strain T22 was omitted from the diagram. Additionally, the deletion includes also the housekeeping genes downstream of the T6SS/1 gene cluster of *E. coli* ECC-1470 indicated by dots in Additional file 3: Figure S7B.

The subtype i1 T6SS/1 of MAEC strain ECC-1470 is the most variable of the virulence-associated regions investigated in more detail in this study, with many repetitive sequence subregions. Typical for T6SSs, it is also adjacent to a highly repetitive *rhs* element. Strain ECA-727 lacks the *yafT* to *impA* genes, because of a putative phage insertion in this region. This phage is not included

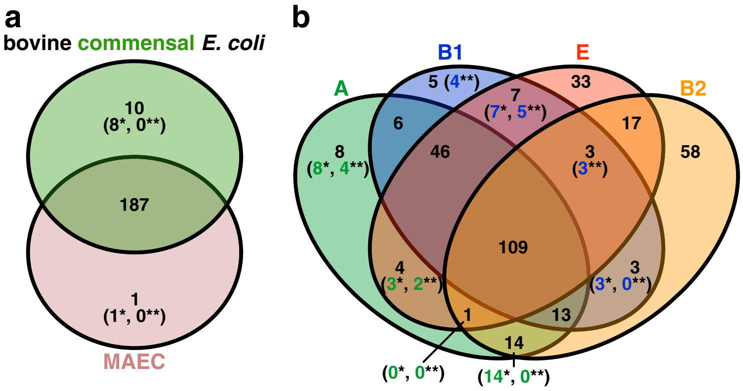


Fig. 5 Venn diagrams of virulence-associated gene enrichment in pathotypes and phylogroups. Enriched virulence-associated genes (numbers without parentheses) were identified with 70% inclusion and 30% exclusion group cutoffs for the bovine-associated *E. coli* classified either by **a** pathotype (MAEC or commensal) or **b** phylogenetic groups (A, B1, B2, or E). Statistical significance of VF association was tested with Fisher's exact test ($p < 0.05$, numbers with a *single asterisk*) and Bonferroni corrected (numbers with *two asterisks*). In the phylogroups association was only tested for the multi-genome phylogroups A versus B1. As with the OG enrichment analysis, phylogenetic lineage of the strains dominates VF content and only very few virulence-associated genes were enriched in the pathotypes

Table 2 Virulence-/fitness-associated genes significantly associated in MAEC or commensal isolates, as well as phylogroups A or B1

Gene/locus tag	Accession number	VF class	Phylogroup association, enrichment
MAEC-enriched virulence-/fitness-associated gene			
eprI	YP_006097353	T3SS/ETT2	significantly A-associated
Commensal-enriched virulence-/fitness-associated genes			
EC042_1639	YP_006095949	CU fimbriae	significantly B1-associated
ydeT	YP_006095947	CU fimbriae	significantly B1-associated
iucA	NP_755502	Iron uptake	no hit
iucB	NP_755501	Iron uptake	no hit
iucC	NP_755500	Iron uptake	no hit
iucD	NP_755499	Iron uptake	no hit
iutA	NP_755498	Iron uptake	no hit
shiF	NP_755503	Iron uptake	no hit

MAEC mastitis-associated *E. coli*, *CU* chaperone usher, *T3SS* type III secretion system, *ETT2 E. coli* T3SS 2

in the figure and the truncation is indicated by dots in the diagram. The T6SS determinants in MAEC strains 1303, MPEC4839, D6-117.29, D6-113.11, and commensal RiKo 2305/09 are most likely not functional because of their small sizes. Overall, we could not find any features of this gene cluster which are associated with phylogeny or pathogenicity of the strains.

Discussion

This is the first study which investigates *E. coli* genomes in relation to bovine mastitis including two closed genomes of finished quality [44], MAEC 1303 and ECC-1470 [45]. Closed genomes of a finished quality allow insights into the genome organization, synteny, and detection of MGEs. We additionally sequenced six bovine fecal commensal and six MAEC draft genomes and supplemented these with publicly available reference bovine *E. coli* strains. With this strain panel of 16 bovine mastitis and nine bovine *E. coli* commensal isolates we were able to analyze differences in the gene content between MAEC and commensal strains in relation to the phylogenetic as well as genomic diversity of bovine *E. coli* in general. Bovine strains are phylogenetically diverse and do not show a virulence-related gene content that is associated with either pathotype. This has implications for the definition of mastitis-related VFs and a bovine mastitis *E. coli* pathotype.

The assembly statistics of the draft genomes of this study indicate a suitable quality for the purposes of our analyses, with 24 to 290 contigs and N50 values ranging from 79 to 358 kb for contigs larger than 500 bp (Additional file 5: Table S3). There are four apparent exceptions: First, the genome of commensal reference strain AA86 has gone through multiple gap closure steps and has only five contigs with an N50 of 2860 kb [46]. Two of these five contigs are plasmids, making AA86 the only strain with resolved plasmid sequences in the strain panel in conjunction with the finished 1303 and ECC-1470 genomes [45]. Second, the three MAEC reference draft genomes D6-117.29, ECA-727, and ECA-O157 are highly fragmented with more than 500 contigs each. However, their coding percentage and overall CDS numbers are in the range of other *E. coli* genomes and thus they were included in the strain panel (Additional file 1: Table S1). Also, overall presence of VFs in the strain panel did not relate to contig number (Additional file 14: Dataset S7).

Bovine-associated *E. coli* originate mostly from phylogroups A and B1

E. coli phylogroup A is traditionally associated with commensal strains, while its sister taxon B1 is associated with commensals and different IPEC including ETEC, EAEC, and EHEC [17, 18, 42]. ECOR phylogroup E includes the genetically closely related O157:H7 EHEC and O55:H7 EPEC [47]. Interestingly, the bovine commensal O157:H43 isolate T22, even though belonging to the O157 serotype, is not a member of phylogroup E, but of group B1 [48], providing an example for horizontal transfer of O-antigen genes. Finally, phylogroup B2 is the most diverse phylogroup, based on nucleotide and gene content. This group also includes most of the ExPEC, like UPEC, APEC, and MNEC [12, 17, 42, 49]. However, with the accumulation of *E. coli* sequencing data, the traditional association of phylogroups with pathotypes have softened, as many pathotypes were shown to have emerged in parallel in different lineages [14, 18, 42].

The phylogenetic placement of the bovine isolates used in this study is in agreement with previous studies where MAEC and bovine commensals were also enriched in phylogroups A and B1, while other phylogroups play only a minor role [9, 11]. Depending on the study and the respective analyzed strain panel, MAEC isolates are either more common in phylogroup A [6, 22, 23] or phylogroup B1 [11, 27, 50]. Also, the WGA phylogeny shows that bovine MAEC and commensals do not cluster together, but rather originate from diverse lineages within phylogroups (Fig. 1) [11, 26, 27]. The discrepancies of MAEC phylogroup associations between the previous studies might be a result of country-specific differences or differences in sampling and phylotyping techniques. The polyphyletic evolutionary history of bovine *E. coli* (both MAEC or commensals)

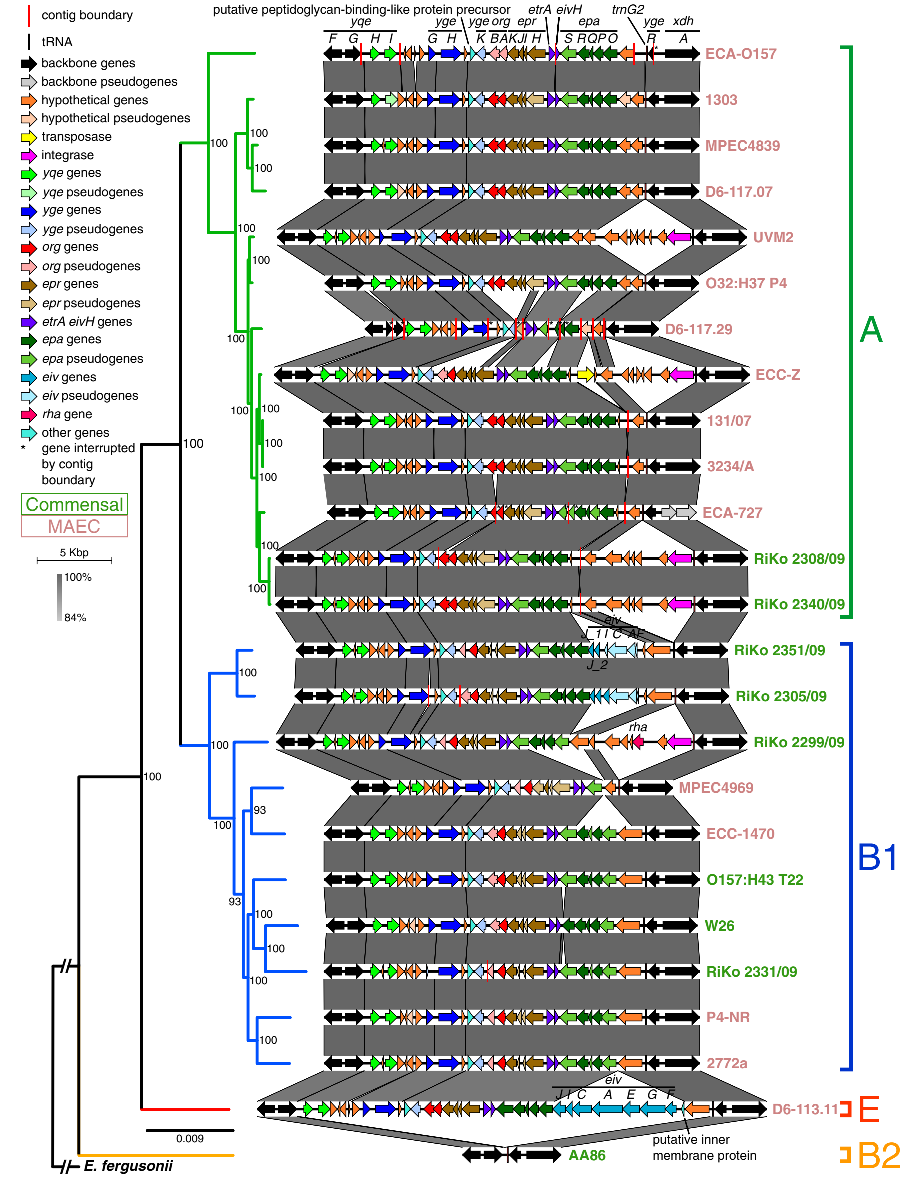


Fig. 6 Gene organization of the ETT2 gene cluster in the bovine-associated *E. coli* genomes. Comparison of the ETT2 gene cluster in the *E. coli* of the strain panel based on BLASTN+. Homologous regions are connected via grey vertices and colored by nucleotide identity. The genomes are ordered according to the WGA core genome phylogeny (Additional file 3: Figure S2A), which is attached on the left side (bootstrap support values below 50 were removed). Phylogroups are indicated correspondingly. MAEC strain names are colored in *light red* and commensals in *green*. Gene names are indicated above genomes encoding for these. The respective contigs of the draft genomes containing the gene cluster were concatenated (contig boundaries are indicated by *red vertical lines*) and CDS spanning contig borders reannotated if needed (indicated by *asterisks*). ETT2 contigs of genome D6-117.29 were difficult to concatenate, because of its high fragmentation. Backbone genes not belonging to ETT2 are colored *black*. Genes within the ETT2 region have different colors (see the legend) to be able to evaluate their presence. Pseudogenes have a lighter color fill. ETT2 shows a large number of different mutational isoforms. Nevertheless, ETT2 composition follows phylogenetic history rather than pathotype affiliation

is substantiated by their high genotypic and phenotypic plasticity [5, 6, 9, 11, 24]. In light of these studies and the genealogy of the bovine-associated *E. coli* in this work (Fig. 1) the strain panel is suitable and sufficiently diverse in its phylogeny for more detailed comparative analyses of MAEC and commensal bovine *E. coli* genomes. Two possible explanations for the phylogenetic diversity of MAEC and bovine commensals can be considered. On the one hand, the ability to cause mastitis could have been developed in parallel on several independent occasions during the evolutionary history of *E. coli* by selecting forces [26]. On the other hand, MAEC might be recruited from the normal intestinal commensal microbiota and the ability to cause mastitis is facultative, as has been proposed for ExPEC [12, 18, 19, 51].

Gene content of bovine-associated *E. coli* mirrors phylogeny rather than pathotype

Several studies have shown that recombination between extant *E. coli* phylogroups is limited by phylogenetic diversity [29, 30]. Thus, the phylogenetic background of *E. coli* has a big impact on possible recombination events and most importantly on the gene content of the flexible genome [15, 18]. Nevertheless, there are examples of convergent evolution in *E. coli,* especially in IPEC pathotypes from multiple parallel phylogenetic origins that typically contain a specific set of VFs, e.g. the occurrence of EHEC in the distant phylogroups B1 and E mediated by HGT of MGEs [29, 47]. Our study demonstrates that there is no evidence for HGT of large mastitis-specific genomic regions, and that the phylogenetic background of the strains has a deciding impact on the overall gene content (Fig. 2 and Additional file 3: Figure S2B). A clustering of strains according to pathotypes would have hinted towards a common gene content and a difference in ecological lifestyles and habitats, as a result of positive selection on the ancestral genomes [29]. However, our results demonstrate that the flexible and the core genome appear to coevolve.

Two previous studies with similar methodology came to two different conclusions. Blum et al. [25] reasoned that three mastitis strains (O32:H37 P4, VL2874, VL2732) were much more closely related in gene content compared to an environmental (commensal fecal) strain (K71), based on the different pathotypes. However, the MAEC in this study are phylogenetically strongly related (phylogroup A) whereas the single commensal strain belongs to phylogroup B1. Thus, as we observed in our study, the phylogenetic relationship had a strong impact on the gene content dendrogram. Kempf et al. [27] comparing four phylogroup A MAEC (D6-117.07, O32:H37 P4, VL2874, VL2732), one phylogroup E MAEC (D6-113.11), and the K71 commensal, achieved results comparable to ours. The authors argued that mastitis pathogens with different phylogenetic histories might employ different virulence strategies to cause mastitis, similar to the variable VF repertoire of ExPEC. A hypothesis we tested in this study by searching for well-known *E. coli* VFs in the bovine-associated *E. coli* strains discussed below.

MAEC cannot be distinguished from bovine commensal *E. coli* based on the presence of virulence-associated genes

It was suggested that the genome content of MAEC is distinct from bovine commensals and not random, as a result of selective pressure. VFs important for MAEC pathogenicity would then supposedly be positively selected within the bovine udder [25, 26]. Several virulence-associated properties have been proposed for MAEC pathogenicity [3, 27, 52]: multiplication and persistence in milk and the udder [10, 53], resistance to serum components and neutrophil neutralization mechanisms [7, 54, 55], adhesion to (and invasion of) mammary epithelial cells [4, 6, 10, 56], and stimulation of the innate immune response by PAMPs [57, 58]. Against this background, a myriad of previous publications have tried to identify VFs specific for MAEC with varying degrees of success [6, 11, 22–24, 50, 55, 59–61]. However, the results of these studies do not agree upon the identified VFs, which is due to the diversity of MAEC and bovine *E. coli*, generally. The aforementioned publications followed a classical diagnostic typing procedure by using PCR assays for virulence-associated gene detection. Only Kempf and colleagues applied a bioinformatic approach similar to ours with a candidate VF panel of 302 genes [27]. However, our larger strain and selected VF panels enabled a more detailed analysis. In our study, commensal strains and MAEC exhibited a similar average virulence-associated gene presence (250 and 237, respectively), also comparable to averages of the *E. coli* genomes categorized by the different phylogroups (phylogroup A: 233, B1: 254, B2: 227, and E: 235). Additionally, we used a bottom-up approach to identify overall OGs associated with MAEC. For these analyses we allocated the analyzed strains into pathotypes (MAEC or commensal) based on their source of isolation and used 70%/30% inclusion/exclusion cutoffs to detect OG/VF association with either group. Furthermore, we applied Fisher's exact test to determine statistically significant associations between OGs or VFs and pathotypes or phylogroups. Both our comparisons of the association of OGs and VFs with mastitis or commensal genomes did not reveal a significant correlation between the presence of individual virulence-related genes and the mastitis-associated isolates (Figs. 3a and 5a, Additional file 7: Dataset S3 and Additional file 16: Dataset S8). Using 70%/30% inclusion/exclusion cutoffs, we recovered only one significantly MAEC- and eight significantly commensal-enriched VF genes (Table 2). As expected from the OG analysis (Fig. 3b and Additional file 8: Dataset S4), the phylogroups had also a strong impact on VF enrichment (Fig. 5b and Additional file 17: Dataset S9).

The aerobactin gene cluster together with the *iutA* and *shiF* genes were detected as significantly associated and enriched in the commensal strains in both our OG and VF analyses (Table 2 and Additional file 8: Dataset S4). The siderophore system aerobactin is considered an ExPEC VF needed for iron uptake under limiting conditions, e.g. in the urinary tract or serum [62]. These genes are often encoded by plasmids harboring additional traits, like colicins and other iron transport systems, e.g. in APEC colicin plasmids [62, 63]. Thus, its distribution might also be due to positive selection of beneficial traits for commensalism, which are encoded by the same plasmid. Another group of commensal-enriched virulence genes included fimbriae-associated genes (EC042_1639, *ydeT*) (Table 2). However, these two genes have in common, that they are enriched in phylogroup B1.

The *eprI* gene was determined as the only significantly mastitis-associated and -enriched virulence-associated gene (Table 2). This gene, which was also shown to be significantly associated with phylogroup A strains, belongs to the ETT2 determinant, a large gene cluster with frequent deletion isoforms in *E. coli* [43]. The ETT2 type III secretion system is contained on GI8 of *E. coli* 1303 and on GI9 of strain ECC-1470 (Additional file 10: Dataset S5 and Additional file 11: Dataset S6). ETT2 has not only been discussed as a VF during mastitis [64], but has also been implicated in being involved in invasion and intracellular survival of blood-brain barrier cells of MNEC K1 strains [65] and in serum resistance of APEC O78:H19 strain 789, besides its degenerate form in the strain [62]. Its prevalence has been analyzed in bovine mastitis *E. coli* isolates and was determined to be approximately 50% [64]. ETT2 has different mutational attrition isoforms in our bovine-associated strain panel, supporting the results of an earlier study [64]. However, overall ETT2 presence was not related to MAEC (Fig. 6). Based on the comparative analysis, and in accordance with Blum et al. [10] these results suggest that serum resistance is not an essential trait for the ability of MAEC to cause intramammary infections. Thus, a role of ETT2 in MAEC is debatable, especially since only *eprI* and none of the other ETT2 genes were MAEC-enriched. In conclusion, MAEC are characterized by a lack of “bona fide” VFs [11, 24, 27]. Instead, the VF variety observed rather mirrors the genome plasticity of bovine-associated *E. coli*, regardless of pathotype. Although many of these putative VFs are not connected to mastitis virulence, they are still maintained within the genomes. This suggests that they serve as FFs for gastrointestinal colonization and propagation, rather than VFs.

Large virulence regions and intraphylogroup comparisons of putative VFs are also not pathotype-specific

An alternative flagellar system (Flag-2) is encoded on 1303 GI1 [43]. The Flag-2 locus encodes also for a type III secretion system in addition to the alternative flagellar system, which might be in cross-talk with ETT2. In contrast to the typical *E. coli* peritrichous flagella 1 gene cluster (Flag-1), which is a polar system for swimming in the liquid phase, the lateral Flag-2 most likely has its functionality in swarming ability over solid surfaces [43]. Flagella are important for motility, but also for adherence during host colonization and biofilm formation [16]. Additionally, flagella might play an important role in the udder for dissemination from the teat and counteracting washing out during milking [10]. MAEC ECC-1470 also carries two T6SS determinants located on GI1 (designated as the first ECC-1470 T6SS, T6SS/1) and on GI8 (ECC-1470 T6SS/2), respectively. *E. coli* ECC-1470 T6SS/1 is classified as subtype i1 [66] or the second *E. coli* T6SS-2 phylogenetic cluster [67] and T6SS/2 as subtype i2 [66] or the first *E. coli* T6SS-1 cluster [67]. Subtype i1 T6SSs generally participate in interbacterial competition, subtype i2 T6SSs target eukaryotic cells and play a role in the infection process of pathogens. All T6SS are implicated in mediating adherence and biofilm formation [67]. The GI1-encoded T6SS/1 was consistently present in strains ECA-O157, ECA-727, and ECC-Z, but only sporadically in human reference commensal strains, and thus associated with MAEC in a preceding study [28]. Nevertheless, the corresponding phenotypes of these systems are mainly unknown and their function, especially any putative role in mastitis, might well be indirect [67]. Because of a low prevalence of T6SS genes in the five included MAEC genomes and presence in commensal strain K71, another previous study questioned the role of T6SS systems in mastitis [27]. We can support this study, as even our detailed analysis of the ETT2, Flag-2, and T6SS/1 regions did not reveal any association with MAEC isolates in our strain panel. These regions of the flexible genome mirror the underlying genomic and phylogenetic diversity of bovine *E. coli*.

Several studies argue that MAEC strains from divergent phylogenetic backgrounds might use different VF subsets and virulence strategies to elicit bovine mastitis [10, 22, 26, 27]. We tested this hypothesis exemplarily, by analyzing the 31 genes of the *fec*, *paa*, and *pga* regions for pathotype enrichment within the multi-genome phylogroups of our strain panel, A and B1. These three regions were detected as being essential in phylogroup A MAEC [26], but they have not been analyzed in other phylogroups. Five of the *pga* and all *fec* genes were significantly associated with MAEC, however *fecBCDE* also with phylogroup A (Additional file 18: Dataset S10). Also, all genes of the three regions were included in the all-strain soft core with the 70% inclusion threshold (Additional file 15: Table S7). Thus, none of the genes were associated with pathotype and only the *paa* phenylacetic acid degradation pathway determinant was missing in the single-genome ECOR phylogroups B2 and E. This might have tipped the scales in the analysis of phylogroup A genomes by Goldstone and co-workers. The 13 phylogroup A strains of our strain panel contain eleven MAEC and two commensal isolates. The ten strains of phylogenetic lineage B1 comprise four MAEC and six commensal strains. Due to ongoing sequencing efforts, the number of suitable reference genomes for more detailed analyses is likely to increase in the near future. However, this is the first study to be able to perform such an analysis. In the ECOR group A genomes, the *fec*, *paa*, and *pga* regions were not pathotype-enriched (with the 70%/30% inclusion/exclusion cutoffs), but were present in the group soft core (except for *paaB* in the unspecific category; Additional file 19: Dataset S11) and none were statistically significantly associated via Fisher's exact test. In a similar way, the PGA biosynthesis and Fec-system encoding regions were also mostly categorized into the group soft core of the analysis with B1 strains (Additional file 20: Dataset S12) and again with no significant Fisher's exact test *p*-values. Only *fecBCDE* were in the unspecific category, because these genes are missing in the genomes of the commensal isolates RiKo 2331/09, O157:H43 T22, and W26. However, the whole seven-gene *pga* region was MAEC-enriched in our phylogroup B1 strain set (albeit without significance), present in all four MAEC, but only in two of the six commensals. We want to stress that this result depends highly on the strain collection used and more bovine *E. coli* strains, especially commensals, from all available phylogroups need to be incorporated for an in-depth analysis. As all three regions are present in the all-strain soft core genome of our whole strain panel analysis, these results illustrate the drawbacks of inferring general observations from low numbers of strains (especially when focusing only on pathogenic strains) considering the genome plasticity of bovine *E. coli*.

Conclusions

This is the first publication to include a phylogenetically diverse bovine *E. coli* strain panel incorporating both MAEC and commensal isolates for genomic content comparisons. Besides the two closed bovine MAEC 1303 and ECC-1470 [45], that can serve as high sequence and annotation quality references, this study includes the largest collection of bovine *E. coli* commensals from fecal origin of udder-healthy cows [68]. As we could not identify any genes significantly associated with MAEC that were not also present in commensal strains or correlated with the strains' phylogenetic background, an MPEC pathotype characterized by specific VFs could not be defined. It is more likely that virulence-associated genes, which have been previously implicated in facilitating mastitis, have their principal function in colonization and persistence of the gastrointestinal habitat. Thus, like ExPEC, MAEC are facultative and opportunistic pathogens basically of naturally occurring commensal ("environmental") *E. coli* origin [12, 18, 19, 23, 24, 51]. As a consequence, we propose to use the term mastitis-associated *E. coli* (MAEC) instead of mammary pathogenic *E. coli* (MPEC).

The genome content of certain bovine *E. coli* strains seems not to support the ability to elicit mastitis in udder-healthy cows as was shown in the case of the commensal strain K71 [25]. The large pan-genome of bovine *E. coli* isolates offers many gene combinations to increase bacterial fitness by utilization of milk nutrients and evasion from the bovine innate immune system, thus resulting in sufficient bacterial intra-mammary growth and consequently infection of the mammary gland [10, 53, 69, 70]. Isolates with an increased potential to cause mastitis can colonize the udder by chance depending on suitable environmental conditions and the cow's immune status. Our data also demonstrate, that there is no positive selection in MAEC for the presence of virulence-associated genes required for causing mastitis. This has implications for vaccine development and diagnostics. Reverse vaccinology may not be suitable for the identification of specific MAEC vaccine candidates, and the utilization of marker genes for improved diagnostics and prediction of the severity and outcome of an *E. coli* bovine mastitis might fail. Herd management and hygiene are still the two most important factors for preventing *E. coli* mastitis incidents. Several studies have shown a dramatic decrease in the bovine udder microbiome during mastitis, even after recovery [71–73]. It might be worthwhile to consider alternative prevention strategies like strengthening the natural udder microbiota that competes with pathogens [74].

We urge the research community to not fall into the same trap with whole genome studies as with the previous typing studies. Mastitis researchers need to consolidate their efforts and, as Zadoks et al. eloquently put it, not to waste precious resources on "YATS" (yet another typing study) [5]. It is necessary to step away from the reductionist approach and adapt an integrated course of action by examining the host-pathogen interaction simultaneously. Synergistic application of techniques, like dual RNA-Seq of host and bacteria [75], Tn-Seq to test virulence association of genes in vivo, comparative SNP analysis of orthologous genes and intergenic regions, proteomics, and metabolomics, are readily available to correlate physiological traits with genomic information. Additionally, the comparison of closed genomes offers the possibility to comprehensively analyze the complete genomic context of strains including genomic architecture, rearrangements and movement of mobile genetic elements.

Methods

A detailed method section can be found in Additional file 21.

Bacterial strains, isolation, and published reference genome acquisition

All fourteen isolates in this study were collected using routine clinical practices from the bovine habitat. Commensal strains were isolated from fecal samples of udder-healthy and mastitis-associated strains from the serous udder exudate of mastitis-afflicted cows. Mastitis strains were acquired from different veterinary diagnostic laboratories in the indicated countries, listed in the genomes feature overview table (Additional file 1: Table S1). Additionally, eleven draft bovine-associated *E. coli* reference genomes were downloaded from NCBI to be used in the analyses. See Table 1 for the respective reference publications. The corresponding accession numbers are given in Additional file 1: Table S1.

Library preparation and sequencing

The strains with closed genomes, 1303 and ECC-1470, were sequenced on a 454 Titanium FLX genome sequencer with GS20 chemistry as described in [45].

These two strains were additionally and the draft strains [68] solely sequenced with a 101-bp PE sequencing run on a HiScan SQ sequencer (Illumina, San Diego, CA, USA).

Assembly of the genomes

Both 454 read sets for the genomes of *E. coli* 1303 and ECC_1470 were *de novo* assembled with Newbler (Roche) (v2.0.00.20 for 1303 and v2.3 for ECC-1470) [76]. Additionally, these reads were assembled in a hybrid *de novo* approach in combination with the respective Illumina reads using MIRA (v3.4.0.1) [77]. Afterwards, each 454 Newbler assembly was combined with the respective hybrid assembly in Gap4 (v4.11.2) of the Staden software package [78]. The remaining gaps in the assembly were closed by primer walking via directed PCR and Sanger sequencing utilizing BigDye Terminator chemistry with ABI 3730 capillary sequencers. The closed genomes were edited to the “finished” standard [44].

The Illumina reads from the draft *E. coli* genomes were each randomly subsampled to an approximate 70-fold coverage with seqtk (v1.0-r32; https://github.com/lh3/seqtk). Afterwards, the PE reads were *de novo* assembled with SPAdes (v3.1.1) [79] and only contigs > = 500 bp retained. At last, the assembled contigs were ordered against the respective *E. coli* 1303 or *E. coli* ECC-1470 reference genomes, according to the ECOR phylogroup affiliation of the draft genomes. All Sequence Read Archive (SRA) study accession numbers for the Illumina and 454 raw reads of the *E. coli* genomes of this study can be found in Additional file 5: Table S3. This file also includes the assembly statistics for all 23 bovine-associated *E. coli* draft genomes. The draft genomes of this study are in the “high-quality draft” standard [44].

Annotation of the genomes

The complete genome sequences of *E. coli* 1303 and ECC-1470 were initially automatically annotated with Prokka (v1.9) [80] and the annotations subsequently supplemented with further databases. This automatic annotation was manually curated with Artemis (v15.1.1) [81] and tbl2tab (v0.1) [82]. Additionally, the annotations of *E. coli* strains 1303 and ECC-1470 were compared (ACT, v12.1.1 [83], and BLASTN+, v2.2.28 [84]) and adapted to each other for a uniform annotation. The high quality annotation of the *E. coli* 1303 genome was then used as reference for the ECOR phylogroup A strains and the ECC-1470 genome annotation for the ECOR B1 strains during the Prokka annotation of the 12 draft genomes of this study.

All eleven reference strains were also automatically reannotated with Prokka to have a uniform ORF-finding and facilitate comparative genomics. The annotations of the references were shortly manually curated in the three putative virulence regions ETT2, Flag-2, and strain ECC-1470’s T6SS/1 by comparisons to the 1303 and ECC-1470 genomes as mentioned above. GENBANK files for these reannotations can be found in Additional file 22: Dataset S13 and Additional file 23: Dataset S14. For an overview of the annotations see the genome feature table created with genomes_feature_table (v0.5) [82] (Additional file 1: Table S1). This table also includes the reference *E. coli* genomes for the phylogenetic analysis (see below), however their annotation features are listed as downloaded from NCBI.

Phylogenetic analysis

A WGA of selected *E. coli* genomes was done with the default parameter settings of Mugsy (v1.2.3) [85] and with *E. fergusonii* as outgroup. The MAF alignment file was further processed to contain only locally colinear blocks without gaps present in all aligned genomes utilizing the software suite Phylomark (v1.3) [86]. The concatenated and filtered alignment was then subjected to RAxML (v8.1.22) [87] to infer the best scoring ML phylogeny with 1000 bootstrap resamplings for local support values. The resulting tree was visualized with Dendroscope (v3.4.4) [88]. This phylogeny was used to classify the bovine-associated strains into ECOR phylogroups according to the included reference strains (with a known phylogeny) and monophyletic clades. The same procedure was followed including only the 25 bovine-associated *E. coli* strains. This tree was visualized with FigTree (v1.4.1; http://tree.bio.ed.ac.uk/software/figtree/) midpoint rooted.

STs were assigned with ecoli_mlst (v0.3) [82] according to the Achtman *E. coli* MLST scheme [13] employing NUCmer with default parameters. PHYLOViZ (v1.1) [89] was used to create a MST with the goeBURST algorithm [90] to classify the STs into CCs. CC numbers were allocated according to the Achtman *E. coli* MLST database. All allele, ST, and CC numbers can be found in Additional file 2: Table S2.

Detection of genomic islands and prophages, and generation of circular genome diagrams

GIs and prophages were predicted in the two closed genomes. GIs were predicted with the three prediction methods of IslandViewer 3 [91]: the two sequence composition methods SIGI-HMM [92] and IslandPath-DIMOB, and the comparative genomic prediction method IslandPick [93]. Only predicted GIs with a size greater than 8 kb were retained. Prophages were predicted with the PHAge Search Tool (PHAST) [94]. Circular genome views were created with the BLAST Ring Image Generator (BRIG, v0.95) using BLASTP+ (v2.2.28) [84] with a disabled low complexity filter (option '-seg no') and upper/lower identity thresholds set to 90 and 70%, respectively. The location of the predicted GIs and prophages are visualized in these diagrams.

Identifying serotypes

The SerotypeFinder (v1.0) database was used to determine serotypes *in silico* [95]. For some strains SerotypeFinder could not resolve the O- or H-antigen uniquely, in these cases both are listed.

Ortholog/paralog analysis

Orthologous and paralogous proteins in all 25 bovine-associated genomes were identified with Proteinortho (v5.11) [96, 97] with a 1 x 10^{-5} *E*-value and 70% coverage/identity cutoffs. This resulted in a total number of 13,481 OGs from the overall 116,535 CDSs in the bovine-associated strain panel.

To identify pathotype- (mastitis/commensal) or phylogroup-enriched (ECOR phylogroups A/B1/B2/E) OGs we employed Fisher's Exact test provided in R (v3.2.5) and tested for OGs which are significantly ($p <$ 0.05 and with a Bonferroni correction) associated with the different pathotype or phylogenetic groups. Additionally, OGs were considered enriched if they are minimally present in 70% of the genomes of one genome group (inclusion cutoff) and in maximally 30% of genomes of all other groups (exclusion cutoff) using po2group_stats (v0.1.1) [82]. The Fisher's exact test *p*-values were visualized as Manhattan plots with R package ggplot2 (v2.2.0) [98] (Additional file 4: Dataset S1). An "all-strain soft core genome" over all genomes with the 70% inclusion cutoff (rounded 18 genomes of the total 25) was determined.

The resulting pathotype-enriched OGs were further evaluated by comparing their representative proteins to the representative proteins in the phylogroup-enriched categories and the all-strain soft core. For this purpose, the prot_finder pipeline with BLASTP+ was used, as described below in the VF workflow, with the pathotype-enriched representative proteins as queries and the phylogroup-enriched or all-strain/phylogroup soft core proteins as subjects.

Finally, a gene content tree was calculated using RAxML (v8.0.26) and 1000 resamplings with the Proteinortho presence/absence matrix of OGs (included in Additional file 4: Dataset S1). The clustering tree was visualized midpoint rooted with Figtree.

Screening of the genomes for known virulence factors

VF reference protein sequences were collected from the VFDB [99–101] and the primary literature. For an overview of the VF panel see Additional file 12: Table S5 and the GitHub repository https://github.com/aleimba/ecoli_VF_collection (v0.1) [41].

The VF panel was used to assess the presence/absence of the 1069 virulence-associated genes in the annotated bovine-associated strains with the prot_finder pipeline (v0.7.1) [82] using BLASTP+ (v2.2.29) with the following options: 1 x 10^{-10} *E*-value cutoff ('-evalue 1e-10'), 70% query identity and coverage cutoffs (options '-i' and '-cov_q'), and the best BLASTP hits option ('-b'). As with the gene content tree, a ML RAxML BINGAMMA search was done to cluster the results in the VF binary matrix with 1000 resamplings. Additionally, the binary VF hit matrix was visualized with function heatmap.2 of the R package gplots and R package RColorBrewer (v1.1-2) [102]. The RAxML cladogram was attached to this heatmap with R package ape (v3.4) [103]. The binary matrix, the cladogram NEWICK file, and the R script are included in Additional file 14: Dataset S7.

VF associations with either pathotypes or phylogenetic groups were tested with a two-tailed Fisher's exact test for significance ($p < 0.05$) and with a Bonferroni-corrected significance value. Manhattan plots were created with R package ggplot2 (Additional file 4: Dataset S1). Again, inclusion and exclusion cutoffs were set to 70 and 30%, respectively, using binary_group_stats (v0.1) [82]. Also, an all-strain soft core VF set was calculated over the virulence-associated gene hits of all genomes with a 70% (18 genome) inclusion cutoff. Pathotype-enriched VF proteins were compared to phylogroup-enriched VF proteins for evaluation.

The same prot_finder pipeline and binary_groups_stats workflow was also used for two putative MAEC-specific regions in ECOR phylogroup A genomes [26], which are not included in the VF panel. The first region is the biofilm-associated polysaccharide synthesis locus (*pgaABCD-ycdT-ymdE-ycdU*). The locus tags in *E. coli* genome 1303 are EC1303_c10400 to EC1303_c10440, EC1303_c10470, and EC1303_c10480. The second region encodes proteins involved in the phenylacetic acid degradation pathway (*feaRB-tynA-paaZABCDEF-GHIJKXY*; MG1655 locus tags b1384 to b1400). The third region (the Fec uptake system, *fecIRABCDE*) is already included in the VF panel of this study. For this analysis the resulting binary BLASTP+ hit matrix was also tested with binary_groups_stats for pathotype association within the ECOR A and B1 phylogroups of the bovine-associated strain panel (with the 70% inclusion and 30% exclusion cutoffs). Associations were additionally tested with Fisher's exact test for significance.

Analysis of large structural putative virulence regions

The composition of the large virulence regions ETT2, Flag-2, and the T6SS/1 subtype i1 determinant of *E. coli* ECC-1470 as well as the antimicrobial multidrug resistance element of 1303 (AMR-SSuT in GI4) was compared in more detail for the bovine-associated strain panel with Easyfig (v2.2.2) [104].

General data generation and figure editing

Dendroscope was used to create tanglegrams between the cladograms of the bovine-associated strain panel WGA phylogeny, gene content, or VF clustering trees. All figures were created either with R (v3.2.5) [105] for the heatmap, Manhattan plots, or venn diagrams, Dendroscope or FigTree for phylogenetic trees, PHYLOViZ for the MLST MST, or Easyfig for the genome diagrams and color edited, labelled, or scaled with Inkscape (v0.91) without changing data representation. The only exception are the BRIG circular genome diagrams which were edited with Gimp (v2.8.16).

Additional files

Additional file 1: Table S1. Genome feature table for the 64 *E. coli*, four *Shigella* spp., and the one *Escherichia fergusonii* genomes plus accession numbers. (XLSX 18 kb)

Additional file 2: Table S2. MLST allele profiles, ST and CC numbers for the 64 *E. coli* and four *Shigella* spp. strains. (XLSX 10 kb)

Additional file 3: Figure S1. Minimum spanning tree (MST) of the MLST results. **Figure S2.** Phylograms and tanglegrams for the 25 bovine-associated *E. coli* genomes based on WGA core genome, gene and VF content. **Figure S3.** Manhattan plots of Fisher's exact test *p*-values for the OG/VF pathotype (MAEC/commensal isolates) and phylogroup (A/B1) associations. **Figure S4.** Gene organization of the AMR-SSuT/Tn*10* gene cluster. **Figure S5.** Circular genome diagrams for the MAEC 1303 and ECC-1470 replicons with GI and prophage positions. **Figure S6.** Heatmap of VF presence/absence, including gene names/locus tags. **Figure S7.** Gene organization of the Flag-2 and ECC-1470 T6SS/1 gene clusters. (PDF 18 Mb)

Additional file 4: Dataset S1. This zip archive contains the binary presence/absence matrix of 13,481 OGs in the 25 bovine-associated *E. coli* genomes and the R script for the Fisher's exact tests to test the associations of OGs/VFs with either pathotype or phylogroup (A/B1). The R script includes also code to create the Manhattan plots in Additional file 3: Figure S3. The binary presence/absence matrix of virulence-associated genes needed as second input for the R script is enclosed in Additional file 14: Dataset S7. (ZIP 16 kb)

Additional file 5: Table S3. This file includes the SRA study accession numbers for the Illumina and 454 raw reads of the 14 *E. coli* genomes of this study. Additionally it lists the assembly statistics for all 23 bovine-associated *E. coli* draft genomes. (XLSX 8 kb)

Additional file 6: Dataset S2. Singleton OGs in the 25 bovine-associated *E. coli* genomes. (XLSX 248 kb)

Additional file 7: Dataset S3. This file includes the pathotype-enriched OGs (MAEC or commensal isolates) with a 70% inclusion and 30% exclusion cutoff and their potential association with phylogroup-enriched categories or soft core genomes. OGs significantly associated according to Fisher's exact test and with a Bonferroni correction are indicated. It also specifies the pathotype group soft core genome and OGs classified as underrepresented and unspecific. For each OG the locus tag and annotation of one representative protein from one *E. coli* genome of the group is shown (or in the case of paralogs several representative proteins). (XLSX 523 kb)

Additional file 8: Dataset S4. Phylogroup-enriched OGs (A, B1, B2, or E) with Fisher exact test *p*-values (and Bonferroni correction) for phylogroup A versus B1 associations and vice versa. Furthermore the file includes the phylogroup group soft core, underrepresented, and unspecific OGs. (XLSX 541 kb)

Additional file 9: Table S4. All-strain soft core genome with 70% inclusion cutoff. (XLSX 185 kb)

Additional file 10: Dataset S5. Predicted GIs and prophages of MAEC 1303. (XLSX 68 kb)

Additional file 11: Dataset S6. Predicted GIs and prophages of MAEC ECC-1470. (XLSX 46 kb)

Additional file 12: Table S5. This file contains the overview of the VF panel. Presence ('1') and absence ('0') of the virulence-associated genes in the 25 bovine-associated *E. coli* genomes is indicated in column "present_in_strain_panel". Virulence-associated genes were collected from the Virulence Factors Database (VFDB) or from the primary literature ('own' in column "source"). (XLSX 55 kb)

Additional file 13: Table S6. BLASTP+ hit results for the VF panel in the 25 bovine-associated *E. coli* genomes. (XLSX 365 kb)

Additional file 14: Dataset S7. This zip archive contains the binary presence/absence matrix of virulence-associated genes in the 25 bovine-associated *E. coli* genomes, the VF content clustering cladogram in NEWICK format, and the R script to create the heatmaps in Fig. 4 and Additional file 3: Figure S4. (ZIP 6 kb)

Additional file 15: Table S7. All-strain soft core VF set with 70% inclusion cutoff. (XLSX 10 kb)

Additional file 16: Dataset S8. This file includes virulence-associated genes with significant Fisher's exact test *p*-values (and Bonferroni correction), which tested the association of VFs with either pathotype (MAEC or commensal strains). Additionally, the pathotype-enriched virulence-associated genes (MAEC or commensal isolates) with a 70% inclusion and 30% exclusion cutoff and their potential association with phylogroup-enriched categories or soft core genomes are listed. It also specifies the pathotype group soft core VF set and virulence-associated genes classified as underrepresented and unspecific. (XLSX 29 kb)

Additional file 17: Dataset S9. Significant virulence associated genes (with and without Bonferroni correction) with phylogroup A versus B1 according to Fisher's exact test. Moreover, phylogroup-enriched virulence-associated genes (A, B1, B2, or E), phylogroup group soft core

VF set, underrepresented, and unspecific virulence-associated genes are specified. (XLSX 46 kb)

Additional file 18: Dataset S10. BLASTP+ hit results for the *pga* and *paa* gene regions and binary presence/absence matrix in the 25 bovine-associated *E. coli* genomes. The spreadsheet file includes also Fisher's exact test *p*-values for significant associated genes to either pathotype or phylogroup. (XLSX 40 kb)

Additional file 19: Dataset S11. Pathotype group soft core and unspecific categorisation of the *fec*, *paa*, and *pga* gene regions in the 13 phylogroup A bovine-associated *E. coli* genomes. (XLSX 8 kb)

Additional file 20: Dataset S12. Pathotype-enriched (with Fisher's exact test *p*-values), group soft core, and unspecific categorisation of the *fec*, *paa*, and *pga* gene regions in the ten phylogroup B1 bovine-associated *E. coli* genomes. (XLSX 9 kb)

Additional file 21: Detailed Material & Methods description. (DOCX 108 kb)

Additional file 22: Dataset S13. This zip archive contains the GENBANK files with the reannotations of five of the eleven reference bovine-associated *E. coli* genomes. Included are *E. coli* strains AA86, D6-113.11, D6-117.07, D6-117.29, and ECA-727. (ZIP 15 Mb)

Additional file 23: Dataset S14. This zip archive contains the GENBANK files with the reannotations of six of the eleven reference bovine-associated *E. coli* genomes. Included are *E. coli* strains ECA-O157, ECC-Z, O32:H37 P4, P4-NR, O157:H43 T22, and W26. (ZIP 18 Mb)

Abbreviations

AIEC: Adherent invasive *E. coli*; AMR-SSuT: Antimicrobial multidrug resistance to streptomycin, sulfonamide, and tetracycline; APEC: Avian pathogenic *E. coli*; BRIG: BLAST Ring image generator; CC: Clonal complex; CDS: Coding DNA sequence; CU: Chaperone usher pathway fimbriae; EAEC: Enteroaggregative *E. coli*; ECOR: *E. coli* collection of reference strains; ECP: *E. coli* common pilus; EHEC: Enterohaemorrhagic *E. coli*; EPEC: Enteropathogenic *E. coli*; ETEC: Enterotoxigenic *E. coli*; ETT2: *E. coli* type III secretion system 2; ExPEC: Extraintestinal pathogenic *E. coli*; Fec: Ferric iron(III)-dicitrate uptake system; FF: Fitness factor; Flag-1: *E. coli* peritrichous flagella 1 gene cluster; Flag-2: *E. coli* lateral flagella 2 gene cluster; G4C: Group 4 capsule; GI: Genomic island; GTR: Generalized time-reversible; HGT: Horizontal gene transfer; IPEC: Intestinal pathogenic *E. coli*; IS: Insertion sequence; LEE: Locus of enterocyte effacement; LPS: Lipopolysaccharide; MAEC: Mastitis-associated *E. coli*; MGE: Mobile genetic element; ML: Maximum likelihood; MLST: Multi-locus sequence typing; MNEC: Newborn meningitis-associated *E. coli*; MPEC: Mammary pathogenic *E. coli*; MST: Minimum spanning tree; OG: Orthologous group; OMP: Outer membrane protein; ORF: Open reading frame; PAI: Pathogenicity island; PAMP: Pathogen-associated molecular pattern; PE: Paired-end; PHAST: PHAge Search Tool; PTS: Phosphotransferase system; Rhs: Rearrangement hotspot; SLV: Single locus variant; SPATE: Serine protease autotransporters of *Enterobacteriaceae*; SRA: Sequence read archive; ST: Sequence type; T2SS: Type II secretion system; T3SS: Type III secretion system; T5SS: Type V secretion system; T6SS: Type VI secretion system; UPEC: Uropathogenic *E. coli*; VF: Virulence factor; VFDB: Virulence factors database; WGA: Whole genome nucleotide alignment

Acknowledgements

We thank Alan McNally for hosting Andreas Leimbach for a week and showing him some of the comparative genomics ropes. We would also like to thank David E. Kerr, Wolfram Petzl, Ynte Schukken, Nahum Shpigel, Olga Wellnitz, Lothar Wieler, and Holm Zerbe for supplying strains.

Funding

This work, including the efforts of AL and UD, was funded by Deutsche Forschungsgemeinschaft (DFG) (DO 789/3-1 and DO 789/4-1).

Availability of data and materials

All raw reads of this study (454 and Illumina) can be accessed from NCBI's SRA. The corresponding SRA study accession numbers are listed in Additional file 5: Table S3. The assembled and annotated genomes of this study have been deposited at DDBJ/ENA/GenBank under the accession numbers listed in Additional file 1: Table S1. The reannotated sequence files of the eleven bovine-associated reference *E. coli* are available in Additional file 22: Dataset S13 and Additional file 23: Dataset S14. The *E. coli* VF panel is available in the GitHub repository ecoli_VF_collection, https://github.com/aleimba/ecoli_VF_collection [41]. The R scripts used in this study can be found in Additional file 4: Dataset S1 and Additional file 14: Dataset S7. R packages used are mentioned in the methods chapter. Perl scripts are stored in GitHub repository bac-genomics-scripts, https://github.com/aleimba/bac-genomics-scripts [82]. These are licensed under GNU GPLv3. Many of these depend on the BioPerl (v1.006923) module collection [106]. All other data sets supporting the results of this article are included within the article and its additional files.

Authors' contributions

Conceptualization: AL UD. Data curation: AL AP. Formal analysis: AL JV DG. Funding acquisition: UD. Investigation: AL. Methodology: AL. Project administration: AL. Resources: AL RD UD. Software: AL. Supervision: AL UD. Validation: AL. Visualization: AL. Writing – original draft: AL. Writing – review & editing: AL UD JV RD AP DG. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

This study did not involve experimental research on animals or animal material. Genome sequences included in this study have either been retrieved from NCBI's Genbank database or generated from bacterial isolates previously collected for diagnostic purposes. Approval by an ethics board was not required.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Institute of Hygiene, University of Münster, Mendelstrasse 7, 48149 Münster, Germany. [2]Department of Genomic and Applied Microbiology, Göttingen Genomics Laboratory, Institute of Microbiology and Genetics, Georg-August-University of Göttingen, Göttingen, Germany. [3]Institute for Molecular Infection Biology, Julius-Maximilians-University of Würzburg, Würzburg, Germany. [4]Leibniz Institute DSMZ, German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany. [5]Institute of Biostatistics and Clinical Research, University of Münster, Münster, Germany.

Received: 27 December 2016 Accepted: 27 April 2017
Published online: 08 May 2017

References

1. Hogeveen H, Huijps K, Lam TJ. Economic aspects of mastitis: new developments. N Z Vet J. 2011;59(1):16–23.
2. Petzl W, Zerbe H, Günther J, Yang W, Seyfert HM, Nürnberg G, et al. *Escherichia coli*, but not *Staphylococcus aureus* triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow. Vet Res. 2008;39(2):18.
3. Shpigel NY, Elazar S, Rosenshine I. Mammary pathogenic *Escherichia coli*. Curr Opin Microbiol. 2008;11(1):60–5.
4. Dogan B, Klaessig S, Rishniw M, Almeida RA, Oliver SP, Simpson K, et al. Adherent and invasive *Escherichia coli* are associated with persistent bovine mastitis. Vet Microbiol. 2006;116(4):270–82.
5. Zadoks RN, Middleton JR, McDougall S, Katholm J, Schukken YH. Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans. J Mammary Gland Biol Neoplasia. 2011;16(4):357–72.
6. Dogan B, Rishniw M, Bruant G, Harel J, Schukken YH, Simpson KW. Phylogroup and lpfA influence epithelial invasion by mastitis associated *Escherichia coli*. Vet Microbiol. 2012;159(1–2):163–70.
7. Burvenich C, Van Merris V, Mehrzad J, Diez-Fraile A, Duchateau L. Severity of *E. coli* mastitis is mainly determined by cow factors. Vet Res. 2003;34(5):521–64.
8. Porcherie A, Cunha P, Trotereau A, Roussel P, Gilbert FB, Rainard P, et al. Repertoire of *Escherichia coli* agonists sensed by innate immunity receptors of the bovine udder and mammary epithelial cells. Vet Res. 2012;43:14.

9. Houser BA, Donaldson SC, Padte R, Sawant AA, DebRoy C, Jayarao BM. Assessment of phenotypic and genotypic diversity of *Escherichia coli* shed by healthy lactating dairy cattle. Foodborne Pathog Dis. 2008;5(1):41–51.
10. Blum S, Heller ED, Krifucks O, Sela S, Hammer-Muntz O, Leitner G. Identification of a bovine mastitis *Escherichia coli* subset. Vet Microbiol. 2008;132(1–2):135–48.
11. Blum SE, Leitner G. Genotyping and virulence factors assessment of bovine mastitis *Escherichia coli*. Vet Microbiol. 2013;163(3–4):305–12.
12. Tenaillon O, Skurnik D, Picard B, Denamur E. The population genetics of commensal *Escherichia coli*. Nat Rev Microbiol. 2010;8(3):207–17.
13. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 2006;60(5):1136–51.
14. Chaudhuri RR, Henderson IR. The evolution of the *Escherichia coli* phylogeny. Infect Genet Evol. 2012;12(2):214–26.
15. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genet. 2009;5(1):e1000344.
16. Croxen MA, Finlay BB. Molecular mechanisms of *Escherichia coli* pathogenicity. Nat Rev Microbiol. 2010;8(1):26–38.
17. Köhler CD, Dobrindt U. What defines extraintestinal pathogenic *Escherichia coli*? Int J Med Microbiol. 2011;301(8):642–7.
18. Leimbach A, Hacker J, Dobrindt U. *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. Curr Top Microbiol Immunol. 2013;358:3–32.
19. Le Gall T, Clermont O, Gouriou S, Picard B, Nassif X, Denamur E, et al. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. Mol Biol Evol. 2007;24(11):2373–84.
20. Kaas RS, Friis C, Ussery DW, Aarestrup FM. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes. BMC Genomics. 2012;13:577.
21. Dobrindt U, Hochhut B, Hentschel U, Hacker J. Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol. 2004;2(5):414–24.
22. Fernandes JB, Zanardo LG, Galvao NN, Carvalho IA, Nero LA, Moreira MA. *Escherichia coli* from clinical mastitis: serotypes and virulence factors. J Vet Diagn Investig. 2011;23(6):1146–52.
23. Suojala L, Pohjanvirta T, Simojoki H, Myllyniemi AL, Pitkala A, Pelkonen S, et al. Phylogeny, virulence factors and antimicrobial susceptibility of *Escherichia coli* isolated in clinical bovine mastitis. Vet Microbiol. 2011;147(3–4):383–8.
24. Wenz JR, Barrington GM, Garry FB, Ellis RP, Magnuson RJ. *Escherichia coli* isolates' serotypes, genotypes, and virulence genes and clinical coliform mastitis severity. J Dairy Sci. 2006;89(9):3408–12.
25. Blum SE, Heller ED, Sela S, Elad D, Edery N, Leitner G. Genomic and phenomic study of mammary pathogenic *Escherichia coli*. PLoS One. 2015;10(9):e0136387.
26. Goldstone RJ, Harris S, Smith DG. Genomic content typifying a prevalent clade of bovine mastitis-associated *Escherichia coli*. Sci Rep. 2016;6:30115.
27. Kempf F, Slugocki C, Blum SE, Leitner G, Germon P. Genomic comparative study of bovine mastitis *Escherichia coli*. PLoS One. 2016;11(1):e0147954.
28. Richards VP, Lefebure T, Pavinski Bitar PD, Dogan B, Simpson KW, Schukken YH, et al. Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic *Escherichia coli*. PLoS One. 2015;10(3):e0119799.
29. Didelot X, Meric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics. 2012;13:256.
30. Leopold SR, Sawyer SA, Whittam TS, Tarr PI. Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*. BMC Evol Biol. 2011;11:183.
31. Rasko DA, Rosovitz MJ, Myers GS, Mongodin EF, Fricke WF, Gajer P, et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J Bacteriol. 2008;190(20):6881–93.
32. Łobocka MB, Rose DJ, Plunkett 3rd G, Rusin M, Samojedny A, Lehnherr H, et al. Genome of bacteriophage P1. J Bacteriol. 2004;186(21):7032–68.
33. Khachatryan AR, Besser TE, Call DR. The streptomycin-sulfadiazine-tetracycline antimicrobial resistance element of calf-adapted *Escherichia coli* is widely distributed among isolates from Washington state cattle. Appl Environ Microbiol. 2008;74(2):391–5.
34. Ziebell K, Johnson RP, Kropinski AM, Reid-Smith R, Ahmed R, Gannon VP, et al. Gene cluster conferring streptomycin, sulfonamide, and tetracycline resistance in *Escherichia coli* O157:H7 phage types 23, 45, and 67. Appl Environ Microbiol. 2011;77(5):1900–3.
35. Feng L, Liu B, Liu Y, Ratiner YA, Hu B, Li D, et al. A genomic islet mediates flagellar phase variation in *Escherichia coli* strains carrying the flagellin-specifying locus flk. J Bacteriol. 2008;190(13):4470–7.
36. Liu B, Hu B, Zhou Z, Guo D, Guo X, Ding P, et al. A novel non-homologous recombination-mediated mechanism for *Escherichia coli* unilateral flagellar phase variation. Nucleic Acids Res. 2012;40(10):4530–8.
37. Lügering A, Benz I, Knochenhauer S, Ruffing M, Schmidt MA. The Pix pilus adhesin of the uropathogenic *Escherichia coli* strain X2194 (O2:K(–):H6) is related to Pap pili but exhibits a truncated regulatory region. Microbiology. 2003;149(Pt 6):1387–97.
38. Schneider G, Dobrindt U, Brüggemann H, Nagy G, Janke B, Blum-Oehler G, et al. The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536. Infect Immun. 2004;72(10):5993–6001.
39. Barondess JJ, Beckwith J. bor gene of phage lambda, involved in serum resistance, encodes a widely conserved outer membrane lipoprotein. J Bacteriol. 1995;177(5):1247–53.
40. Johnson TJ, Wannemuehler YM, Nolan LK. Evolution of the *iss* gene in *Escherichia coli*. Appl Environ Microbiol. 2008;74(8):2360–9.
41. Leimbach A. ecoli_VF_collection: v0.1. Zenodo. 2016. http://dx.doi.org/10.5281/zenodo.56686.
42. Croxen MA, Law RJ, Scholz R, Keeney KM, Wlodarska M, Finlay BB. Recent advances in understanding enteric pathogenic *Escherichia coli*. Clin Microbiol Rev. 2013;26(4):822–80.
43. Ren CP, Beatson SA, Parkhill J, Pallen MJ. The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*. J Bacteriol. 2005;187(4):1430–40.
44. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D, et al. Genomics. Genome project standards in a new era of sequencing. Science. 2009;326(5950):236–7.
45. Leimbach A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R, et al. Complete genome sequences of *Escherichia coli* Strains 1303 and ECC-1470 isolated from Bovine Mastitis. Genome Announc. 2015;3(2):e00182–15.
46. Yi H, Cho YJ, Hur HG, Chun J. Genome sequence of *Escherichia coli* AA86, isolated from cow feces. J Bacteriol. 2011;193(14):3681.
47. Cooper KK, Mandrell RE, Louie JW, Korlach J, Clark TA, Parker CT, et al. Comparative genomics of enterohemorrhagic *Escherichia coli* O145:H28 demonstrates a common evolutionary lineage with *Escherichia coli* O157:H7. BMC Genomics. 2014;15:17.
48. Sváb D, Horváth B, Szucs A, Maróti G, Tóth I. Draft Genome Sequence of an *Escherichia coli* O157:H43 Strain Isolated from Cattle. Genome Announc. 2013;1(3):e00263–13.
49. Johnson TJ, Wannemuehler Y, Johnson SJ, Stell AL, Doetkott C, Johnson JR, et al. Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens. Appl Environ Microbiol. 2008;74(22):7043–50.
50. Liu Y, Liu G, Liu W, Ali T, Chen W, Yin J, et al. Phylogenetic group, virulence factors and antimicrobial resistance of *Escherichia coli* associated with bovine mastitis. Res Microbiol. 2014;165(4):273–7.
51. Tourret J, Denamur E. Population phylogenomics of extraintestinal pathogenic *Escherichia coli*. Microbiol Spectr. 2016;4:1.
52. Bradley A. Bovine mastitis: an evolving disease. Vet J. 2002;164(2):116–28.
53. Kornalijnslijper JE, Daemen AJ, van Werven T, Niewold TA, Rutten VP, Noordhuizen-Stassen EN. Bacterial growth during the early phase of infection determines the severity of experimental *Escherichia coli* mastitis in dairy cows. Vet Microbiol. 2004;101(3):177–86.
54. Boulanger V, Bouchard L, Zhao X, Lacasse P. Induction of nitric oxide production by bovine mammary epithelial cells and blood leukocytes. J Dairy Sci. 2001;84(6):1430–7.
55. Kaipainen T, Pohjanvirta T, Shpigel NY, Shwimmer A, Pyorala S, Pelkonen S. Virulence factors of *Escherichia coli* isolated from bovine clinical mastitis. Vet Microbiol. 2002;85(1):37–46.
56. Döpfer D, Almeida RA, Lam TJ, Nederbragt H, Oliver SP, Gaastra W. Adhesion and invasion of *Escherichia coli* from single and recurrent clinical cases of bovine mastitis in vitro. Vet Microbiol. 2000;74(4):331–43.
57. Günther J, Koy M, Berthold A, Schuberth HJ, Seyfert HM. Comparison of the pathogen species-specific immune response in udder derived cell types and their models. Vet Res. 2016;47:22.
58. Schukken YH, Günther J, Fitzpatrick J, Fontaine MC, Goetze L, Holst O, et al. Host-response patterns of intramammary infections in dairy cows. Vet Immunol Immunopathol. 2011;144(3–4):270–89.

59. Fairbrother JH, Dufour S, Fairbrother JM, Francoz D, Nadeau E, Messier S. Characterization of persistent and transient *Escherichia coli* isolates recovered from clinical mastitis episodes in dairy cows. Vet Microbiol. 2015;176(1–2):126–33.
60. Ghanbarpour R, Oswald E. Phylogenetic distribution of virulence genes in *Escherichia coli* isolated from bovine mastitis in Iran. Res Vet Sci. 2010;88(1):6–10.
61. Lehtolainen T, Pohjanvirta T, Pyorala S, Pelkonen S. Association between virulence factors and clinical course of *Escherichia coli* mastitis. Acta Vet Scand. 2003;44(3–4):203–5.
62. Huja S, Oren Y, Trost E, Brzuszkiewicz E, Biran D, Blom J, et al. Genomic avenue to avian colisepticemia. MBio. 2015;6(1):e01681–14.
63. Johnson TJ, Siek KE, Johnson SJ, Nolan LK. DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains. J Bacteriol. 2006;188(2):745–58.
64. Cheng D, Zhu S, Su Z, Zuo W, Lu H. Prevalence and isoforms of the pathogenicity island ETT2 among *Escherichia coli* isolates from colibacillosis in pigs and mastitis in cows. Curr Microbiol. 2012;64(1):43–9.
65. Yao Y, Xie Y, Perace D, Zhong Y, Lu J, Tao J, et al. The type III secretion system is involved in the invasion and intracellular survival of *Escherichia coli* K1 in human brain microvascular endothelial cells. FEMS Microbiol Lett. 2009;300(1):18–24.
66. Li J, Yao Y, Xu HH, Hao L, Deng Z, Rajakumar K, et al. SecReT6: a web-based resource for type VI secretion systems found in bacteria. Environ Microbiol. 2015;17(7):2196–202.
67. Journet L, Cascales E. The type VI secretion system in *Escherichia coli* and related species. EcoSal Plus. 2016;7:1.
68. Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W, et al. Whole-genome draft sequences of six commensal fecal and six mastitis-associated *Escherichia coli* strains of Bovine Origin. Genome Announc. 2016;4(4):e00753–16.
69. Rainard P, Riollet C. Innate immunity of the bovine mammary gland. Vet Res. 2006;37(3):369–400.
70. Vangroenweghe F, Rainard P, Paape M, Duchateau L, Burvenich C. Increase of *Escherichia coli* inoculum doses induces faster innate immune response in primiparous cows. J Dairy Sci. 2004;87(12):4132–44.
71. Falentin H, Rault L, Nicolas A, Bouchard DS, Lassalas J, Lamberton P, et al. Bovine teat Microbiome analysis revealed reduced alpha diversity and significant changes in taxonomic profiles in quarters with a history of Mastitis. Front Microbiol. 2016;7:480.
72. Ganda EK, Bisinotto RS, Lima SF, Kronauer K, Decter DH, Oikonomou G, et al. Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin. Sci Rep. 2016;6:37565.
73. Oikonomou G, Machado VS, Santisteban C, Schukken YH, Bicalho RC. Microbial diversity of bovine mastitic milk as described by pyrosequencing of metagenomic 16 s rDNA. PLoS One. 2012;7(10):e47671.
74. Bouchard DS, Seridan B, Saraoui T, Rault L, Germon P, Gonzalez-Moreno C, et al. Lactic acid bacteria isolated from bovine mammary microbiota: potential allies against bovine Mastitis. PLoS One. 2015;10(12):e0144831.
75. Westermann AJ, Förstner KU, Amman F, Barquist L, Chao Y, Schulte LN, et al. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. Nature. 2016;529(7587):496–501.
76. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–80.
77. Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. 1999. German conference on bioinformatics. bioinfo.de. Available from: http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html
78. Staden R, Beal KF, Bonfield JK. The Staden package, 1998. Methods Mol Biol. 2000;132:115–30.
79. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77.
80. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9.
81. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944–5.
82. Leimbach A. bac-genomics-scripts: Bovine *E. coli* mastitis comparative genomics edition. Zenodo. 2016. http://dx.doi.org/10.5281/zenodo.215824.
83. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the artemis comparison tool. Bioinformatics. 2005;21(16):3422–3.
84. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf. 2009;10:421.
85. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27(3):334–42.
86. Sahl JW, Matalka MN, Rasko DA. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl Environ Microbiol. 2012;78(14):4884–92.
87. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
88. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61(6):1061–7.
89. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinf. 2012;13:87.
90. Francisco AP, Bugalho M, Ramirez M, Carrico JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinf. 2009;10:152.
91. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F, et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 2015;43(W1):W104–8.
92. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinf. 2006;7:142.
93. Langille MG, Hsiao WW, Brinkman FS. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinf. 2008;9:329.
94. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39(Web Server issue):W347–52.
95. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy *In Silico* Serotyping of *Escherichia coli* Isolates by use of whole-genome sequencing data. J Clin Microbiol. 2015;53(8):2410–26.
96. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinf. 2011;12:124.
97. Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thevenin A, Stoye J, et al. Orthology detection combining clustering and synteny for very large datasets. PLoS One. 2014;9(8):e105015.
98. Wickham H. Elegant graphics for data analysis. New York: Springer; 2009.
99. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33(Database issue):D325–8.
100. Yang J, Chen L, Sun L, Yu J, Jin Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res. 2008;36:D539–42.
101. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67(11):2640–4.
102. Neuwirth E. RColorBrewer: ColorBrewer palettes. 2014.
103. Popescu AA, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics. 2012;28(11):1536–7.
104. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27(7):1009–10.
105. R Core Team. R: A language and environment fro statistical computing. Vienna: R Foundation for Statistical Computing; 2016.
106. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, et al. The Bioperl toolkit: perl modules for the life sciences. Genome Res. 2002;12(10):1611–8.
107. Kim M, Yi H, Cho YJ, Jang J, Hur HG, Chun J. Draft genome sequence of *Escherichia coli* W26, an enteric strain isolated from cow feces. J Bacteriol. 2012;194(18):5149–50.

5.3.4.3 *Supplementary information*

Leimbach et al. (2017) has extensive supplementary information – of these Figures S1–S7 (Additional file 3) are shown on pages 164–175. Also the detailed supplementary materials and methods section (Additional file 21) is included on pages 176–191. However, all supplementary tables (Tables S1–S7; Additional files 1–2, 5, 9, 12–13, and 15) and datasets (Datasets S1–S15; Additional files 4, 6–8, 10–11, 14, 16–20, and 22–23) are either too extensive or not suitable for a reprint here. The content of these files is described on pages 159–160 and can be found with the original source files on the BMC Genomics manuscript website:
https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-017-3739-x

Alternatively, the bioRχiv server hosts the identical supplemental files:
http://biorxiv.org/content/early/2017/04/21/096479.figures-only

Figure S1

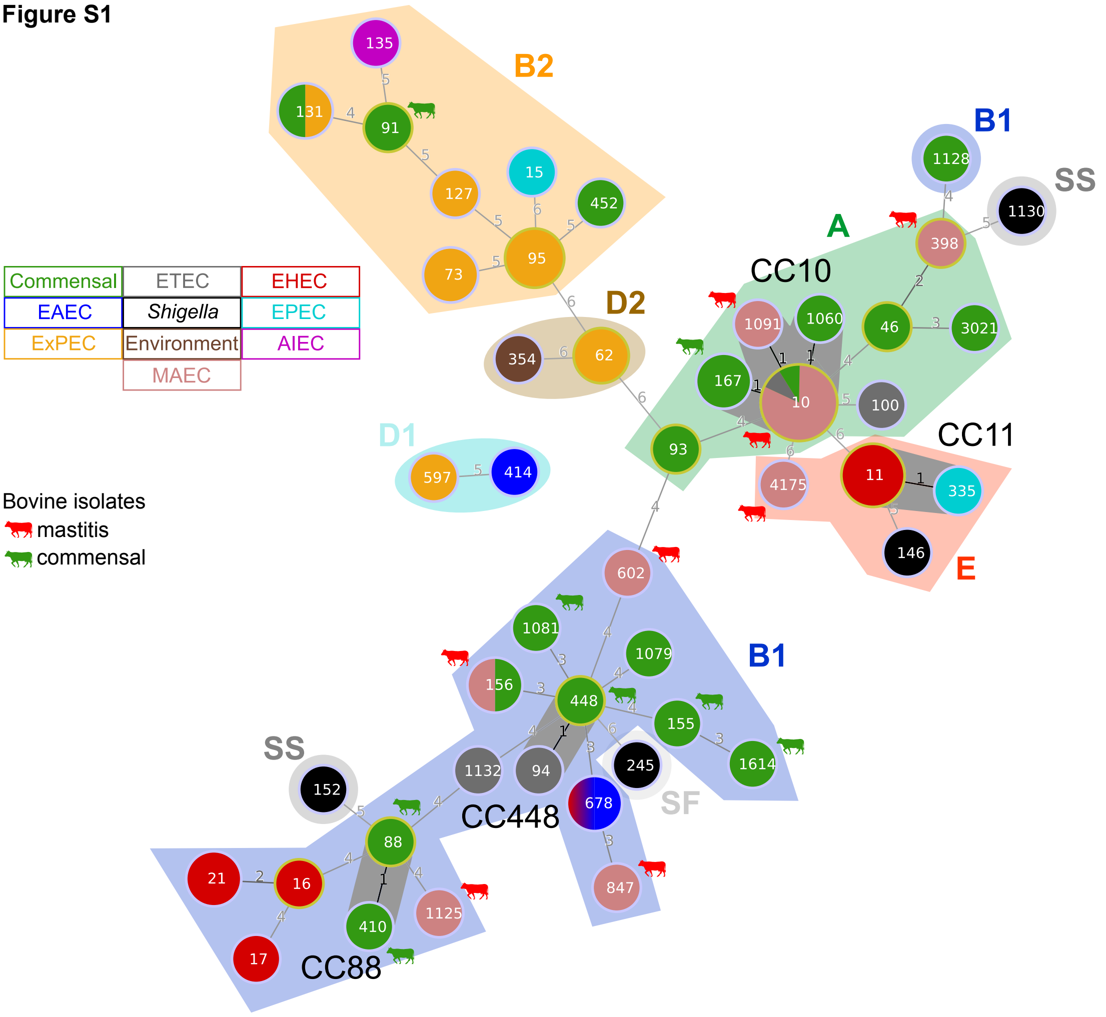


Figure S1. Minimum spanning tree (MST) of the multi-locus sequence typing (MLST) results generated with the goeBURST algorithm and visualized with PHYLOViZ. Sequence types (STs) are depicted by circles (nodes) with the respective ST number. The size of each ST node is proportional to the number of isolates allocated to the ST. Additionally, ST nodes are colored pie charts indicating the proportion of included pathotypes. Number of variants between each ST node are indicated on the edges. Clonal complexes (CCs) were designated for STs differing only by a single locus/allele (single locus variant, SLV). The STs belonging to a CC are highlighted by gray vertices and the CCs are named by the corresponding founder ST. Colored cows indicate STs with MAEC and commensal isolates. ECOR phylogenetic groups are highlighted by colored fields.

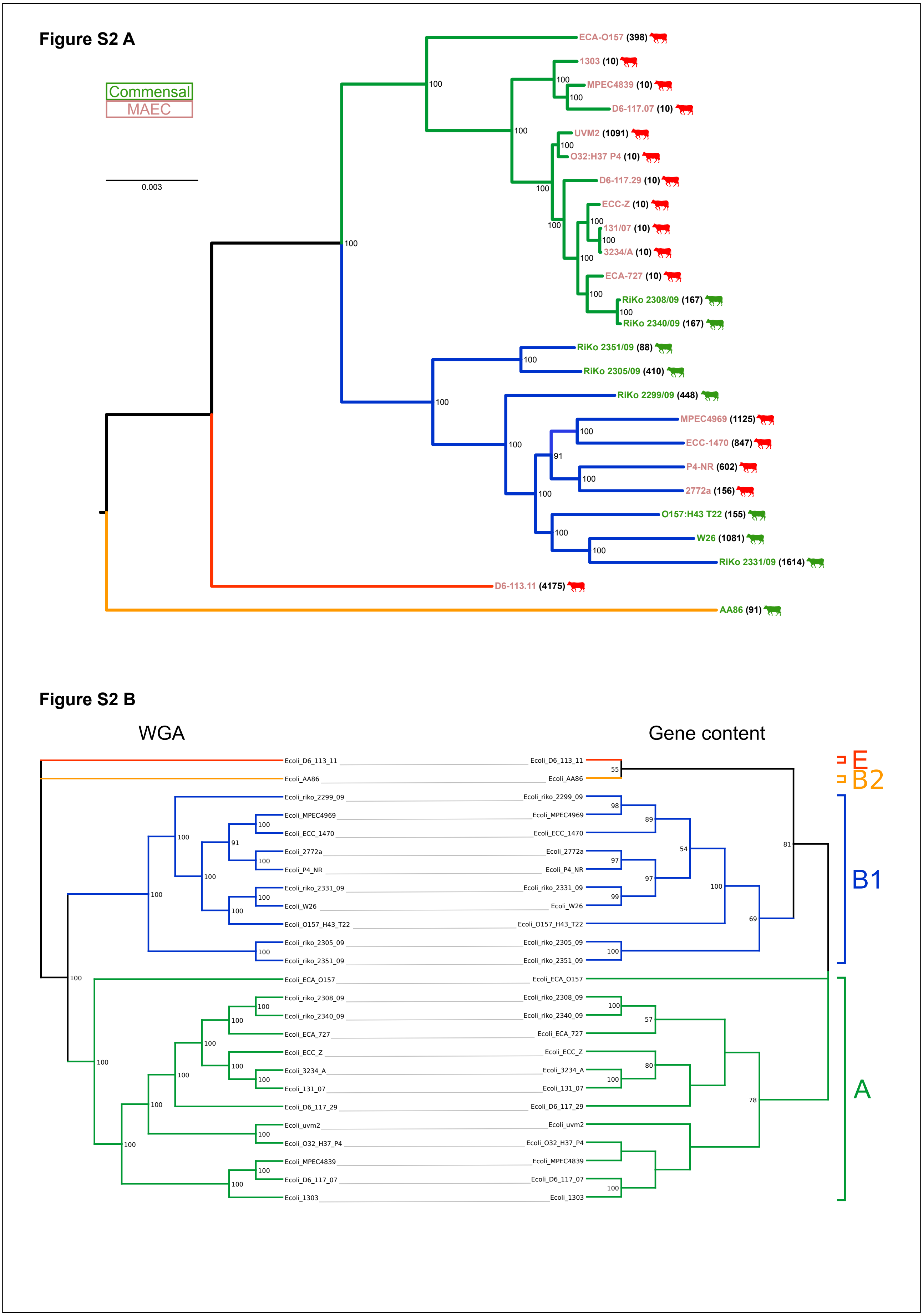


Figure S2 A
Commensal
MAEC
0.003
ECA-O157 (398)
1303 (10)
MPEC4839 (10)
D6-117.07 (10)
UVM2 (1091)
O32:H37 P4 (10)
D6-117.29 (10)
ECC-Z (10)
131/07 (10)
3234/A (10)
ECA-727 (10)
RiKo 2308/09 (167)
RiKo 2340/09 (167)
RiKo 2351/09 (88)
RiKo 2305/09 (410)
RiKo 2299/09 (448)
MPEC4969 (1125)
ECC-1470 (847)
P4-NR (602)
2772a (156)
O157:H43 T22 (155)
W26 (1081)
RiKo 2331/09 (1614)
D6-113.11 (4175)
AA86 (91)
100
91
Figure S2 B
WGA
Gene content
Ecoli_D6_113_11
Ecoli_AA86
Ecoli_riko_2299_09
Ecoli_MPEC4969
Ecoli_ECC_1470
Ecoli_2772a
Ecoli_P4_NR
Ecoli_riko_2331_09
Ecoli_W26
Ecoli_O157_H43_T22
Ecoli_riko_2305_09
Ecoli_riko_2351_09
Ecoli_ECA_O157
Ecoli_riko_2308_09
Ecoli_riko_2340_09
Ecoli_ECA_727
Ecoli_ECC_Z
Ecoli_3234_A
Ecoli_131_07
Ecoli_D6_117_29
Ecoli_uvm2
Ecoli_O32_H37_P4
Ecoli_MPEC4839
Ecoli_D6_117_07
Ecoli_1303
55
98
89
54
81
97
97
100
99
69
100
100
57
80
100
78
E
B2
B1
A

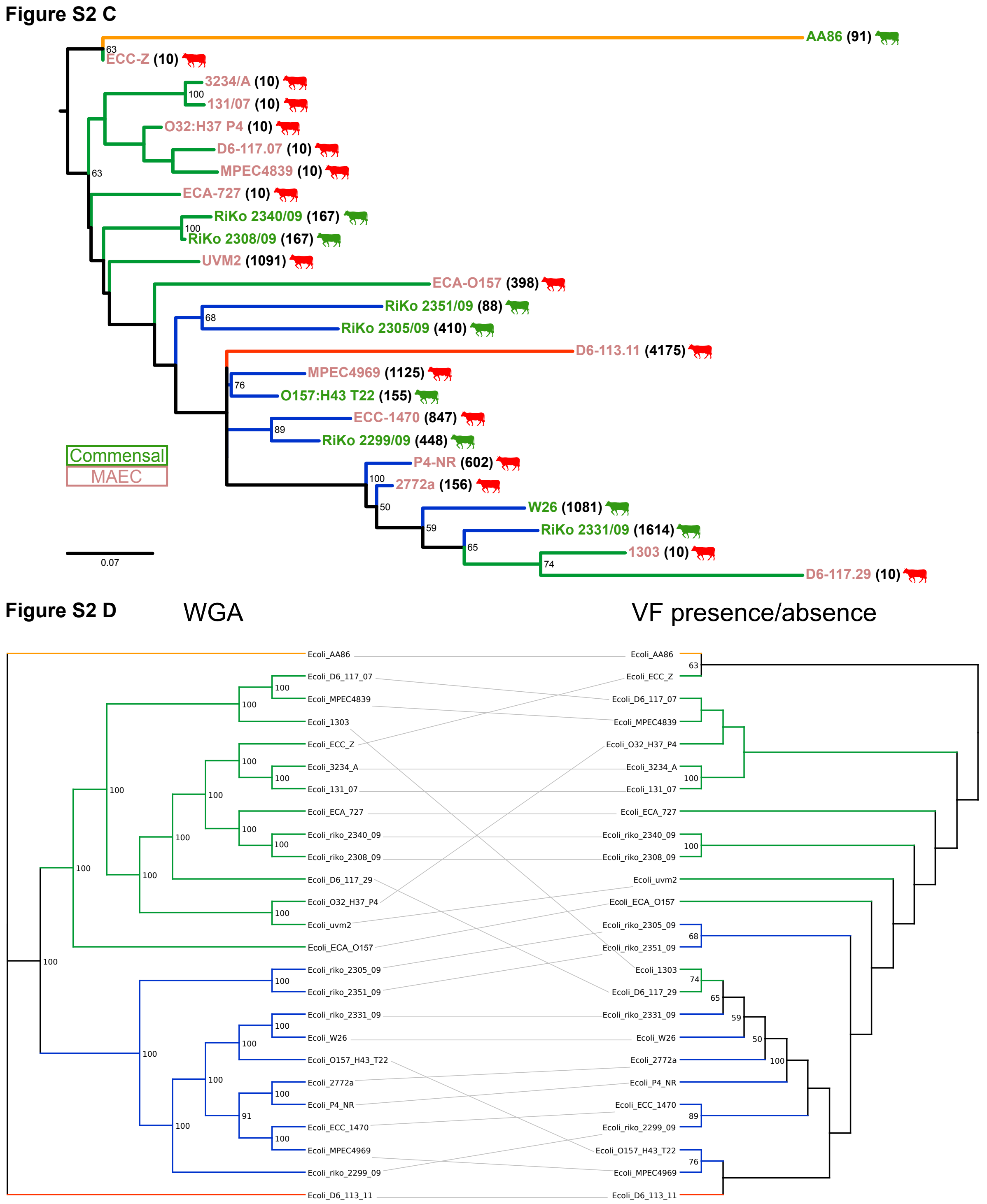


Figure S2. Phylograms and tanglegrams for the 25 bovine-associated *E. coli* genomes. ECOR phylogenetic groups are indicated by color: A (green), B1 (blue), E (red), and B2 (orange). The ST for each strain is given in parentheses. Bootstrap values below 50 were removed from all trees. **(A)** Whole genome alignment (WGA) phylogeny. The best scoring maximum likelihood (ML) phylogeny was inferred with RAxML's GTRGAMMA model from the core alignment length of 3,393,864 bp with 1,000 bootstrap resamplings. The tree was visualized with FigTree and midpoint rooted. **(B)** Tanglegram between the WGA genealogy and the gene content tree generated with Dendroscope. **(C)** Best scoring ML dendrogram (RAxML BINGAMMA) based on the presence/absence of 556 virulence-associated genes, with 1000 resamplings for bootstrap support values. The tree was visualized with FigTree and midpoint rooted. **(D)** Tanglegram between the WGA genealogy and the VF content tree generated with Dendroscope.

Figure S3 A

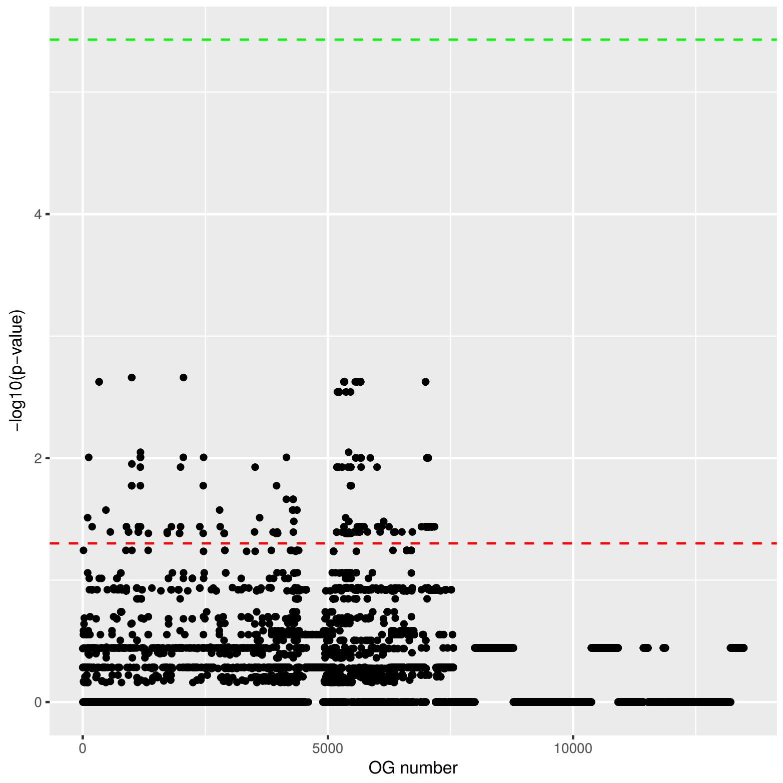


Figure S3 B

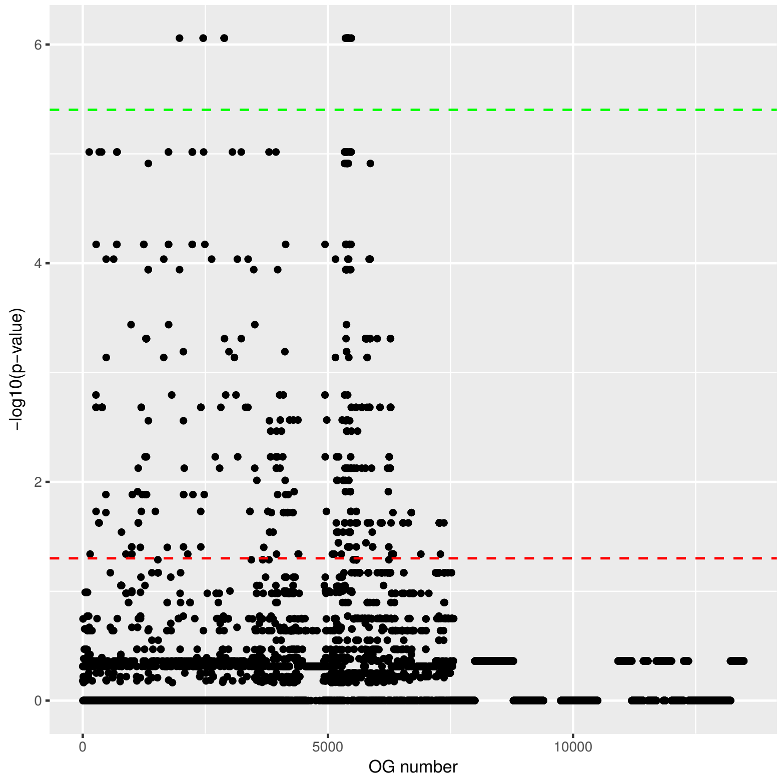


Figure S3 C

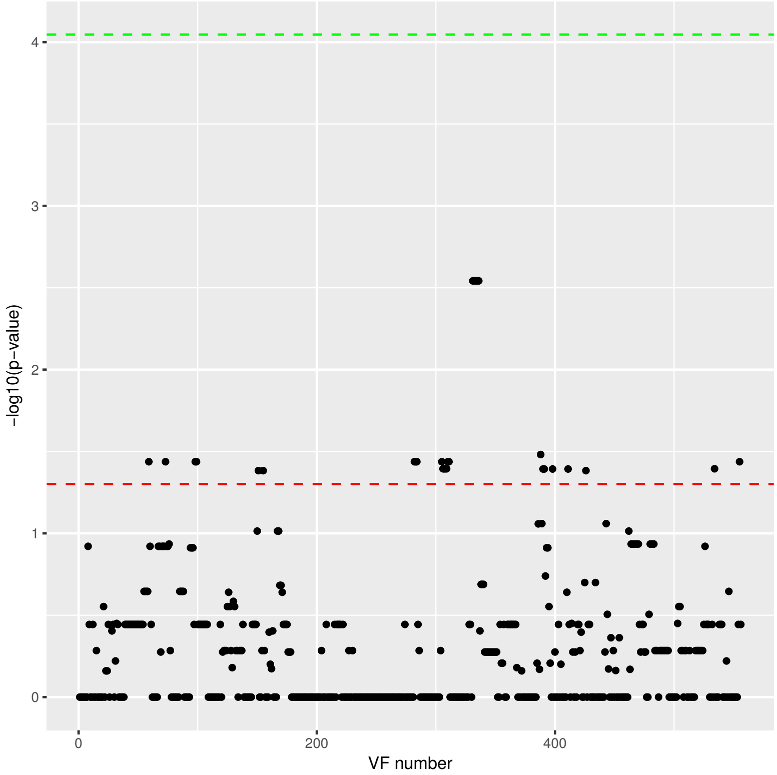


Figure S3 D

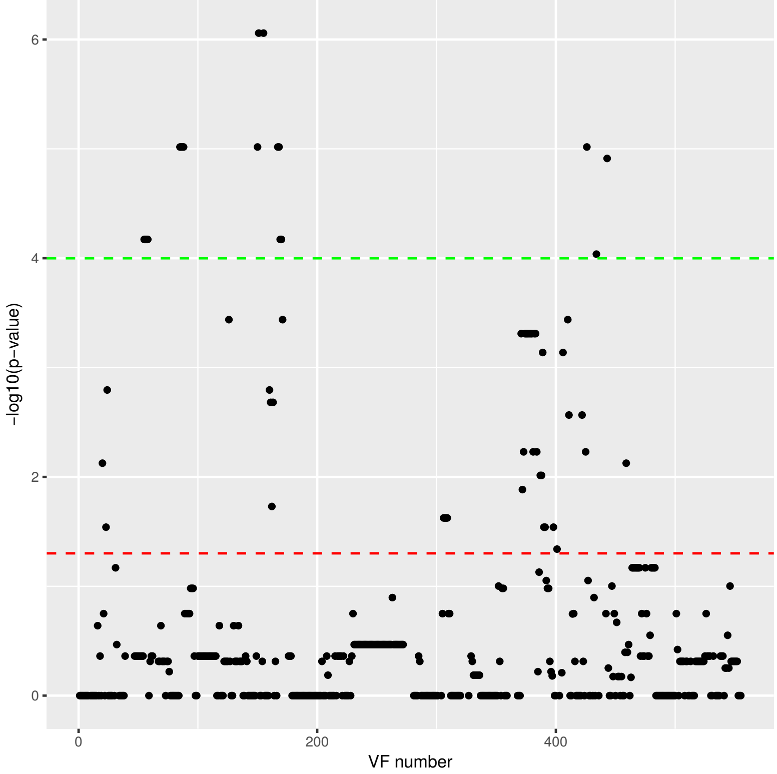


Figure S3. Manhattan plots of Fisher exact p-values for orthologous group (OG) or virulence factor (VF) associations of the 25 bovine-associated *E. coli* genomes. The red dotted line indicates the significance threshold (0.05), the green line the Bonferroni-corrected threshold. **(A)** OG association with pathotype (MAEC, commensal isolates) or **(B)** phylogroup (A, B1). **(C)** VF association with pathotype (MAEC, commensal) or **(D)** phylogroup (A, B1).

Figure S4

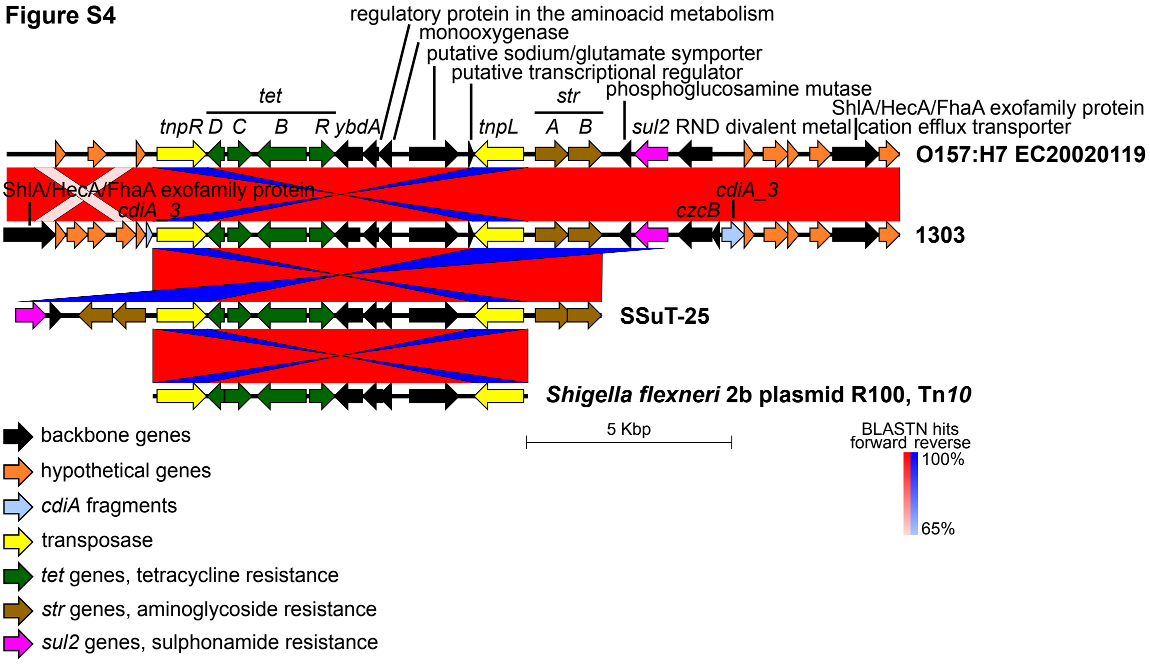

regulatory protein in the aminoacid metabolism
monooxygenase
putative sodium/glutamate symporter
putative transcriptional regulator
phosphoglucosamine mutase
ShlA/HecA/FhaA exofamily protein
tet
str
tnpR D C B R ybdA
tnpL A B
sul2 RND divalent metal cation efflux transporter
O157:H7 EC20020119
ShlA/HecA/FhaA exofamily protein
cdiA_3
cdiA_3
czcB
1303
SSuT-25
Shigella flexneri 2b plasmid R100, Tn10
backbone genes
hypothetical genes
cdiA fragments
transposase
tet genes, tetracycline resistance
str genes, aminoglycoside resistance
sul2 genes, sulphonamide resistance
5 Kbp
BLASTN hits
forward reverse
100%
65%

Figure S4. Gene organization of the AMR-SSuT/Tn*10* gene cluster. The comparison was done with the AMR-SSuT entities in strains MAEC 1303 (encoded on GI4), *E. coli* O157:H7 strain EC20020119, and *E. coli* strain SSuT-25, as well as Tn*10* of *Shigella flexneri* 2b plasmid R100. The diagram was created with Easyfig utilizing BLASTN+. Homologous regions are connected via red vertices for forward and blue vertices for inverted regions, and colored by nucleotide identity. Gene names are indicated above genomes encoding for these. Genes are colored according to their functions.

Figure S5 A

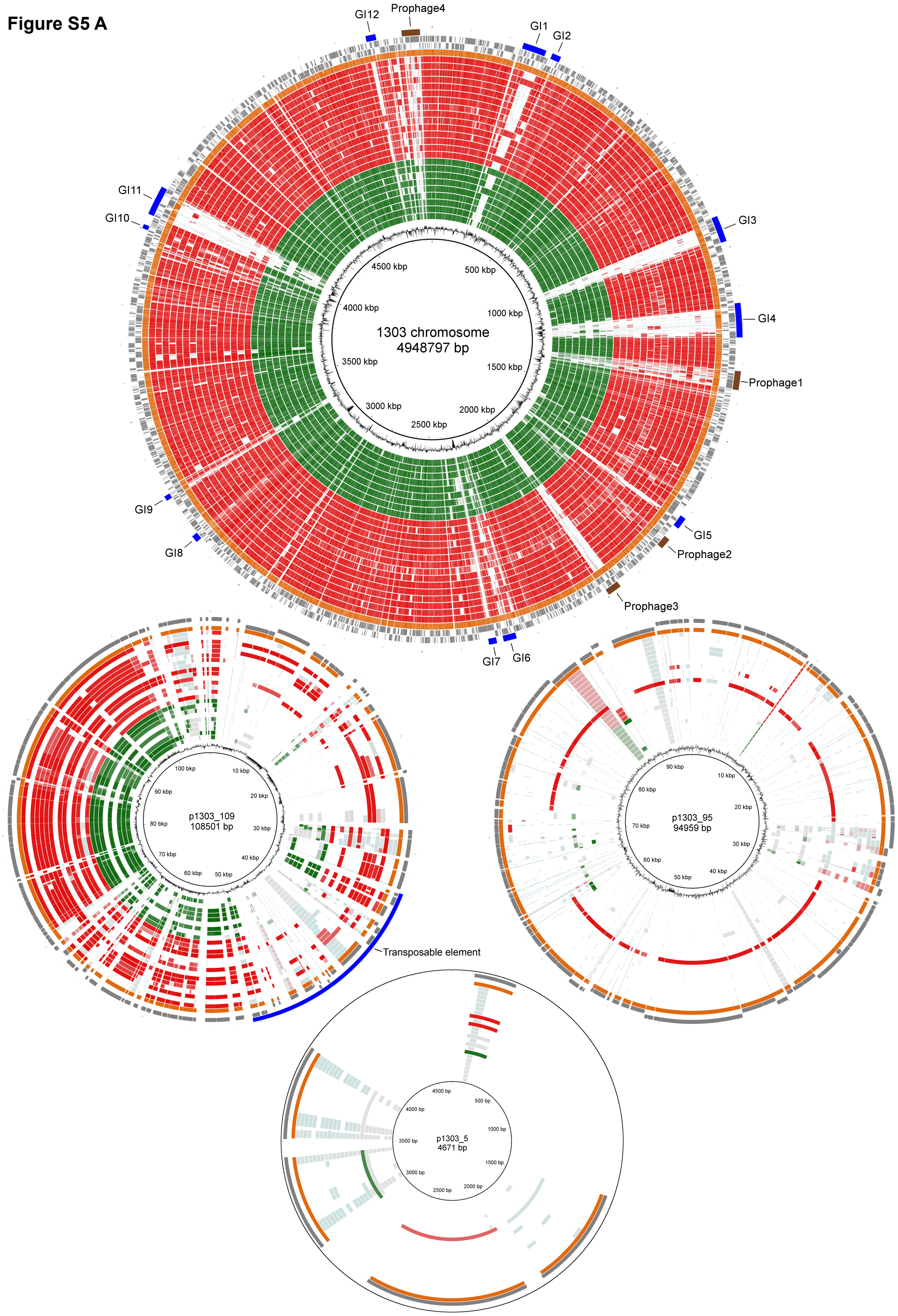

GI12
Prophage4
GI1
GI2
GI11
GI10
GI3
GI4
Prophage1
4500 kbp
500 kbp
4000 kbp
1000 kbp
1303 chromosome
4948797 bp
3500 kbp
1500 kbp
3000 kbp
2000 kbp
2500 kbp
GI9
GI8
GI5
Prophage2
Prophage3
GI7
GI6
100 bkp
10 kbp
90 kbp
20 bkp
p1303_109
108501 bp
80 bkp
30 kbp
70 kbp
40 kbp
60 kbp
50 kbp
Transposable element
90 kbp
10 kbp
80 kbp
20 kbp
p1303_95
94959 bp
70 kbp
30 kbp
60 kbp
40 kbp
50 kbp
4500 bp
500 bp
4000 bp
1000 bp
3500 bp
p1303_5
4671 bp
1500 bp
3000 bp
2000 bp
2500 bp

Figure S5 B

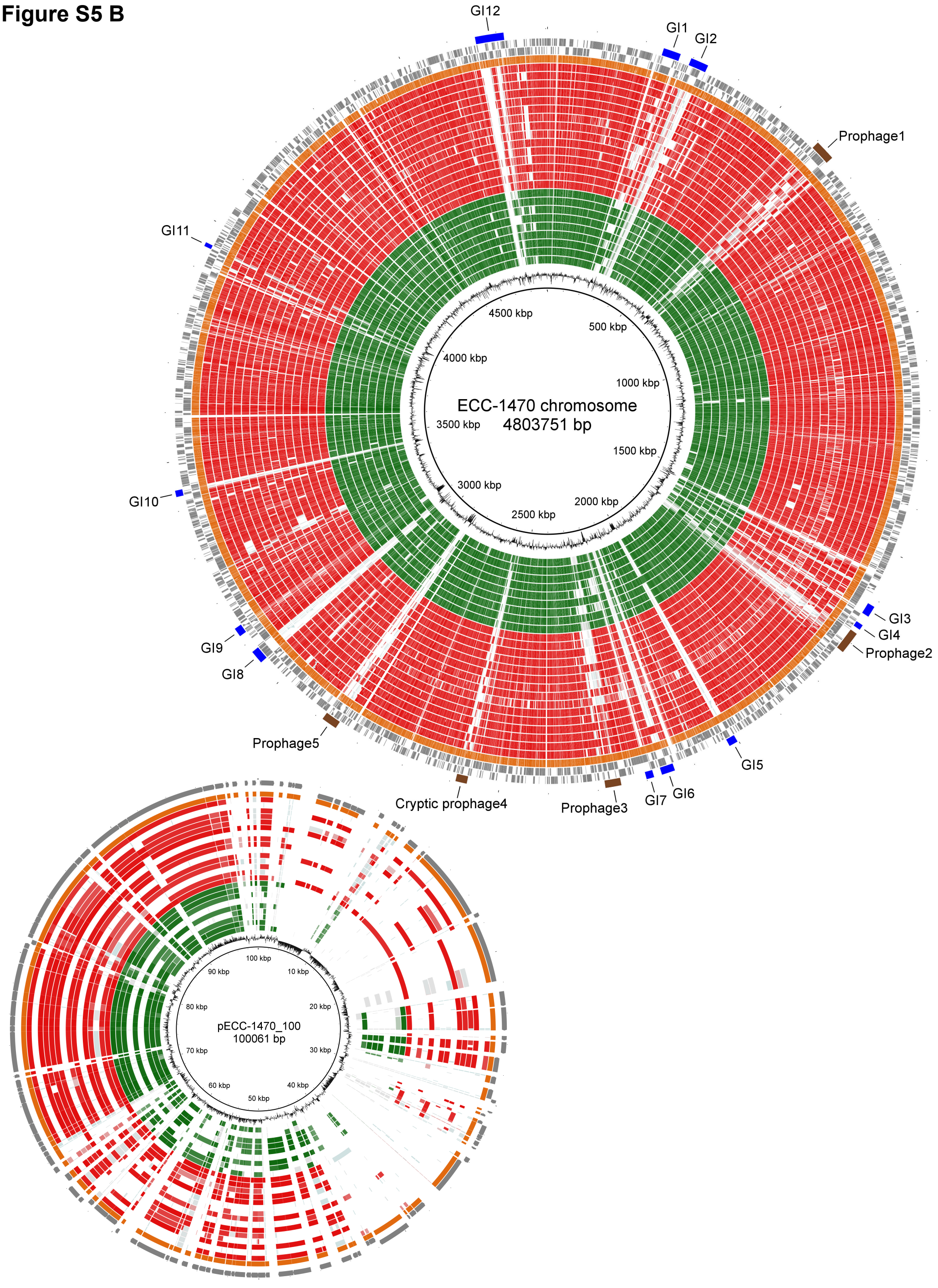


Figure S5. Circular genome diagrams created with BRIG. All figures have the following ring order (from inner to outer). **Black:** GC content. **Green** commensal strains: AA86, O157:H43 T22, RiKo 2299/09, RiKo 2305/09, RiKo 2308/09, RiKo 2331/09, RiKo 2340/09, RiKo 2351/09, and W26. **Red** MAEC: (1303 only in the ECC-1470 diagrams), 131/07, 2772a, 3234/A, D6-113.11, D6-117.07, D6-117.29, ECA-727, ECA-O157, (ECC-1470 only in the 1303 diagrams), ECC-Z, MPEC4839, MPEC4969, O32:H37 P4, P4-NR, and UVM2. **Orange:** The respective reference replicon from 1303 or ECC-1470. **Grey:** CDSs on the lagging and on the leading strand. **Blue or Brown:** Genomic islands/mobile elements or prophages. **(A)** Circular genome diagrams of all MAEC 1303 replicons: Chromosome, F-plasmid (p1303_109), circularized P1 bacteriophage (p1303_95), and small cryptic plasmid (p1303_5). **(B)** Circular genome diagrams of all MAEC ECC-1470 replicons: Chromosome and F-plasmid (pECC-1470_100).

Figure S6

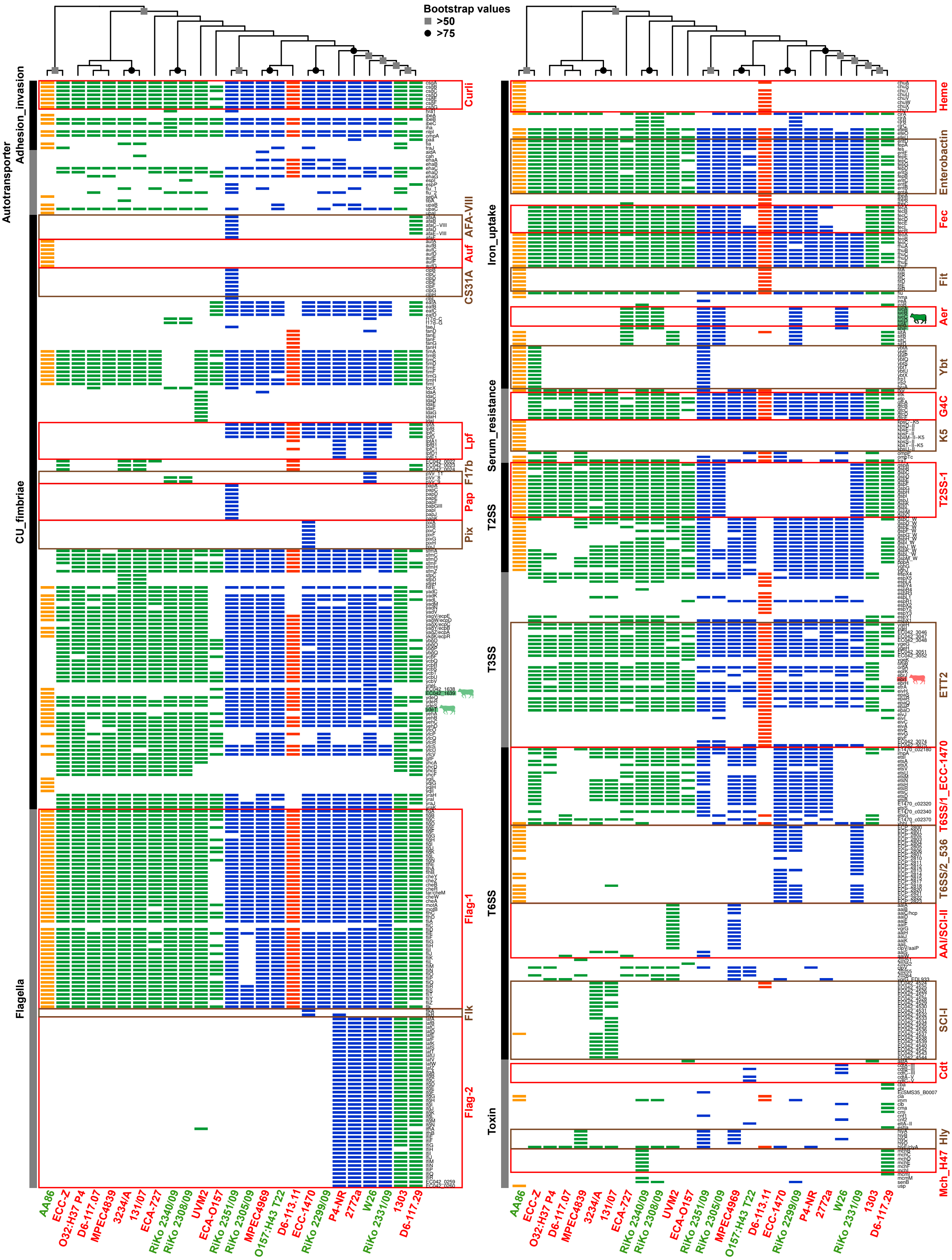
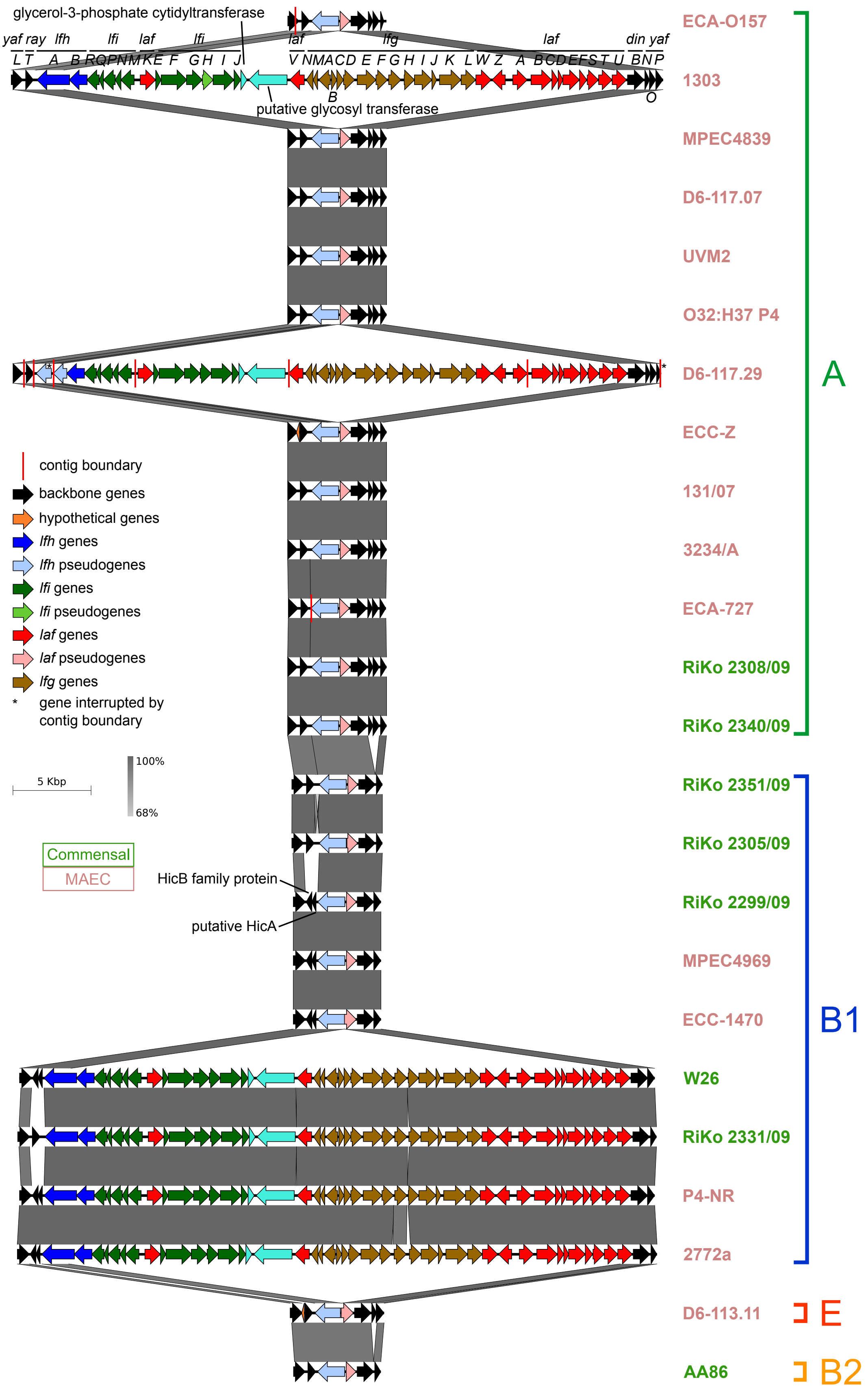


Figure S6. Heatmap of VF presence/absence. For the figure description see Figure 4. The only difference is the inclusion of virulence-associated gene names/locus tags.

Figure S7 A


glycerol-3-phosphate cytidyltransferase
yaf ray lfh lfi laf lfi
L T A B RQPNMKE F G H I J
laf lfg laf din yaf
V NMACD E F G H I J K L W Z A BCDEFSTU BNP
B
O
putative glycosyl transferase
ECA-O157
1303
MPEC4839
D6-117.07
UVM2
O32:H37 P4
D6-117.29
ECC-Z
131/07
3234/A
ECA-727
RiKo 2308/09
RiKo 2340/09
A
RiKo 2351/09
RiKo 2305/09
RiKo 2299/09
MPEC4969
ECC-1470
W26
RiKo 2331/09
P4-NR
2772a
B1
D6-113.11
E
AA86
B2
contig boundary
backbone genes
hypothetical genes
lfh genes
lfh pseudogenes
lfi genes
lfi pseudogenes
laf genes
laf pseudogenes
lfg genes
* gene interrupted by contig boundary
5 Kbp
100%
68%
Commensal
MAEC
HicB family protein
putative HicA

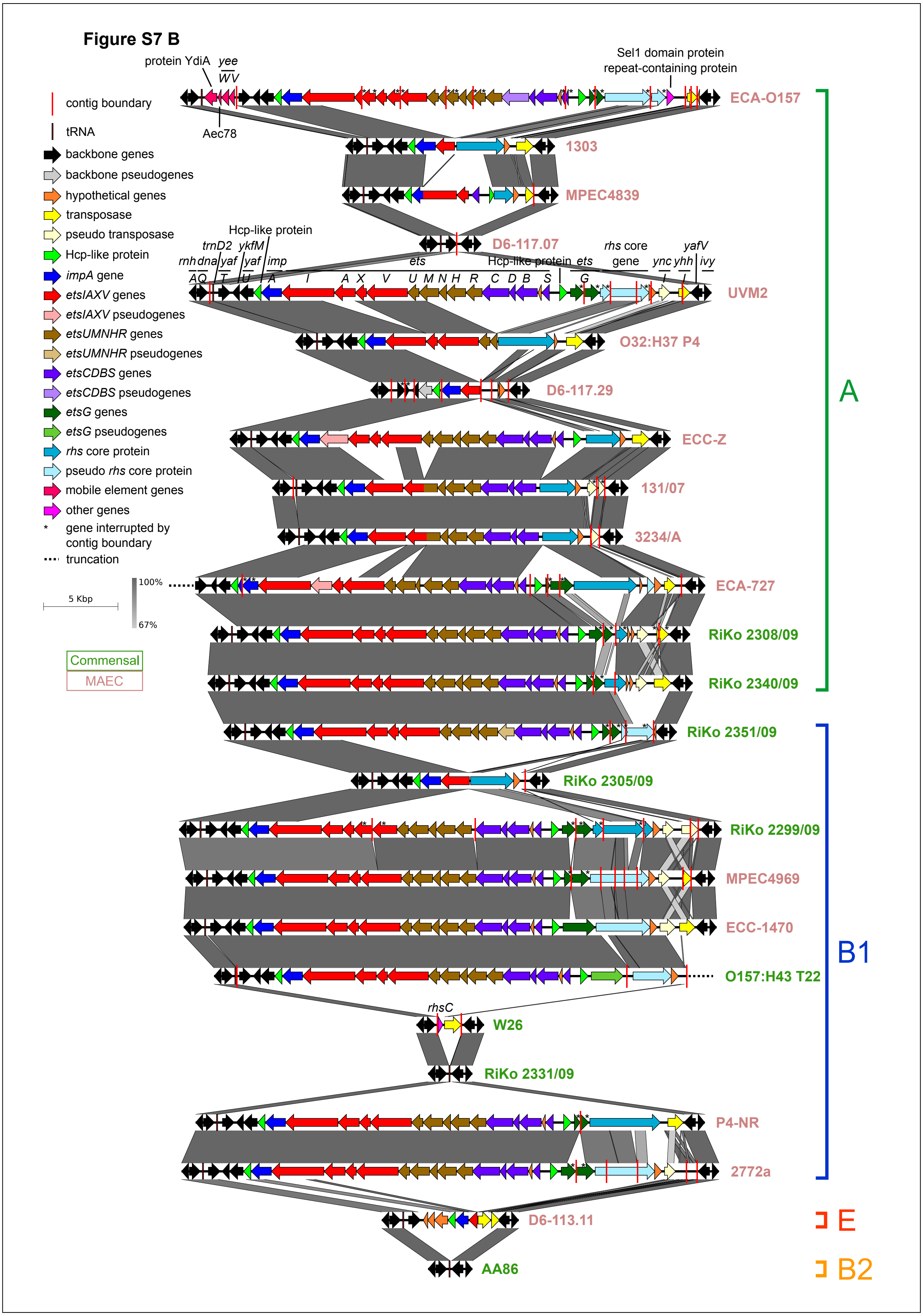
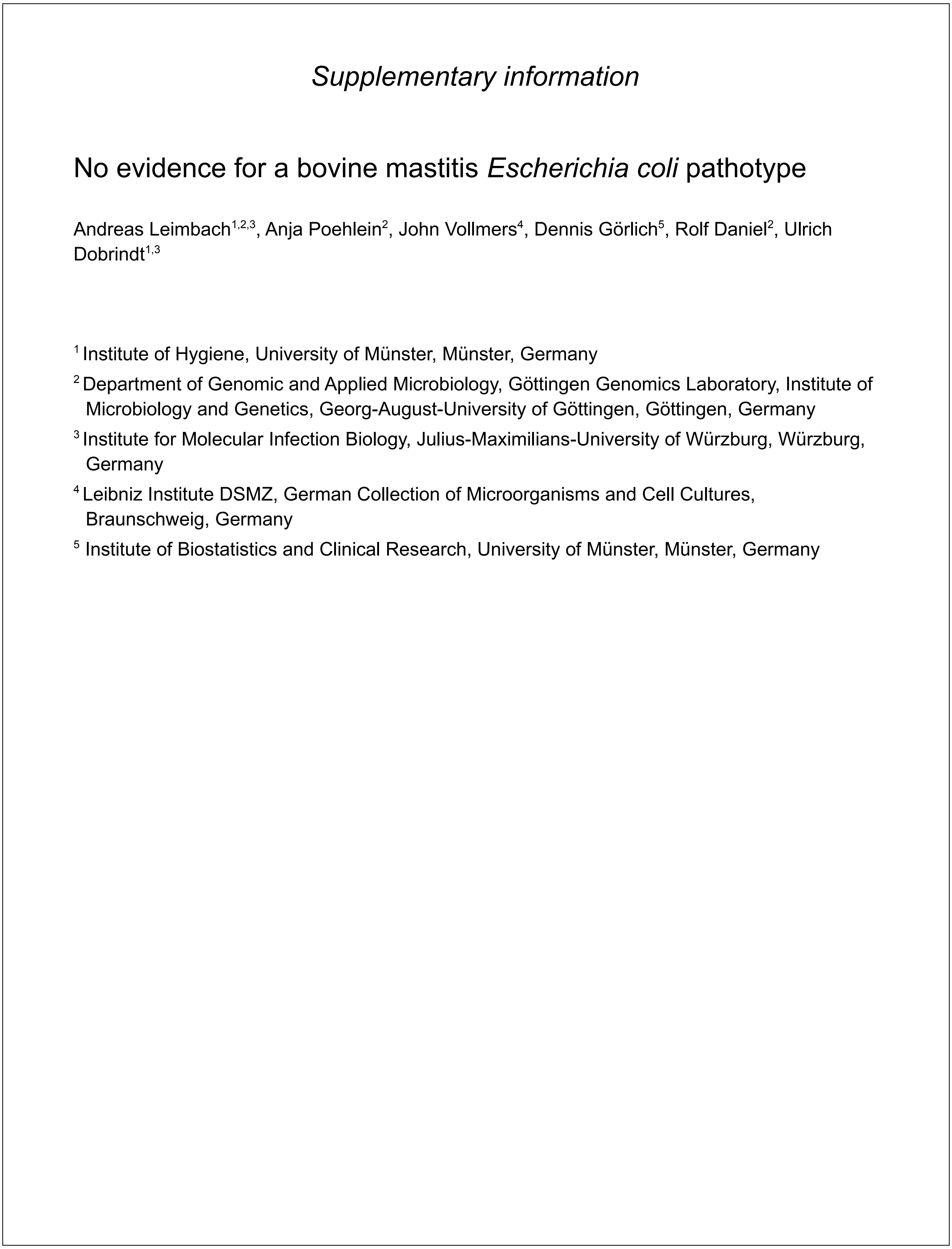


Figure S7 B

contig boundary
tRNA
backbone genes
backbone pseudogenes
hypothetical genes
transposase
pseudo transposase
Hcp-like protein
impA gene
etsIAXV genes
etsIAXV pseudogenes
etsUMNHR genes
etsUMNHR pseudogenes
etsCDBS genes
etsCDBS pseudogenes
etsG genes
etsG pseudogenes
rhs core protein
pseudo rhs core protein
mobile element genes
other genes
* gene interrupted by contig boundary
truncation
5 Kbp
100%
67%
Commensal
MAEC
protein YdiA
yee
WV
Aec78
Sel1 domain protein
repeat-containing protein
Hcp-like protein
trnD2
ykfM
rnh
dna
yaf
yaf
imp
ets
Hcp-like protein
ets
rhs core gene
yafV
ync
yhh
ivy
AQ
T
U
A
I
A
X
V
U
M
N
H
R
C
D
B
S
G
I
I
rhsC
ECA-O157
1303
MPEC4839
D6-117.07
UVM2
O32:H37 P4
D6-117.29
ECC-Z
131/07
3234/A
ECA-727
RiKo 2308/09
RiKo 2340/09
RiKo 2351/09
RiKo 2305/09
RiKo 2299/09
MPEC4969
ECC-1470
O157:H43 T22
W26
RiKo 2331/09
P4-NR
2772a
D6-113.11
AA86
A
B1
E
B2

Figure S7. Gene organization of the Flag-2 and MAEC ECC-1470 subtype i1 T6SS/1 gene clusters. Comparisons were done with Easyfig utilizing BLASTN+. Homologous regions are connected via grey vertices and colored by nucleotide identity. The genomes are ordered according to the WGA core genome phylogeny, ECOR phylogroups are indicated correspondingly. MAEC strain names are colored in light red and commensal strains in green. Gene names are indicated above genomes encoding for these. The respective contigs of the draft genomes containing the gene cluster were concatenated (contig boundaries are indicated by red vertical lines) and CDS spanning contig borders reannotated if needed (indicated by asterisks). Backbone genes not belonging to the putative VFs are colored black. Genes within each respective region have different colors (see the legend) to be able to evaluate their presence. Pseudogenes have a lighter color fill. **(A)** Flag-2 gene cluster comparison. *E. coli* strain O157:H43 T22 is omitted from the diagram because of a large deletion including also the Flag-2 flanking backbone genes. **(B)** MAEC ECC-1470 T6SS/1 gene organization comparison. The T6SS gene region of draft MAEC D6-117.29 could probably not be fully manually assembled, because of its high fragmentation. The *E. coli* strain O157:H43 T22 deletion encompasses also the T6SS downstream housekeeping genes, which is indicated by dots in the figure. Strain ECA-727 lacks the *yafT* to *impA* genes, because of a putative phage insertion in this region. This phage is not included in the figure and the truncation also indicated by dots.



Supplementary information

No evidence for a bovine mastitis *Escherichia coli* pathotype

Andreas Leimbach[1,2,3], Anja Poehlein[2], John Vollmers[4], Dennis Görlich[5], Rolf Daniel[2], Ulrich Dobrindt[1,3]

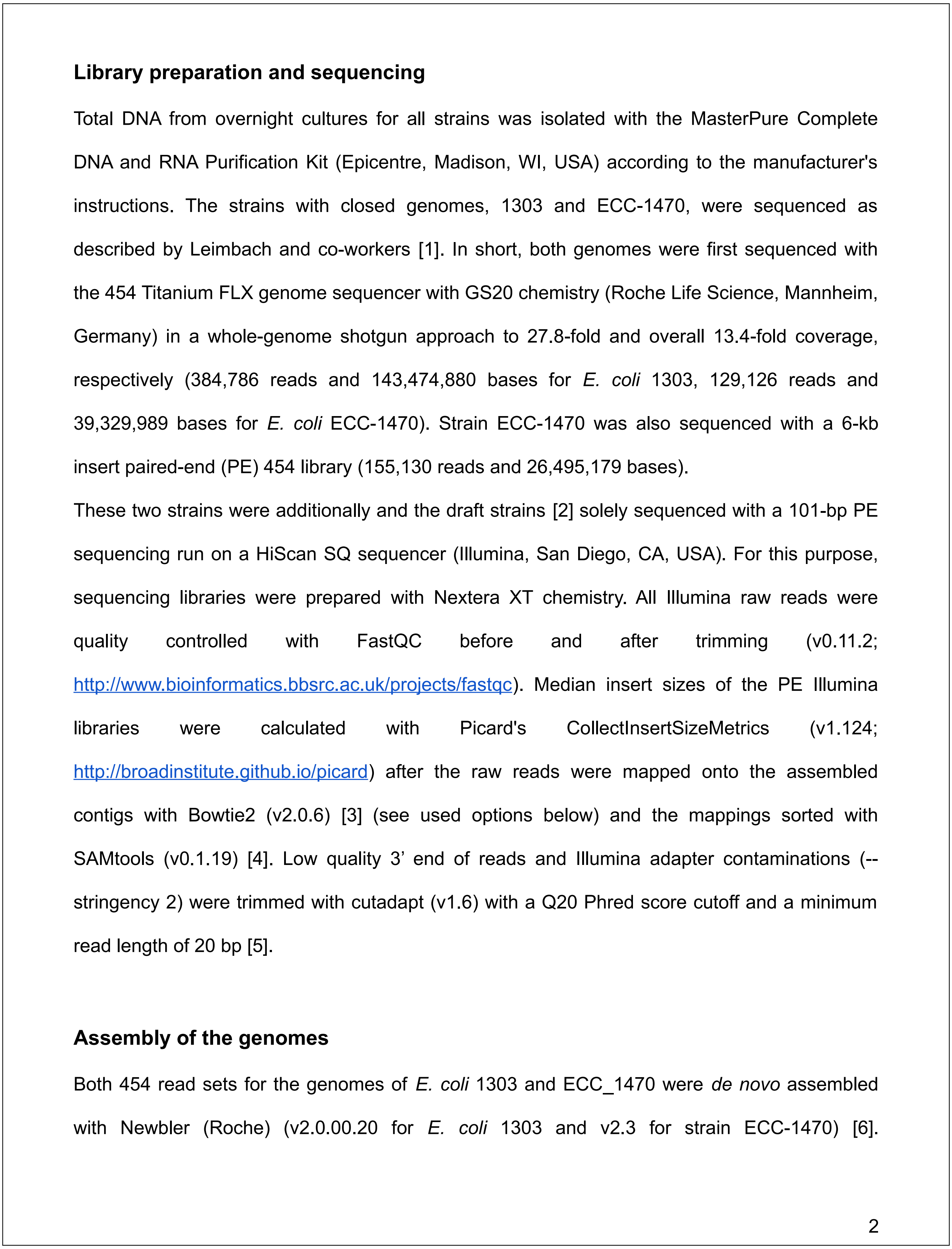
[1] Institute of Hygiene, University of Münster, Münster, Germany

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

[3] Institute for Molecular Infection Biology, Julius-Maximilians-University of Würzburg, Würzburg, Germany

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

[5] Institute of Biostatistics and Clinical Research, University of Münster, Münster, Germany



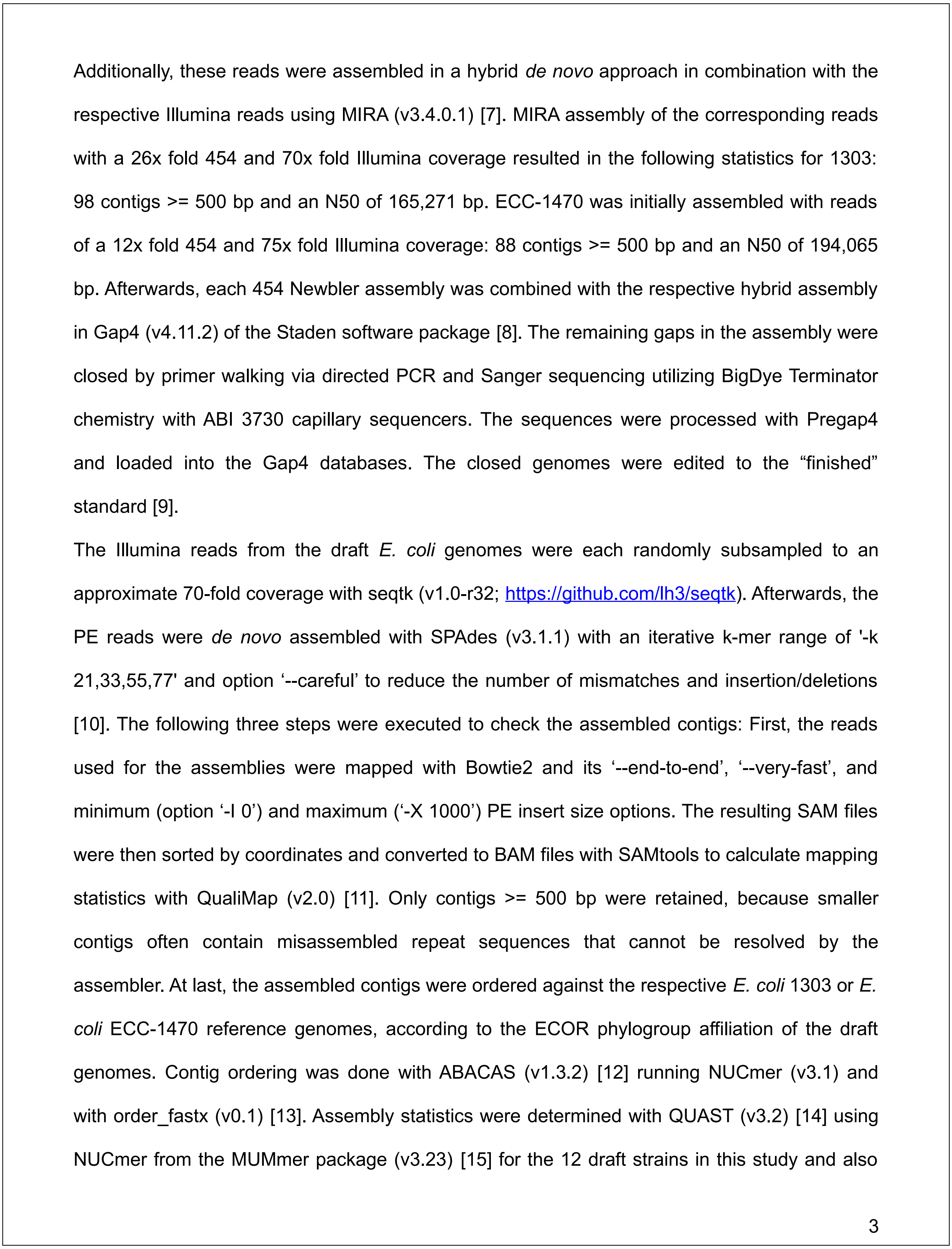
Library preparation and sequencing

Total DNA from overnight cultures for all strains was isolated with the MasterPure Complete DNA and RNA Purification Kit (Epicentre, Madison, WI, USA) according to the manufacturer's instructions. The strains with closed genomes, 1303 and ECC-1470, were sequenced as described by Leimbach and co-workers [1]. In short, both genomes were first sequenced with the 454 Titanium FLX genome sequencer with GS20 chemistry (Roche Life Science, Mannheim, Germany) in a whole-genome shotgun approach to 27.8-fold and overall 13.4-fold coverage, respectively (384,786 reads and 143,474,880 bases for *E. coli* 1303, 129,126 reads and 39,329,989 bases for *E. coli* ECC-1470). Strain ECC-1470 was also sequenced with a 6-kb insert paired-end (PE) 454 library (155,130 reads and 26,495,179 bases).

These two strains were additionally and the draft strains [2] solely sequenced with a 101-bp PE sequencing run on a HiScan SQ sequencer (Illumina, San Diego, CA, USA). For this purpose, sequencing libraries were prepared with Nextera XT chemistry. All Illumina raw reads were quality controlled with FastQC before and after trimming (v0.11.2; http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Median insert sizes of the PE Illumina libraries were calculated with Picard's CollectInsertSizeMetrics (v1.124; http://broadinstitute.github.io/picard) after the raw reads were mapped onto the assembled contigs with Bowtie2 (v2.0.6) [3] (see used options below) and the mappings sorted with SAMtools (v0.1.19) [4]. Low quality 3’ end of reads and Illumina adapter contaminations (--stringency 2) were trimmed with cutadapt (v1.6) with a Q20 Phred score cutoff and a minimum read length of 20 bp [5].

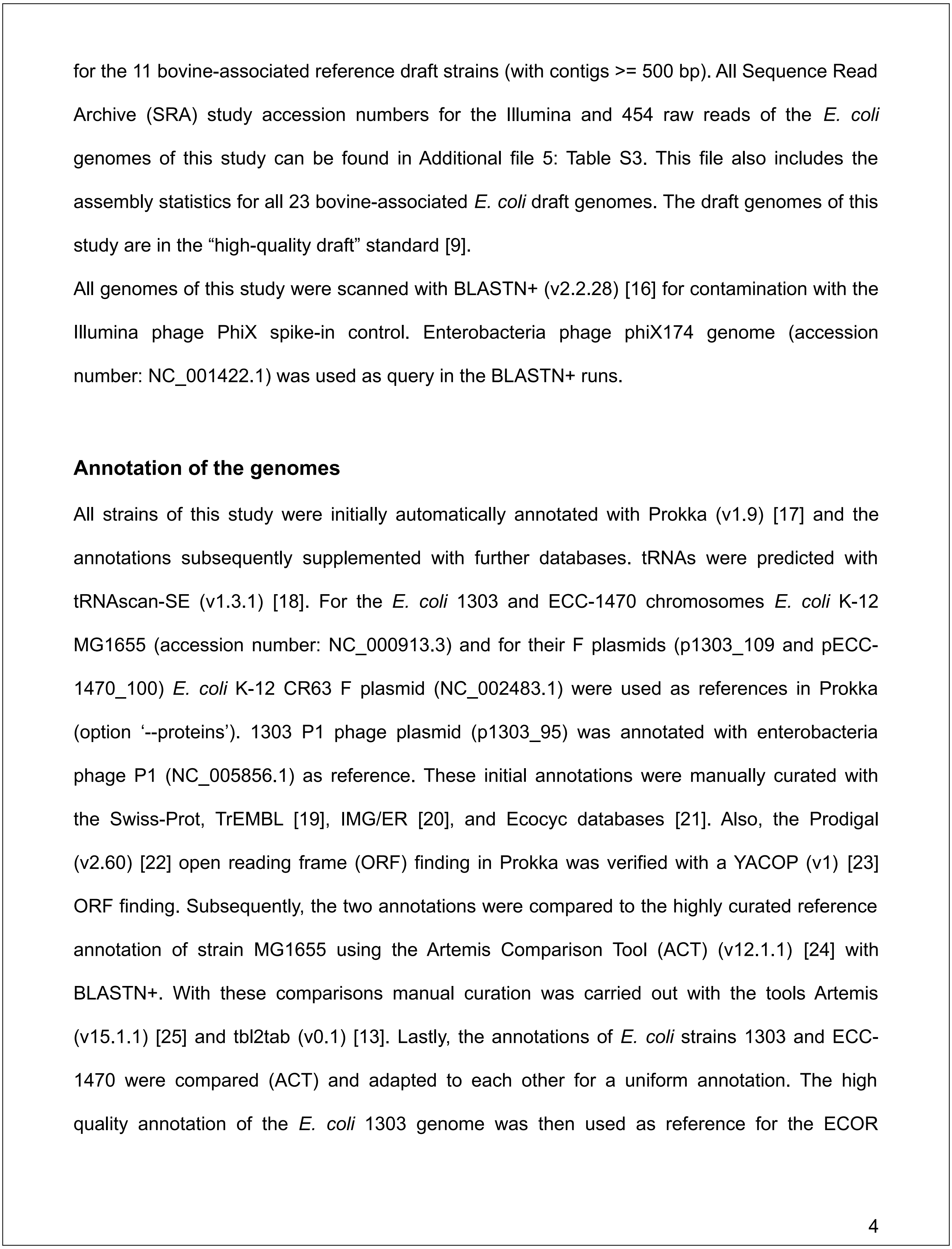
Assembly of the genomes

Both 454 read sets for the genomes of *E. coli* 1303 and ECC_1470 were *de novo* assembled with Newbler (Roche) (v2.0.00.20 for *E. coli* 1303 and v2.3 for strain ECC-1470) [6].



Additionally, these reads were assembled in a hybrid *de novo* approach in combination with the respective Illumina reads using MIRA (v3.4.0.1) [7]. MIRA assembly of the corresponding reads with a 26x fold 454 and 70x fold Illumina coverage resulted in the following statistics for 1303: 98 contigs >= 500 bp and an N50 of 165,271 bp. ECC-1470 was initially assembled with reads of a 12x fold 454 and 75x fold Illumina coverage: 88 contigs >= 500 bp and an N50 of 194,065 bp. Afterwards, each 454 Newbler assembly was combined with the respective hybrid assembly in Gap4 (v4.11.2) of the Staden software package [8]. The remaining gaps in the assembly were closed by primer walking via directed PCR and Sanger sequencing utilizing BigDye Terminator chemistry with ABI 3730 capillary sequencers. The sequences were processed with Pregap4 and loaded into the Gap4 databases. The closed genomes were edited to the “finished” standard [9].

The Illumina reads from the draft *E. coli* genomes were each randomly subsampled to an approximate 70-fold coverage with seqtk (v1.0-r32; https://github.com/lh3/seqtk). Afterwards, the PE reads were *de novo* assembled with SPAdes (v3.1.1) with an iterative k-mer range of '-k 21,33,55,77' and option ‘--careful’ to reduce the number of mismatches and insertion/deletions [10]. The following three steps were executed to check the assembled contigs: First, the reads used for the assemblies were mapped with Bowtie2 and its ‘--end-to-end’, ‘--very-fast’, and minimum (option ‘-I 0’) and maximum (‘-X 1000’) PE insert size options. The resulting SAM files were then sorted by coordinates and converted to BAM files with SAMtools to calculate mapping statistics with QualiMap (v2.0) [11]. Only contigs >= 500 bp were retained, because smaller contigs often contain misassembled repeat sequences that cannot be resolved by the assembler. At last, the assembled contigs were ordered against the respective *E. coli* 1303 or *E. coli* ECC-1470 reference genomes, according to the ECOR phylogroup affiliation of the draft genomes. Contig ordering was done with ABACAS (v1.3.2) [12] running NUCmer (v3.1) and with order_fastx (v0.1) [13]. Assembly statistics were determined with QUAST (v3.2) [14] using NUCmer from the MUMmer package (v3.23) [15] for the 12 draft strains in this study and also



for the 11 bovine-associated reference draft strains (with contigs >= 500 bp). All Sequence Read Archive (SRA) study accession numbers for the Illumina and 454 raw reads of the *E. coli* genomes of this study can be found in Additional file 5: Table S3. This file also includes the assembly statistics for all 23 bovine-associated *E. coli* draft genomes. The draft genomes of this study are in the “high-quality draft” standard [9].

All genomes of this study were scanned with BLASTN+ (v2.2.28) [16] for contamination with the Illumina phage PhiX spike-in control. Enterobacteria phage phiX174 genome (accession number: NC_001422.1) was used as query in the BLASTN+ runs.

Annotation of the genomes

All strains of this study were initially automatically annotated with Prokka (v1.9) [17] and the annotations subsequently supplemented with further databases. tRNAs were predicted with tRNAscan-SE (v1.3.1) [18]. For the *E. coli* 1303 and ECC-1470 chromosomes *E. coli* K-12 MG1655 (accession number: NC_000913.3) and for their F plasmids (p1303_109 and pECC-1470_100) *E. coli* K-12 CR63 F plasmid (NC_002483.1) were used as references in Prokka (option ‘--proteins’). 1303 P1 phage plasmid (p1303_95) was annotated with enterobacteria phage P1 (NC_005856.1) as reference. These initial annotations were manually curated with the Swiss-Prot, TrEMBL [19], IMG/ER [20], and Ecocyc databases [21]. Also, the Prodigal (v2.60) [22] open reading frame (ORF) finding in Prokka was verified with a YACOP (v1) [23] ORF finding. Subsequently, the two annotations were compared to the highly curated reference annotation of strain MG1655 using the Artemis Comparison Tool (ACT) (v12.1.1) [24] with BLASTN+. With these comparisons manual curation was carried out with the tools Artemis (v15.1.1) [25] and tbl2tab (v0.1) [13]. Lastly, the annotations of *E. coli* strains 1303 and ECC-1470 were compared (ACT) and adapted to each other for a uniform annotation. The high quality annotation of the *E. coli* 1303 genome was then used as reference for the ECOR phylogroup A strains and the ECC-1470 genome annotation for the ECOR B1 strains during the Prokka annotation of the 12 draft genomes of this study. These annotations were further manually curated via ortholog/genome synteny analyses with the respective replicons of *E. coli* strains 1303 and ECC-1470 as references with Proteinortho (v5.11) [26, 27] (see options below) and po2anno (v0.2) [13], ACT (v13.0.0) [24] with BLASTN+, and cat_seq (v0.1) [13]. At last, releases 1 (R1) and 2 (R2) of the Virulence Factors Database (VFDB) [28, 29], and the ResFinder (v2.1) [30], VirulenceFinder (v1.2) [31], and SerotypeFinder (v1.0) [32] databases were used to refine the annotations with Artemis (v16.0.0) and tbl2tab (v0.2).

All eleven reference strains were also automatically reannotated with Prokka to have a uniform ORF-finding with Prodigal and facilitate comparative genomics. The draft genomes of D6-113.11 and D6-117_07.11 contain one contig each smaller than 200 bp. These two contigs were skipped by Prokka with the used option '--compliant'. The annotations of the references were shortly manually curated in the three putative virulence regions ETT2, Flag-2, and strain ECC-1470's T6SS/1 by comparisons to the 1303 and ECC-1470 genomes as mentioned above. GENBANK files for these reannotations were created with NCBI's tbl2asn (v24.3; https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/) with option '-V b' and can be found in Additional file 22: Dataset S13 and Additional file 23: Dataset S14. For an overview of the annotations see the genome feature table created with genomes_feature_table (v0.5) [13] (Additional file 1: Table S1). This table also includes the reference *E. coli* genomes for the phylogenetic analysis (see below), however their annotation features are listed as downloaded from NCBI.

Phylogenetic analysis

For the phylogenetic analysis 39 additional reference *E. coli* strains (plus four *Shigella* spp. and one *Escherichia fergusonii* strain) were downloaded from NCBI with a wide variety of known pathotype and ECOR phylogroup affiliations. For the accession numbers see Additional file 1:

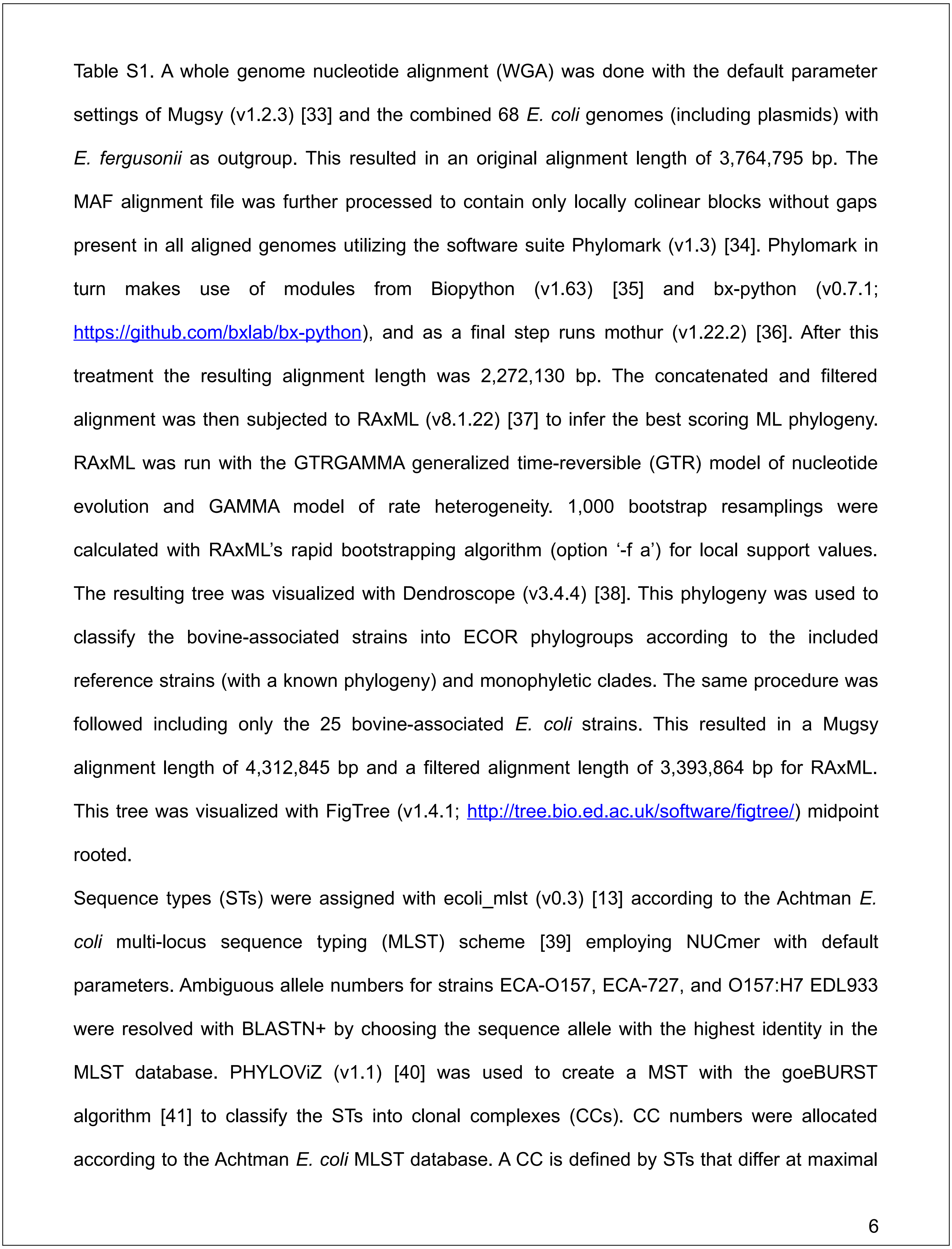
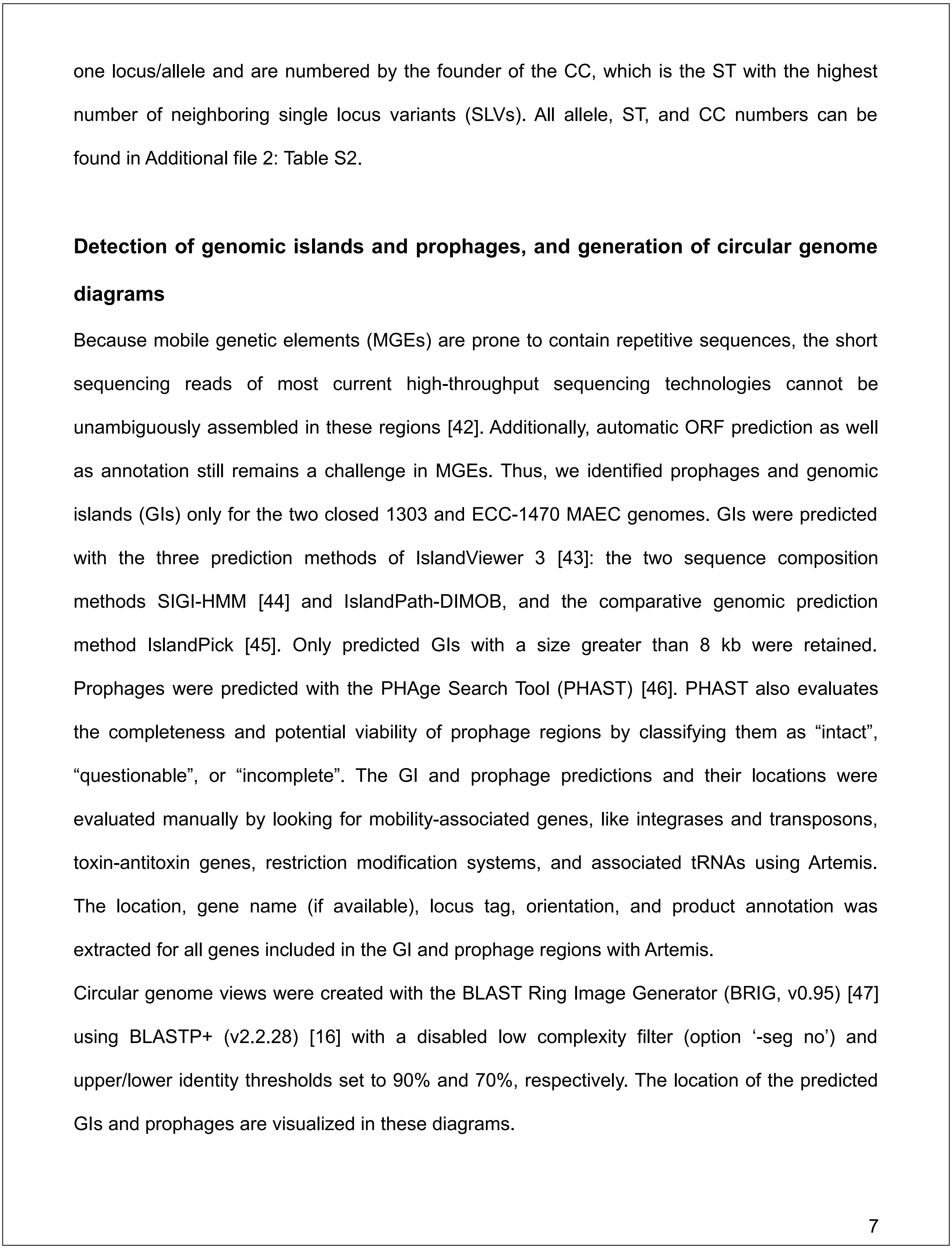


Table S1. A whole genome nucleotide alignment (WGA) was done with the default parameter settings of Mugsy (v1.2.3) [33] and the combined 68 *E. coli* genomes (including plasmids) with *E. fergusonii* as outgroup. This resulted in an original alignment length of 3,764,795 bp. The MAF alignment file was further processed to contain only locally colinear blocks without gaps present in all aligned genomes utilizing the software suite Phylomark (v1.3) [34]. Phylomark in turn makes use of modules from Biopython (v1.63) [35] and bx-python (v0.7.1; https://github.com/bxlab/bx-python), and as a final step runs mothur (v1.22.2) [36]. After this treatment the resulting alignment length was 2,272,130 bp. The concatenated and filtered alignment was then subjected to RAxML (v8.1.22) [37] to infer the best scoring ML phylogeny. RAxML was run with the GTRGAMMA generalized time-reversible (GTR) model of nucleotide evolution and GAMMA model of rate heterogeneity. 1,000 bootstrap resamplings were calculated with RAxML's rapid bootstrapping algorithm (option '-f a') for local support values. The resulting tree was visualized with Dendroscope (v3.4.4) [38]. This phylogeny was used to classify the bovine-associated strains into ECOR phylogroups according to the included reference strains (with a known phylogeny) and monophyletic clades. The same procedure was followed including only the 25 bovine-associated *E. coli* strains. This resulted in a Mugsy alignment length of 4,312,845 bp and a filtered alignment length of 3,393,864 bp for RAxML. This tree was visualized with FigTree (v1.4.1; http://tree.bio.ed.ac.uk/software/figtree/) midpoint rooted.

Sequence types (STs) were assigned with ecoli_mlst (v0.3) [13] according to the Achtman *E. coli* multi-locus sequence typing (MLST) scheme [39] employing NUCmer with default parameters. Ambiguous allele numbers for strains ECA-O157, ECA-727, and O157:H7 EDL933 were resolved with BLASTN+ by choosing the sequence allele with the highest identity in the MLST database. PHYLOViZ (v1.1) [40] was used to create a MST with the goeBURST algorithm [41] to classify the STs into clonal complexes (CCs). CC numbers were allocated according to the Achtman *E. coli* MLST database. A CC is defined by STs that differ at maximal

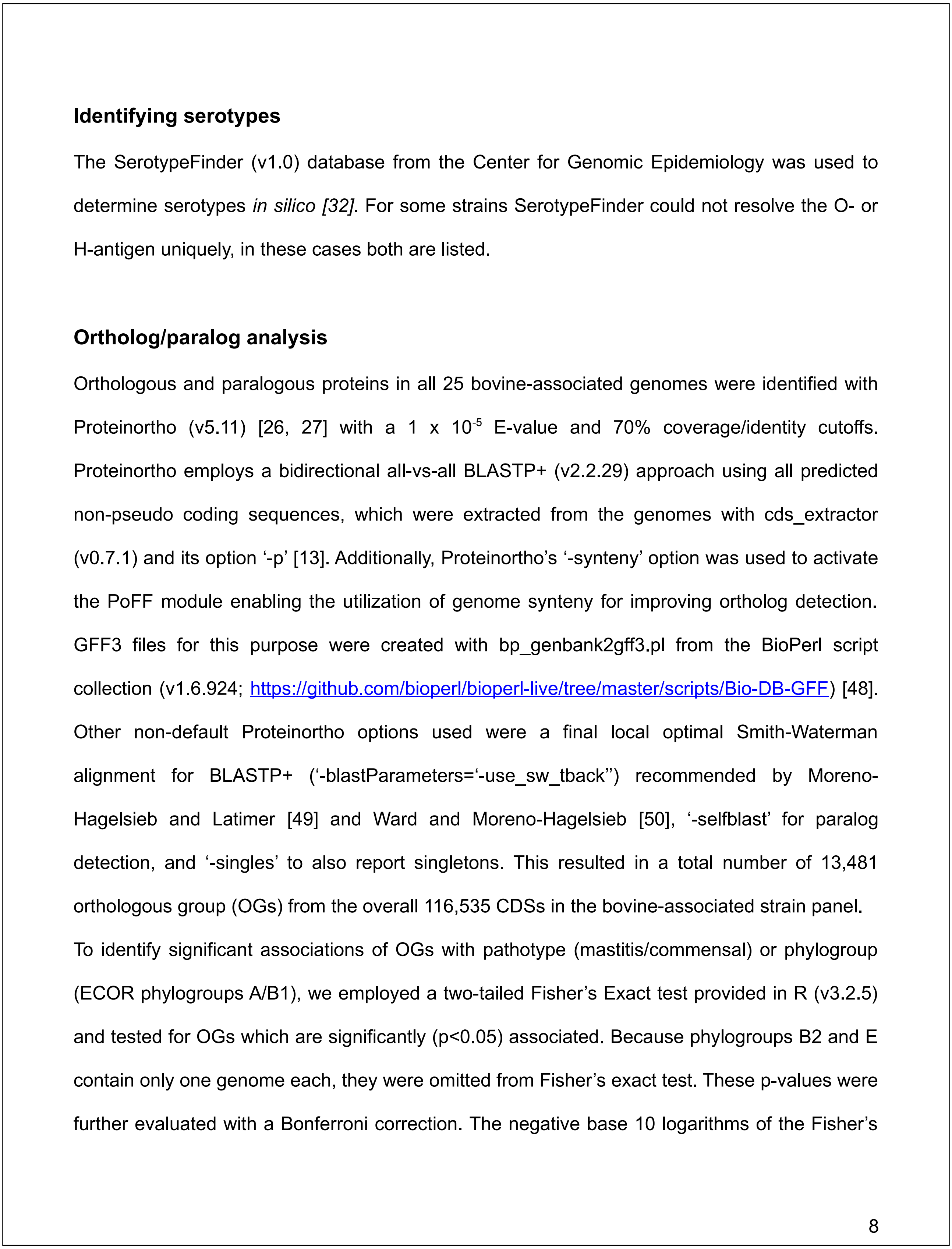


one locus/allele and are numbered by the founder of the CC, which is the ST with the highest number of neighboring single locus variants (SLVs). All allele, ST, and CC numbers can be found in Additional file 2: Table S2.

Detection of genomic islands and prophages, and generation of circular genome diagrams

Because mobile genetic elements (MGEs) are prone to contain repetitive sequences, the short sequencing reads of most current high-throughput sequencing technologies cannot be unambiguously assembled in these regions [42]. Additionally, automatic ORF prediction as well as annotation still remains a challenge in MGEs. Thus, we identified prophages and genomic islands (GIs) only for the two closed 1303 and ECC-1470 MAEC genomes. GIs were predicted with the three prediction methods of IslandViewer 3 [43]: the two sequence composition methods SIGI-HMM [44] and IslandPath-DIMOB, and the comparative genomic prediction method IslandPick [45]. Only predicted GIs with a size greater than 8 kb were retained. Prophages were predicted with the PHAge Search Tool (PHAST) [46]. PHAST also evaluates the completeness and potential viability of prophage regions by classifying them as “intact”, “questionable”, or “incomplete”. The GI and prophage predictions and their locations were evaluated manually by looking for mobility-associated genes, like integrases and transposons, toxin-antitoxin genes, restriction modification systems, and associated tRNAs using Artemis. The location, gene name (if available), locus tag, orientation, and product annotation was extracted for all genes included in the GI and prophage regions with Artemis.

Circular genome views were created with the BLAST Ring Image Generator (BRIG, v0.95) [47] using BLASTP+ (v2.2.28) [16] with a disabled low complexity filter (option ‘-seg no’) and upper/lower identity thresholds set to 90% and 70%, respectively. The location of the predicted GIs and prophages are visualized in these diagrams.



Identifying serotypes

The SerotypeFinder (v1.0) database from the Center for Genomic Epidemiology was used to determine serotypes *in silico [32]*. For some strains SerotypeFinder could not resolve the O- or H-antigen uniquely, in these cases both are listed.

Ortholog/paralog analysis

Orthologous and paralogous proteins in all 25 bovine-associated genomes were identified with Proteinortho (v5.11) [26, 27] with a 1 x 10^{-5} E-value and 70% coverage/identity cutoffs. Proteinortho employs a bidirectional all-vs-all BLASTP+ (v2.2.29) approach using all predicted non-pseudo coding sequences, which were extracted from the genomes with cds_extractor (v0.7.1) and its option '-p' [13]. Additionally, Proteinortho's '-synteny' option was used to activate the PoFF module enabling the utilization of genome synteny for improving ortholog detection. GFF3 files for this purpose were created with bp_genbank2gff3.pl from the BioPerl script collection (v1.6.924; https://github.com/bioperl/bioperl-live/tree/master/scripts/Bio-DB-GFF) [48]. Other non-default Proteinortho options used were a final local optimal Smith-Waterman alignment for BLASTP+ ('-blastParameters='-use_sw_tback'') recommended by Moreno-Hagelsieb and Latimer [49] and Ward and Moreno-Hagelsieb [50], '-selfblast' for paralog detection, and '-singles' to also report singletons. This resulted in a total number of 13,481 orthologous group (OGs) from the overall 116,535 CDSs in the bovine-associated strain panel.

To identify significant associations of OGs with pathotype (mastitis/commensal) or phylogroup (ECOR phylogroups A/B1), we employed a two-tailed Fisher's Exact test provided in R (v3.2.5) and tested for OGs which are significantly ($p<0.05$) associated. Because phylogroups B2 and E contain only one genome each, they were omitted from Fisher's exact test. These p-values were further evaluated with a Bonferroni correction. The negative base 10 logarithms of the Fisher's exact test p-values were visualized as Manhattan plots with R package ggplot2 (v2.2.0) [51]. The binary matrix for the OG presence/absence and the R script for the Fisher's exact test are included in Additional file 4: Dataset S1.

Additionally, we considered OGs as pathotype- or phylogroup-enriched if they are minimally present in 70% of the genomes of one genome group (inclusion cutoff) and in maximally 30% of the genomes of all other groups (the other pathotype or phylogroups; exclusion cutoff) using po2group_stats (v0.1.1) [13]. The 70%/30% inclusion/exclusion cutoffs amount to rounded 6/3 inclusion/exclusion genome cutoffs for the commensal isolates and 11/5 for the mastitis isolates. Similarly, the cutoffs in the phylogroups translate to rounded 9/4 genome inclusion/exclusion cutoffs for phylogroup A, 7/3 for B1, and 1/0 for the single genome groups B2 and E. According to these pathotype or phylogroup cutoffs, OGs are classified in "pathotype-/phylogroup-enriched", "-absent", "group soft core genome", "underrepresented", and "unspecific" (option '-u') categories. OGs that are present >= the inclusion cutoff in the genomes of all groups are categorized in the "group soft core genome" category. The "underrepresented" category includes OGs present in <= genomes than the exclusion cutoff in all groups. Finally, OGs that are present in more genomes than the exclusion, but less than the inclusion cutoff in any group are categorized as "unspecific". For each OG po2group_stats extracts the locus tag and annotation of one representative protein from one *E. coli* strain panel genome of the group (or in the case of paralogs several representative proteins).

The resulting Fisher's exact test significant and pathotype-/phylogroup-enriched OG numbers from po2group_stats were visualized in venn diagrams (po2group_stats option '-p') with the venn function of R package gplots (v3.0.1) [52]. Additionally, singletons (option '-s') were identified with po2group_stats.

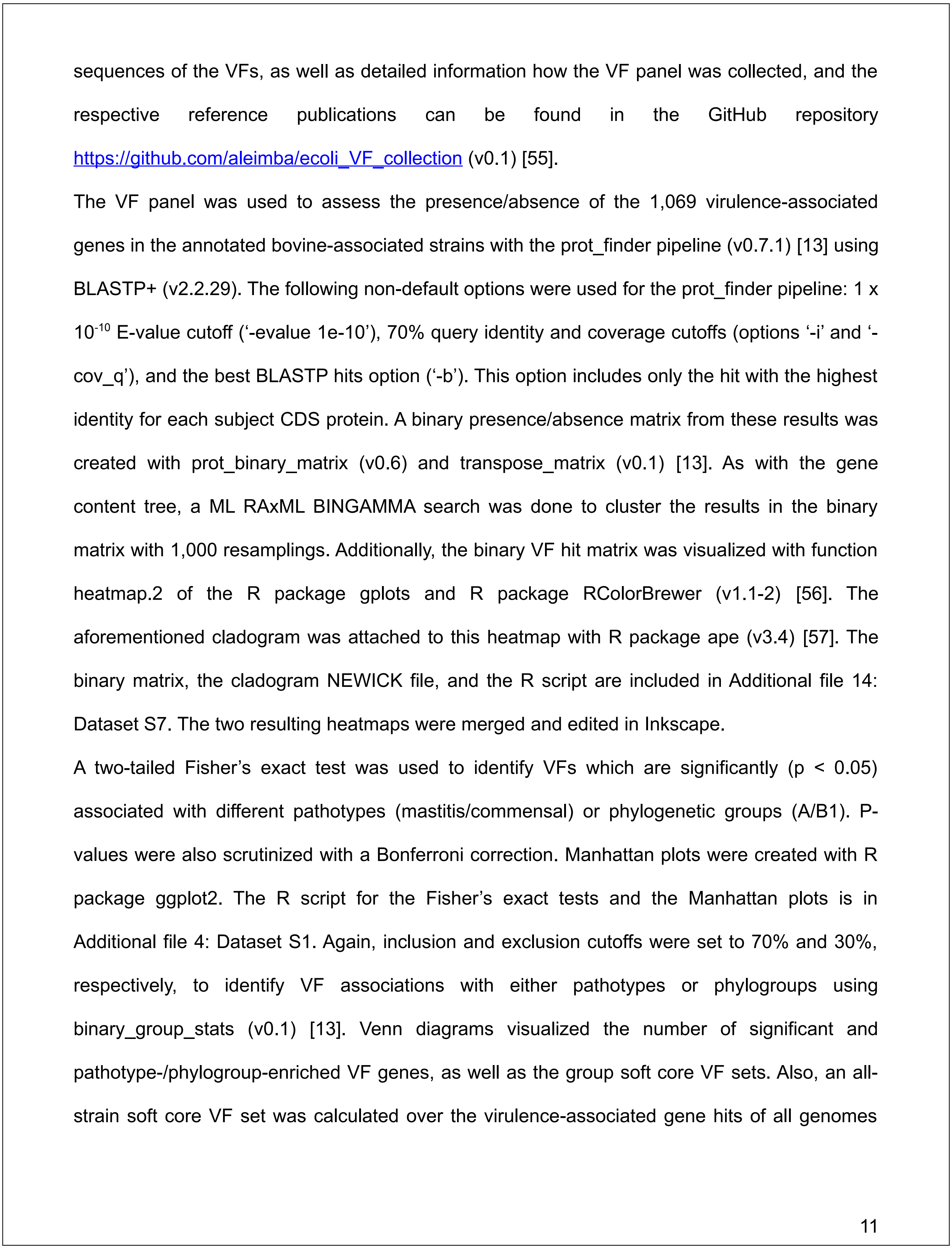
In addition to the pathotype and phylogroup group soft core genomes calculated by po2group_stats, an "all-strain soft core genome" including all genomes with the 70% inclusion cutoff (18 out of 25 genomes) was determined. The all-strain soft core genome always includes more OGs than the pathotype/phylogroup group soft cores, because of the different number of groups the 70% inclusion cutoff is applied to. The difference originates from the inclusion of all OGs which are present in at least 70% of all genomes of each group in comparison to 70% of all genomes.

The resulting pathotype-enriched OGs were further evaluated by comparing their representative proteins to the representative proteins in the phylogroup-enriched categories and the all-strain soft core. The representative protein sequences were extracted from the respective GENBANK files with the locus tags included in the po2group_stats result files using cds_extractor (options '-p' and '-l'). Subsequently, the prot_finder pipeline with BLASTP+ was used, as described in the virulence factore (VF) workflow, with the pathotype-enriched representative proteins as queries (option '-q') and the phylogroup-enriched or all-strain/phylogroup soft core proteins as subjects (option '-s').

Finally, a gene content tree was calculated with the Proteinortho presence/absence matrix of OGs (included in Additional file 4: Dataset S1). First, the matrix was converted to a binary matrix, transposed with transpose_matrix (v0.1) [13], and then converted to FASTA format. This file was used to cluster the results by searching for the best scoring ML tree with RAxML's (v8.0.26) BINGAMMA module (binary substitution model with GAMMA model of rate heterogeneity) and 1000 resamplings. The clustering tree was visualized midpoint rooted with Figtree.

Screening of the genomes for known virulence factors

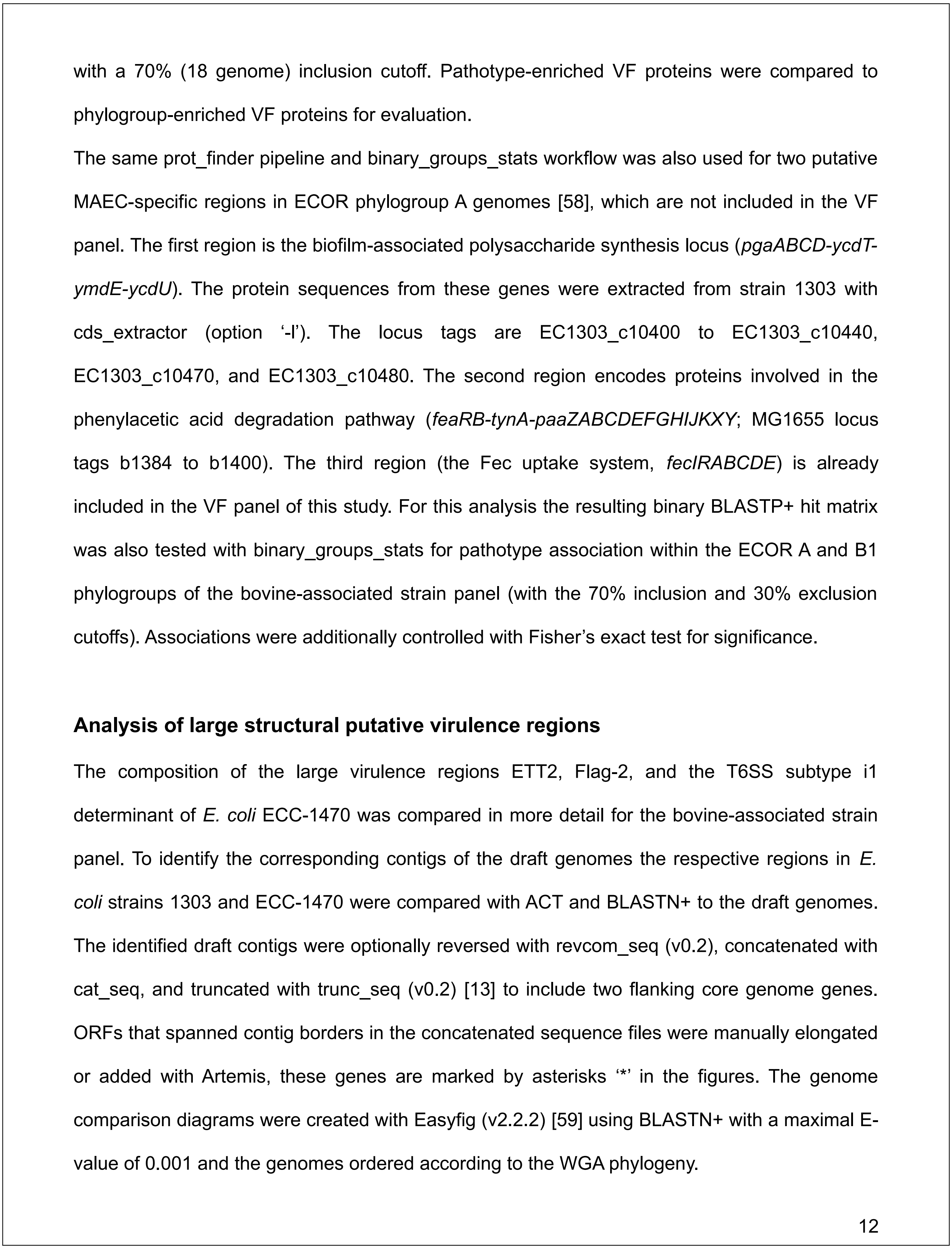
VF reference protein sequences were collected from the VFDB (R1 core dataset with experimentally validated VFs [53], R2 comparative genomics dataset with intra-genera comparisons [29], and R3 VF centric dataset with inter-genera comparisons [28]) and reviewing the primary literature. For an overview of the VF panel see Additional file 12: Table S5. A focus was put on putative ExPEC VFs, because MAEC are considered to be ExPEC [54]. The protein



sequences of the VFs, as well as detailed information how the VF panel was collected, and the respective reference publications can be found in the GitHub repository https://github.com/aleimba/ecoli_VF_collection (v0.1) [55].

The VF panel was used to assess the presence/absence of the 1,069 virulence-associated genes in the annotated bovine-associated strains with the prot_finder pipeline (v0.7.1) [13] using BLASTP+ (v2.2.29). The following non-default options were used for the prot_finder pipeline: 1 x 10^{-10} E-value cutoff ('-evalue 1e-10'), 70% query identity and coverage cutoffs (options '-i' and '-cov_q'), and the best BLASTP hits option ('-b'). This option includes only the hit with the highest identity for each subject CDS protein. A binary presence/absence matrix from these results was created with prot_binary_matrix (v0.6) and transpose_matrix (v0.1) [13]. As with the gene content tree, a ML RAxML BINGAMMA search was done to cluster the results in the binary matrix with 1,000 resamplings. Additionally, the binary VF hit matrix was visualized with function heatmap.2 of the R package gplots and R package RColorBrewer (v1.1-2) [56]. The aforementioned cladogram was attached to this heatmap with R package ape (v3.4) [57]. The binary matrix, the cladogram NEWICK file, and the R script are included in Additional file 14: Dataset S7. The two resulting heatmaps were merged and edited in Inkscape.

A two-tailed Fisher's exact test was used to identify VFs which are significantly ($p < 0.05$) associated with different pathotypes (mastitis/commensal) or phylogenetic groups (A/B1). P-values were also scrutinized with a Bonferroni correction. Manhattan plots were created with R package ggplot2. The R script for the Fisher's exact tests and the Manhattan plots is in Additional file 4: Dataset S1. Again, inclusion and exclusion cutoffs were set to 70% and 30%, respectively, to identify VF associations with either pathotypes or phylogroups using binary_group_stats (v0.1) [13]. Venn diagrams visualized the number of significant and pathotype-/phylogroup-enriched VF genes, as well as the group soft core VF sets. Also, an all-strain soft core VF set was calculated over the virulence-associated gene hits of all genomes

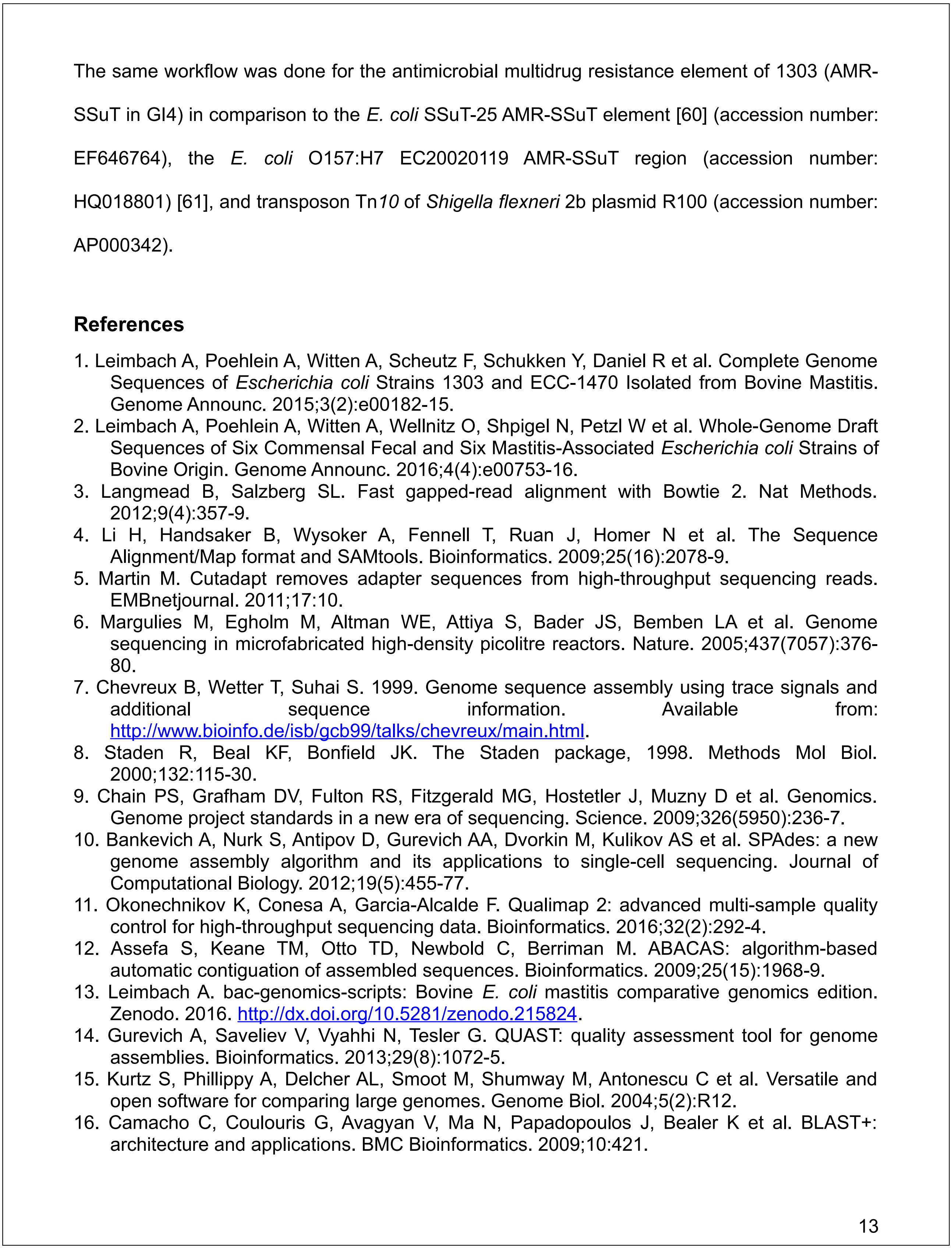


with a 70% (18 genome) inclusion cutoff. Pathotype-enriched VF proteins were compared to phylogroup-enriched VF proteins for evaluation.

The same prot_finder pipeline and binary_groups_stats workflow was also used for two putative MAEC-specific regions in ECOR phylogroup A genomes [58], which are not included in the VF panel. The first region is the biofilm-associated polysaccharide synthesis locus (*pgaABCD-ycdT-ymdE-ycdU*). The protein sequences from these genes were extracted from strain 1303 with cds_extractor (option '-l'). The locus tags are EC1303_c10400 to EC1303_c10440, EC1303_c10470, and EC1303_c10480. The second region encodes proteins involved in the phenylacetic acid degradation pathway (*feaRB-tynA-paaZABCDEFGHIJKXY*; MG1655 locus tags b1384 to b1400). The third region (the Fec uptake system, *fecIRABCDE*) is already included in the VF panel of this study. For this analysis the resulting binary BLASTP+ hit matrix was also tested with binary_groups_stats for pathotype association within the ECOR A and B1 phylogroups of the bovine-associated strain panel (with the 70% inclusion and 30% exclusion cutoffs). Associations were additionally controlled with Fisher's exact test for significance.

Analysis of large structural putative virulence regions

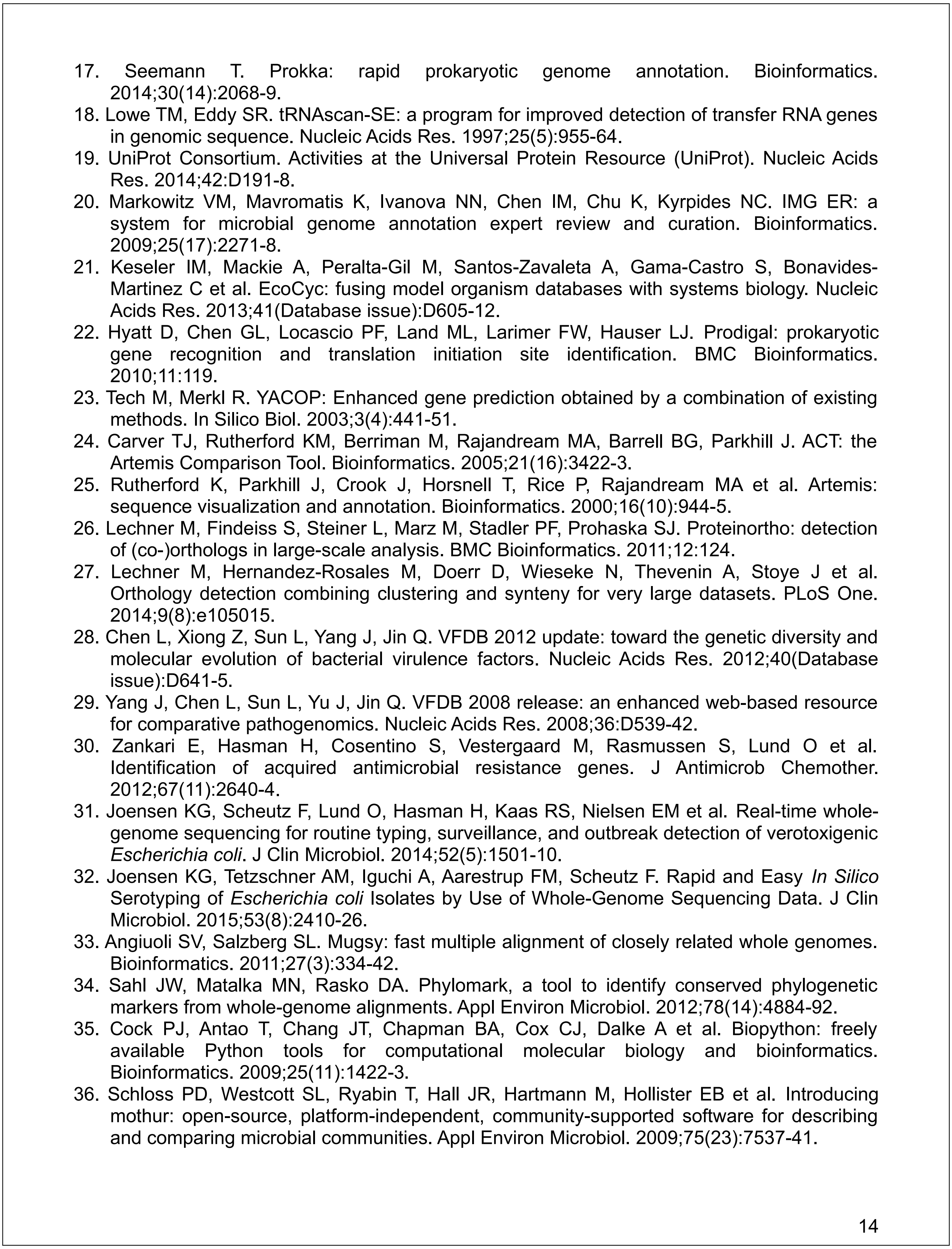
The composition of the large virulence regions ETT2, Flag-2, and the T6SS subtype i1 determinant of *E. coli* ECC-1470 was compared in more detail for the bovine-associated strain panel. To identify the corresponding contigs of the draft genomes the respective regions in *E. coli* strains 1303 and ECC-1470 were compared with ACT and BLASTN+ to the draft genomes. The identified draft contigs were optionally reversed with revcom_seq (v0.2), concatenated with cat_seq, and truncated with trunc_seq (v0.2) [13] to include two flanking core genome genes. ORFs that spanned contig borders in the concatenated sequence files were manually elongated or added with Artemis, these genes are marked by asterisks '*' in the figures. The genome comparison diagrams were created with Easyfig (v2.2.2) [59] using BLASTN+ with a maximal E-value of 0.001 and the genomes ordered according to the WGA phylogeny.



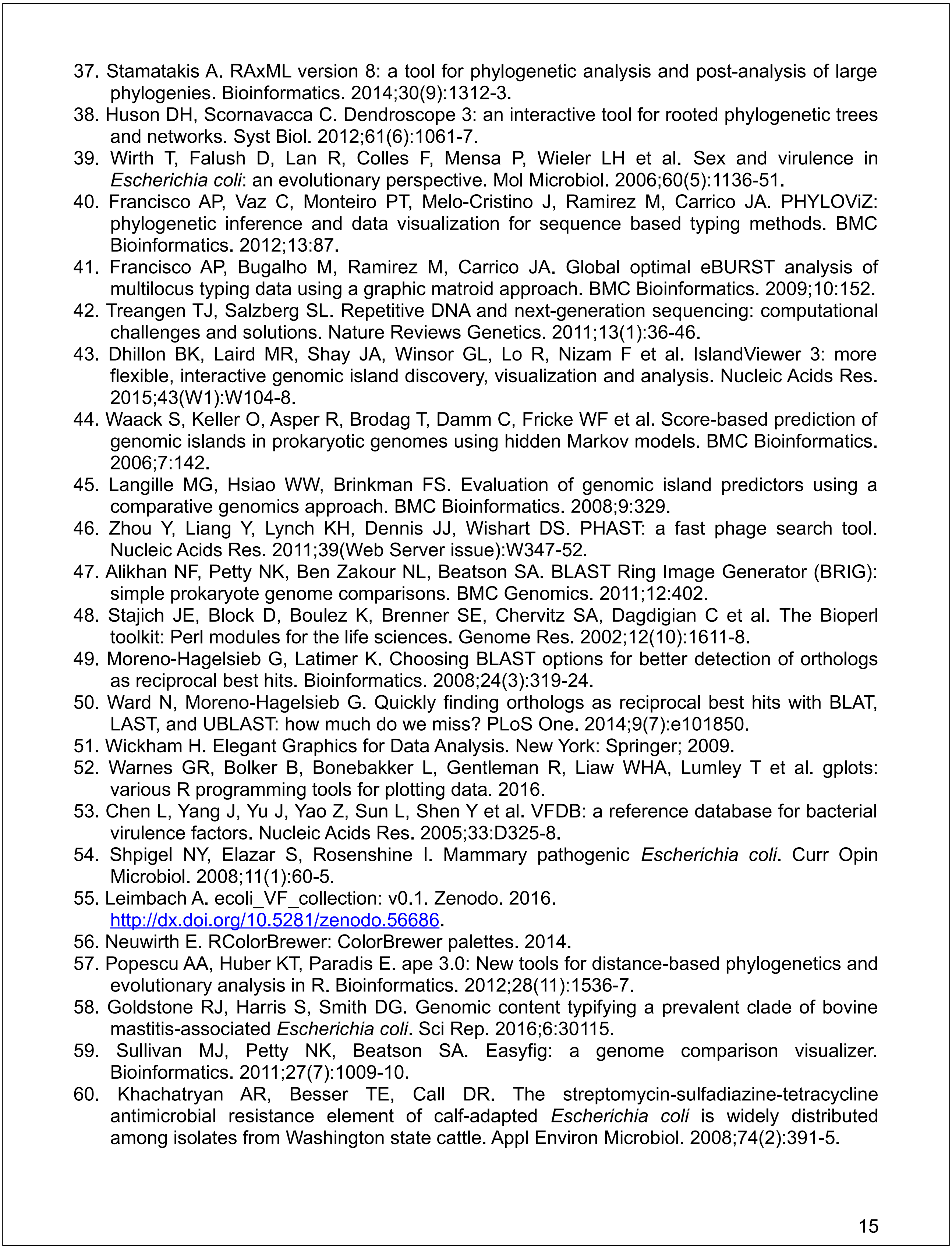
The same workflow was done for the antimicrobial multidrug resistance element of 1303 (AMR-SSuT in GI4) in comparison to the *E. coli* SSuT-25 AMR-SSuT element [60] (accession number: EF646764), the *E. coli* O157:H7 EC20020119 AMR-SSuT region (accession number: HQ018801) [61], and transposon Tn*10* of *Shigella flexneri* 2b plasmid R100 (accession number: AP000342).

References

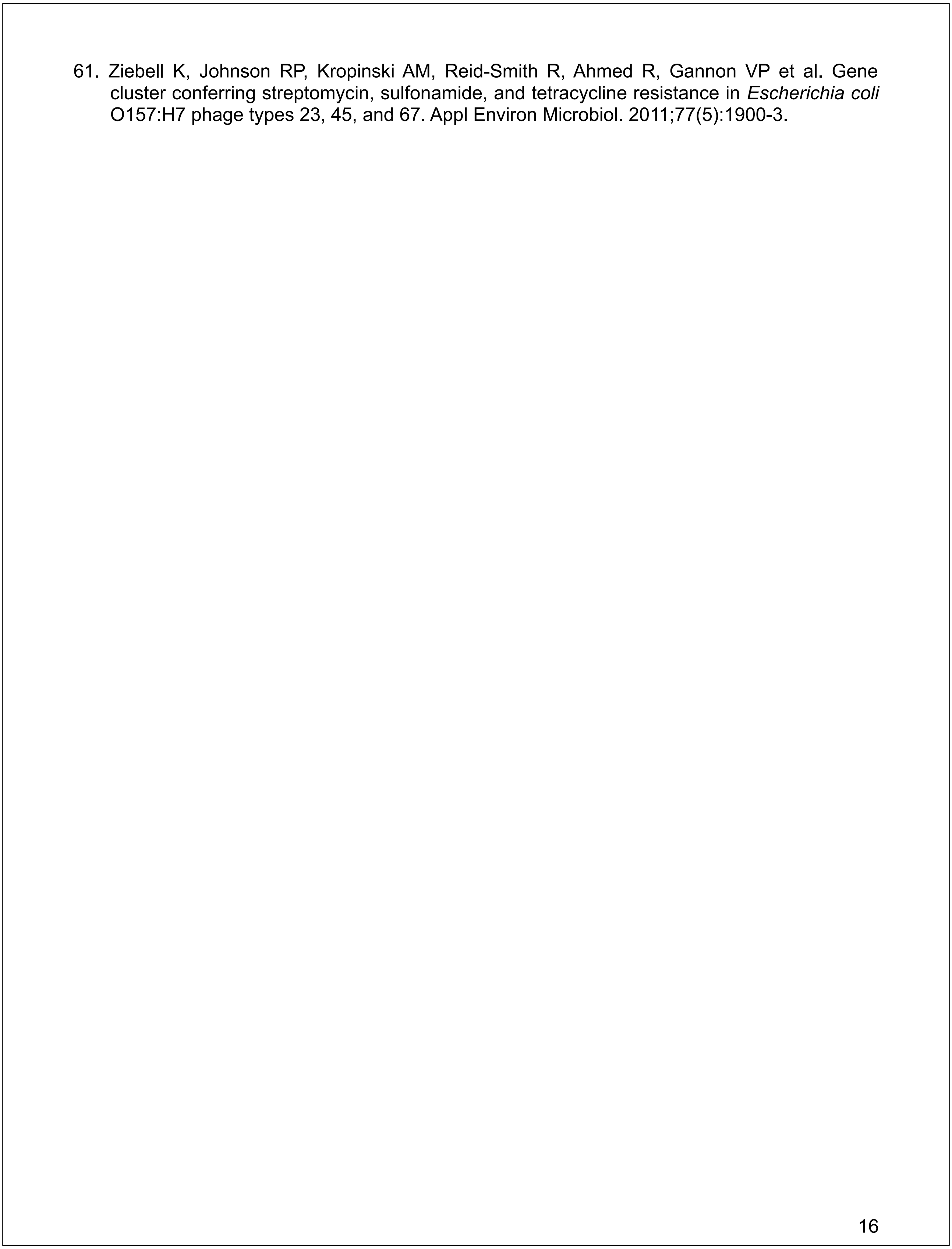
1. Leimbach A, Poehlein A, Witten A, Scheutz F, Schukken Y, Daniel R et al. Complete Genome Sequences of *Escherichia coli* Strains 1303 and ECC-1470 Isolated from Bovine Mastitis. Genome Announc. 2015;3(2):e00182-15.
2. Leimbach A, Poehlein A, Witten A, Wellnitz O, Shpigel N, Petzl W et al. Whole-Genome Draft Sequences of Six Commensal Fecal and Six Mastitis-Associated *Escherichia coli* Strains of Bovine Origin. Genome Announc. 2016;4(4):e00753-16.
3. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357-9.
4. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
5. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal. 2011;17:10.
6. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376-80.
7. Chevreux B, Wetter T, Suhai S. 1999. Genome sequence assembly using trace signals and additional sequence information. Available from: http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html.
8. Staden R, Beal KF, Bonfield JK. The Staden package, 1998. Methods Mol Biol. 2000;132:115-30.
9. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Muzny D et al. Genomics. Genome project standards in a new era of sequencing. Science. 2009;326(5950):236-7.
10. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012;19(5):455-77.
11. Okonechnikov K, Conesa A, Garcia-Alcalde F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 2016;32(2):292-4.
12. Assefa S, Keane TM, Otto TD, Newbold C, Berriman M. ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics. 2009;25(15):1968-9.
13. Leimbach A. bac-genomics-scripts: Bovine *E. coli* mastitis comparative genomics edition. Zenodo. 2016. http://dx.doi.org/10.5281/zenodo.215824.
14. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5.
15. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
16. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.



17. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068-9.
18. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):955-64.
19. UniProt Consortium. Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res. 2014;42:D191-8.
20. Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics. 2009;25(17):2271-8.
21. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C et al. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 2013;41(Database issue):D605-12.
22. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
23. Tech M, Merkl R. YACOP: Enhanced gene prediction obtained by a combination of existing methods. In Silico Biol. 2003;3(4):441-51.
24. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. Bioinformatics. 2005;21(16):3422-3.
25. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000;16(10):944-5.
26. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. Proteinortho: detection of (co-)orthologs in large-scale analysis. BMC Bioinformatics. 2011;12:124.
27. Lechner M, Hernandez-Rosales M, Doerr D, Wieseke N, Thevenin A, Stoye J et al. Orthology detection combining clustering and synteny for very large datasets. PLoS One. 2014;9(8):e105015.
28. Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Res. 2012;40(Database issue):D641-5.
29. Yang J, Chen L, Sun L, Yu J, Jin Q. VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics. Nucleic Acids Res. 2008;36:D539-42.
30. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother. 2012;67(11):2640-4.
31. Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol. 2014;52(5):1501-10.
32. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data. J Clin Microbiol. 2015;53(8):2410-26.
33. Angiuoli SV, Salzberg SL. Mugsy: fast multiple alignment of closely related whole genomes. Bioinformatics. 2011;27(3):334-42.
34. Sahl JW, Matalka MN, Rasko DA. Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments. Appl Environ Microbiol. 2012;78(14):4884-92.
35. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422-3.
36. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009;75(23):7537-41.



37. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
38. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012;61(6):1061-7.
39. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol. 2006;60(5):1136-51.
40. Francisco AP, Vaz C, Monteiro PT, Melo-Cristino J, Ramirez M, Carrico JA. PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods. BMC Bioinformatics. 2012;13:87.
41. Francisco AP, Bugalho M, Ramirez M, Carrico JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009;10:152.
42. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics. 2011;13(1):36-46.
43. Dhillon BK, Laird MR, Shay JA, Winsor GL, Lo R, Nizam F et al. IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis. Nucleic Acids Res. 2015;43(W1):W104-8.
44. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics. 2006;7:142.
45. Langille MG, Hsiao WW, Brinkman FS. Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics. 2008;9:329.
46. Zhou Y, Liang Y, Lynch KH, Dennis JJ, Wishart DS. PHAST: a fast phage search tool. Nucleic Acids Res. 2011;39(Web Server issue):W347-52.
47. Alikhan NF, Petty NK, Ben Zakour NL, Beatson SA. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics. 2011;12:402.
48. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C et al. The Bioperl toolkit: Perl modules for the life sciences. Genome Res. 2002;12(10):1611-8.
49. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008;24(3):319-24.
50. Ward N, Moreno-Hagelsieb G. Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PLoS One. 2014;9(7):e101850.
51. Wickham H. Elegant Graphics for Data Analysis. New York: Springer; 2009.
52. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T et al. gplots: various R programming tools for plotting data. 2016.
53. Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y et al. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res. 2005;33:D325-8.
54. Shpigel NY, Elazar S, Rosenshine I. Mammary pathogenic *Escherichia coli*. Curr Opin Microbiol. 2008;11(1):60-5.
55. Leimbach A. ecoli_VF_collection: v0.1. Zenodo. 2016. http://dx.doi.org/10.5281/zenodo.56686.
56. Neuwirth E. RColorBrewer: ColorBrewer palettes. 2014.
57. Popescu AA, Huber KT, Paradis E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. Bioinformatics. 2012;28(11):1536-7.
58. Goldstone RJ, Harris S, Smith DG. Genomic content typifying a prevalent clade of bovine mastitis-associated *Escherichia coli*. Sci Rep. 2016;6:30115.
59. Sullivan MJ, Petty NK, Beatson SA. Easyfig: a genome comparison visualizer. Bioinformatics. 2011;27(7):1009-10.
60. Khachatryan AR, Besser TE, Call DR. The streptomycin-sulfadiazine-tetracycline antimicrobial resistance element of calf-adapted *Escherichia coli* is widely distributed among isolates from Washington state cattle. Appl Environ Microbiol. 2008;74(2):391-5.



61. Ziebell K, Johnson RP, Kropinski AM, Reid-Smith R, Ahmed R, Gannon VP et al. Gene cluster conferring streptomycin, sulfonamide, and tetracycline resistance in *Escherichia coli* O157:H7 phage types 23, 45, and 67. Appl Environ Microbiol. 2011;77(5):1900-3.

Part IV

GENERAL DISCUSSION

6 DISCUSSION

Two main themes are dominating this thesis: The unexpected high genomic plasticity of the species *E. coli* and the impact of phylogenetic background on the gene content of individual strains. Both will be discussed on several occasions in this chapter, especially their implications on mastitis prevention and treatment options, and current diagnostic and public health microbiology standards.

6.1 THE ELUSIVE *E. COLI* 1303 O70 SEROTYPE

Initially, mastitis-associated *E. coli* (MAEC) 1303 was considered to have a serogroup O5 and the detailed structural O-antigen analysis was published as a new subtype of this serogroup accordingly (Section 5.3.1 on page 119) (Duda et al., 2011). However, a two-way cross-reactivity between the antibodies of O5 and O70 in the agglutination test lead to this false result. A re-serotyping by the World Health Organization (WHO) reference laboratory on *E. coli* typing, the Statens Serum Institute in Copenhagen, revealed conclusively that *E. coli* 1303 actually possesses an O70 LPS O-antigen and a full serotype of O70:K-:H32. Still, Duda et al. (2011) was the first description of the *E. coli* O70 serogroup structure and corresponding gene cluster.

We corrected the mistake in both successive publications Leimbach et al. (2015) (Section 5.3.2 on page 134) and Leimbach et al. (2017) (Section 5.3.4 on page 140). Also, the respective sequence records in the nucleotide databases for the O-antigen gene cluster (NCBI Genbank accession number FN995094) and the MAEC 1303 whole genome sequence entry (CP009166) contain the correct O70 serogroup. These entries supplement the over 196 *E. coli* O-groups and H-types recognized by traditional serotyping (DebRoy et al., 2016; Ingle et al., 2016a).

Nowadays, *E. coli* serotypes can be reliably predicted from WGS data as long as suitable references are available. Our characterized O-antigen gene cluster and finished genome of MAEC O70:H32 1303 (including *fliC*) can now serve as reference for tools like `SerotypeFinder` (Joensen et al., 2015). *In silico* methods have several advantages over traditional serological phenotyping, as they do not rely on typing sera with varying quality and their ability to type strains that autoagglutinate or do not express the respective O-, K-, or H-antigens *in vitro* (Ingle et al., 2016a; Robins-Browne et al., 2016). As in other areas of diagnostics, traditional wet-lab serotyping will most likely be replaced in the near future by *in silico* inferences in public health laboratories (Chapter 7 on page 219). The transfer of O- and H-antigen loci between different *E. coli* chromosomal backbones is common, thus limiting the use of serotyping to identify pathogenic *E. coli* lineages as e. g. in the German O104:H4 STEC epidemic (Ingle et al., 2016a). Universally applying WGS with bioinformatical analyses and high-quality reference databases in clinical microbiology would remedy many of the current drawbacks, e. g. problems in detecting emerging hybrid *E. coli* pathotypes.

6.2 FURTHER CONSEQUENCES OF THE GERMAN 2011 STEC EPIDEMIC

The detection of the hybrid STEC/EAEC pathotype was a milestone in microbial genomics and diagnostics. There are several scientific and clinical consequences during and after the outbreak not touched upon in our genomic analysis (Section 5.2 on page 104) that will be discussed here.

6.2.1 *Clinical effect of the hybrid STEC/EAEC pathotype*

proposed to be designated as EAHEC (Brzuszkiewicz et al., 2011) or STEAEC (Qin et al., 2011)

The hybrid STEC and EAEC pathotype was the cause for a very low infectious dose, the unusual high number of cases, and the high incidence of post-enteritis HUS[1] (about 25% of the cases and more than 900 patients) (Croxen et al., 2013; Frank et al., 2011). Such a high diarrhea-associated HUS rate is untypical for other STEC outbreaks, normally around 1–15% and in over 90% of cases in children (Buchholz et al., 2011; Frank et al., 2011). The augmented adhesion of the outbreak *E. coli* to enterocytes with the aggregative adherence fimbriae (AAF) (instead of the typical LEE) might have lead to higher colonization, higher Stx2 blood absorption, and ultimately the high incidence of HUS (Bielaszewska et al., 2011a; Rasko et al., 2011). In addition, it was shown that the outbreak strain produces a high amount of proinflammatory curli fibre adhesins that may contribute to enhanced Stx absorption into the bloodstream (Richter et al., 2014).

Usual HUS treatments like plasma exchange through dialysis and complement-blocking antibody (© eculizumab), proved to be inconclusive (Croxen et al., 2013; Greinacher et al., 2011). The use of antibiotics in HUS is disadvised, as they can activate the Stx phage to its lytic cycle, and thus promote HUS by increased Stx production (Croxen et al., 2013; Karch et al., 2012). Nevertheless, an additional challenge of the O104:H4 STEC outbreak was the high antibiotic resistance of the strain due to the carriage of plasmid-encoded extended-spectrum β-lactamases (ESBLs) type CTX-M (*bla*$_{\text{CTX-M-15}}$) and TEM (*bla*$_{\text{TEM-1}}$) (Figure 10 on page 201), as well as other drug resistance genes (Section 5.2

STEC O104:H4 multiresistant

[1] HUS is a serious disease characterized by the triad of acute renal failure, hemolytic anemia, and thrombocytopenia.

on page 104) (Bielaszewska et al., 2011a; Brzuszkiewicz et al., 2011; Monecke et al., 2011; Rohde et al., 2011).

6.2.2 *Real-time analysis of the outbreak* E. coli *genome*

The time frame of the epidemic was shortly after the release of several benchtop second-generation HTS machines with reduced run times (454 GS Junior in 2010, and in 2011 Illumina MiSeq and Ion Torrent PGM; Section 1.1.1.1 on page 6) (Loman et al., 2012b). Since the respective vendors were propagating the *democratization* of sequencing with these benchtop machines (i. e. the movement of HTS away from large sequencing centers to individual labs), it was a perfect opportunity to showcase the capabilities of these sequencers. Several labs and collaborations (mostly UKM in cooperation with Ion Torrent Life Tech, University Medical Center Hamburg-Eppendorf (UKE) with the Beijing Genomics Institute (BGI), the Göttingen Genomics Laboratory (G2L), and the Health Protection Agency (HPA)) quickly acquired bacterial *E. coli* outbreak isolates and sequenced the genomes while the outbreak was still ongoing (Brzuszkiewicz et al., 2011; Mellmann et al., 2011; Rohde et al., 2011). Nevertheless, it was a large genome center, the BGI in cooperation with the UKE, that released the first genomic sequence reads to the public with a Creative Commons Public Domain Dedication (CC0 1.0 Universal). This release in the public domain was the catalyst that started a frenzy of enthusiastic microbial genomicists and bioinformaticians across the globe to analyze the bacterial causative agent, with daily analysis updates on blogs, a GitHub repository[2] (see Section 3.1 on page 65 for the purpose of GitHub), and preprints. Major contributors were Nick Loman, Mark Pallen (both at that time University of Birmingham, UK), Kathryn Holt (University of Melbourne, Australia), David Studholme, Konrad Paszkiewicz (both University of Exeter, UK), Marina Manrique, Raquel Tobes, Eduardo Pareja-Tobes (all three Era7 Information Technologies, Spain), Lisa Crossman (University of East Anglia), and many others (Rohde et al., 2011). These *crowdsourced* analyses, as well as the analyses of the teams mentioned above, was the first time a genome of a bacterial pathogen was analyzed during an epidemic (in near real-time), made possible through the advancements in HTS and fast dissemination of information over the internet[3]. This epidemic and the corresponding genomic analyses showed a glimpse of the future of microbial epidemiology and diagnostics (Chapter 7 on page 219). Nevertheless, nearly all of the involved

open data release

[2] The GitHub repository with dates and links to the analyses (unfortunately many hyperlinks are out-dated now) can be found at https://github.com/ehec-outbreak-crowdsourced/BGI-data-analysis/wiki.

[3] Although many journals rushed publications concerning the epidemic through peer-review, instantaneous updates of blogs and websites exemplified how scientists can work collaboratively and how fast scientific information can be disseminated nowadays (Table 6).

scientists and groups published afterwards in traditional journals and satisfied the current scientific reward system (Table 6).

Table 6: Timeline of major genomic and epidemiological events during the 2011 German O104:H4 STEC epidemic.

DATE	EVENT
Begin of May	Outbreak started
20th of May	RKI notes increase in EHEC infections and HUS incidence
21st–22nd of May	Peak of outbreak cases
23rd of May	UKM receives first stool samples
25th of May	UKM typed the strain as O104:H4 based on *gnd* and *fliC*, as well as MLST ST678
25th–26th of May	RKI releases warnings on consuming cucumbers and other vegetables UKM detects Stx (*stx2*) by PCR and sequencing, as well as ESBL phenotype
28th of May	STEC TY2482 DNA arrives at BGI in cooperation with UKE
30th of May	UKM releases rapid multiplex PCR molecular diagnostic test for STEC O104:H4 based on *stx2*, *terD*, *rfb* O104, and *fliC* H4 (Bielaszewska et al., 2011a). Later published in a modified form as a real-time multiplex PCR approach (Zhang et al., 2012). STEC LB226692 WGS on Ion Torrent PGM (Life Tech) begins in cooperation with UKM
2nd of June	BGI releases five Ion Torrent PGM runs of TY2482 on its website with a CC0 license (Li et al., 2011) Nick Loman releases first TY2482 *de novo* assembly with the BGI Ion Torrent data on his blog (3,057 contigs) Sequencing of related historic STEC O104:H4 01-09591 (HUSEC041) ST678 from 2001 begins at UKM
3rd of June	Automatic annotation of Loman's assembly by Raquel Tobes (Manrique et al., 2011) BGI releases two more Ion Torrent PGM runs of TY2482 plus reference-guided (mapping) and *de novo* assembly of unmapped reads (Li et al., 2011) STEC LB226692 sequencing finished at Life Tech and UKM, reads from the eight runs released

Continued on next page.

Table 6: Timeline of major genomic and epidemiological events during the 2011 German O104:H4 STEC epidemic (continued).

DATE	EVENT
4th of June	Reference-guided (mapping) plus *de novo* of unmapped reads draft assembly of LB226692 released from UKM to public on NCBI's Genbank (364 contigs)
5th of June	Fenugreek sprouts suspected as outbreak source by epidemiological detective work
6th of June	BGI releases hybrid *de novo* assembly of TY2482 from Ion Torrent PGM and Illumina HiSeq 2000 (shotgun) data (451 contigs) (Li et al., 2011) UKM announces completion of STEC 01-09591 (HUSEC041) sequencing, but no public release STEC GOS1/2 DNA arrives at the G2L
7th of June	BGI releases PCR typing scheme based on the WGS data with *stx2* and *aggCD* of AAF/I (Qin et al., 2011)
10th of June	HPA releases shotgun and paired-end 454 GS Junior data of STEC H112180280 and *de novo* assembly (13 scaffolds) STEC O104:H4 detected in sprout leftovers via PCR GOS1/2 sequencing and assembly finished at G2L (171/204 contigs)
11th of June	BGI releases 2nd hybrid *de novo* assembly of TY2482 from Ion Torrent PGM and Illumina HiSeq 2000 (shotgun and paired-end, insert size 500 bp) data (452 scaffolds) (Li et al., 2011)
16th of June	BGI releases closed TY2482 genome on website based on Illumina paired-end reads with three different insert sizes (500 bp, 2 kb, 6 kb) (Li et al., 2011) G2L releases sequencing reads and annotated assemblies on FTP server
20th of June	HPA releases four more STEC isolates Illumina MiSeq data and assemblies for STEC H112180280 and four other isolates
29th of June	Brzuszkiewicz et al. (2011) published online
6th of July	PacBio releases data for STEC C227-11 outbreak strain and its *de novo* assembly (33 contigs), and seven diarrhea-associated O104:H4 EAEC and four reference EAEC of other serotypes
4th of July	Last outbreaks cases reported

Continued on next page.

Table 6: Timeline of major genomic and epidemiological events during the 2011 German O104:H4 STEC epidemic (continued).

DATE	EVENT
20th of July	Mellmann et al. (2011) published and release of historic STEC H11218028 genome assembly (mapping and *de novo*) and data (456 contigs)
27th of July	The New England Journal of Medicine publishes the crowdsourced–UKE/BGI initiative (Rohde et al., 2011) and PacBio (Rasko et al., 2011) publications

An EAEC ancestor to the 2011 German outbreak strain must have acquired a phage encoding for the Stx through HGT. The detailed genomic analyses identified some more putative VFs of the outbreak strain (Section 5.2 on page 104) (Brzuszkiewicz et al., 2011). Thus, the combination of specific adhesion, Stx2 production, carriage of three SPATEs (*pic*, *sepA*, and *sigA*) implicated in mucosal damage and colonization, helps to explain the exceptional virulence of the 2011 STEC outbreak strain (Figure 10 on the facing page) (Karch et al., 2012; Rasko et al., 2011). This rare hybrid strain is a prime example of the dynamic nature and high plasticity of the *E. coli* genome and showcases the needed vigilance in emerging and re-emerging food-borne pathogens (Section 6.3.1 on page 203) (Bielaszewska et al., 2011a; Croxen et al., 2013). Luckily, the University of Münster quickly developed a diagnostic PCR scheme for detection of the O104:H4 hybrid STEC pathotype, which was then used in German diagnostic laboratories throughout the epidemic (Table 6 on page 198) (Bielaszewska et al., 2011a).

6.2.3 *Future vigilance for O104:H4 STEC*

historic O104:H4 STEC isolates

Prior to 2011 only few reports were available of a pathogenic agent being a Stx-encoding EAEC, that caused hemorrhagic colitis and HUS (Croxen et al., 2013; Morabito et al., 1998). But the German outbreak was not the first occurrence of O104:H4 STEC strains, actually several smaller outbreaks of HUS cases already occurred before 2011. Examples are the historic *E. coli* 01-09591 (HUSEC041)[4] isolated 2001 in Germany, and several cases in France, Korea, Italy, the Republic of Georgia, Tunisia, and Finland (Ahmed et al., 2012; Bae et al., 2006; Denamur, 2011; Grad et al., 2013; Guy et al., 2013; Karch et al., 2012; Monecke et al., 2011). The *after-epidemic* WGS efforts into these historic STEC/EAEC hybrid pathotype strains allowed insights into population

[4] STEC 01-09591 (HUSEC041) is a member of the HUS-associated *E. coli* collection (HUSEC) (Mellmann et al., 2008) of the national consulting laboratory for HUS at the UKM. It was sequenced during the 2011 German outbreak (see Table 6) (Mellmann et al., 2011).

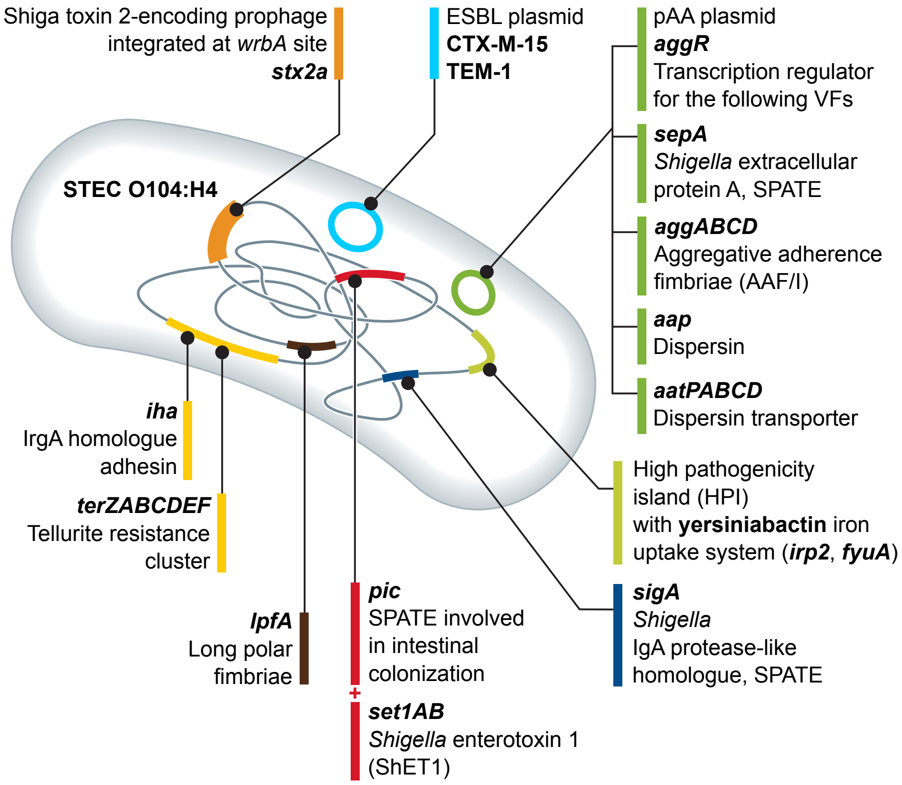


Figure 10: Overview of the main VFs and antibiotic resistance determinants included in the genome of the German 2011 O104:H4 STEC outbreak strain. These genes most likely contributed to the high virulence of the strain. Figure adapted from (Karch et al., 2012) using Inkscape.

genomics, HGT of VFs, and evolutionary processes[5] shaping this highly virulent *E. coli* pathotype (Grad et al., 2012, 2013; Guy et al., 2013). Thus, the STEC/EAEC hybrid was an established pathogen already circulating in the human population most likely in several different O104:H4 STEC lineages, albeit its rarity in causing diseases (Bielaszewska et al., 2011a; Monecke et al., 2011). The circulation of the hybrid pathotype is also exemplified by two O104:H4 STEC sporadic cases different from the German outbreak strain, during the 2011 epidemic, as well as cases in France and Turkey after the outbreak (Grad et al., 2013; Tietze et al., 2015).

All of these O104:H4 lineages most likely originated from a O104:H4 EAEC ancestor and the emergence of the outbreak strain depended on the acquisition of the Stx2 prophage, an antibiotic ESBL resistance plasmid, and replacing the AAF/III fimbriae pAA plasmid with a plasmid encoding the rarer AAF/I fimbriae (Section 5.2 on page 104) (Brzuszkiewicz et al., 2011; Rasko et al., 2011; Rohde et al., 2011). It is, however, not clear what bacterial traits caused the major outbreak in Germany in contrast to other smaller scale cases. But, it emphasizes the possibility of future food-borne *E. coli* O104:H4 outbreaks (Grad et al., 2013). Also, the natural habitat of the outbreak strain is not clear, as typically, the *E. coli* EAEC pathotype is not associated with zoonotic infections but rather has a human reservoir, in contrast to EHEC which are asymptomatically associated with ruminants, especially cattle (Croxen et al., 2013).

6.3 FLEXIBILITY OF *E. COLI* GENOMES AND PATHOTYPES

The genomic flexibility of "modern" *E. coli* strains was and maybe still is highly undervalued. Thus, the famous quote

> "Anything found to be true of *E. coli* must also be true of elephants." – Jacques Monod

can also be interpreted as the vastness of *E. coli*'s possibilities and adaptability, even in comparison to large multicellular organisms. Two good examples are described in this thesis: The unprecedented German STEC outbreak 2011 with an unusual hybrid pathotype of EAEC and EHEC (Section 5.2 on page 104) (Brzuszkiewicz et al., 2011) and the high genomic diversity of bovine *E. coli* isolates, their potential to cause bovine mastitis, and the resulting absence of a mammary pathogenic *E. coli* (MPEC) pathotype (Section 5.3.4 on page 140) (Leimbach et al., 2017).

[5] E. g. STEC 01-09591 (HUSEC041) does not encode for the ESBL CTX-M-15 and contains AAF type III (as does EAEC O104:H4 55989) instead of AAF/I. The *E. coli* 01-09591 pAA plasmid also contains the enteroaggregative heat-stable enterotoxin 1 (EAST1; *astA*), which is not present in the 2011 clonal outbreak strains (Mellmann et al., 2011).

6.3.1 *Ambiguous and emerging (hybrid)* E. coli *pathotypes*

The advancement in sequencing technologies with highly increased throughput and specificity (Section 1.1.1 on page 5) has shattered many older paradigms on *E. coli* pathogenicity. Especially, several traditionally well-defined *E. coli* pathotype definitions have taken a turn to more ambiguous definitions (Croxen et al., 2013). With the availability of thousands of *E. coli* genomes (Figure 5 on page 12) and realization of the unlimited and open *E. coli* pan-genome (Figure 6 on page 14) it is now clear that VFs, that were historically deemed to be pathotype-specific, can actually be integrated into different pathogenic or commensal *E. coli* genomic backgrounds. This is e. g. the case for several autotransporter (AT) proteins that initially were deemed markers for *E. coli* pathotypes, but actually are present in diverse strains of the *E. coli* population and rather correlate with phylogenetic background (Section 5.1 on page 79) (Zude et al., 2014). Even phenotypically defined IPEC pathotypes can be difficult to pin down genetically, if analyzed in adequate genome numbers, as is the case for atypical enteropathogenic *E. coli* (EPEC), diffusely adherent *E. coli* (DAEC), and adherent invasive *E. coli* (AIEC)[6] (Ingle et al., 2016b; O'Brien et al., 2016; Robins-Browne et al., 2016).

The emergence of "new" hybrid pathotypes additionally burdens the already complicated classification system (Croxen et al., 2013): e. g. typical and atypical EPEC[7] (Hazen et al., 2016; Ingle et al., 2016b) or EAEC[8], STEC and EHEC[9], and enteroinvasive *E. coli* (EIEC) and *Shigella*. These accommodations of newly found strains that do not fit the traditional classification are the direct result of the increased awareness of *E. coli*'s unexpected genotypical diversity.

It can be assumed that all hybrid combinations of VFs in individual *E. coli* genomes are possible as long as they can be acquired via HGT, are supported by their natural habitats (mainly animal intestines), and do not succumb to evolutionary pressure like maintenance cost for the strain. Recently, there have been reports on several mixed *E. coli* pathotypes causing disease in humans, e. g. hybrid ETEC/STEC strains causing both HUS or diarrhea (Leonard et al., 2016; Nyholm et al., 2015), UPEC harboring typical IPEC VFs[10] (Toval et al., 2014a,b), or a heteropathogenic O2:H6 STEC expressing both IPEC and ExPEC VFs and causing both diarrhea or urinary tract infections (UTIs) (Bielaszewska et al., 2014). The possession of VFs outside the pathotype classification

[6] AIEC are associated with Crohn's disease, which is a chronic inflammatory bowels disease.

[7] Strains encoding for LEE, but being either *bfp*-positive or -negative EPEC. *bfp* is the plasmid-encoded bundle-forming pilus operon.

[8] EAEC encoding for the master regulator *aggR* or not.

[9] EHEC are classically defined as being associated with hemorrhagic colitis and HUS, but on the molecular level as LEE-positive and Stx-encoding *E. coli*. STEC are LEE-negative.

[10] These UPEC strains include VFs of EAEC, STEC, and atypical EPEC.

might even enhance virulence properties, as is the case for EAEC causing a UTI outbreak in Denmark (Boll et al., 2013; Olesen et al., 2012). It is not surprising then, that natural infections in individuals can be caused by multiple phylogenetically diverse pathogenic *E. coli* with variable VF-content, e. g. cholera-like diarrhea caused by diverse ETEC in Dhaka Bangladesh (Sahl et al., 2015).

Many hybrid strains might circulate in human or non-human habitats in small enough numbers to evade detection by conventional diagnostic techniques. Thus, there is a needed vigilance for newly evolved or already circulating hybrid *E. coli* pathotypes that can become pathogenic or make the jump to the human host (Croxen et al., 2013). Nevertheless, the majority of *E. coli* infections are still caused by known culprits (e. g. EHEC O157:H7 or UPEC with traditional VFs) (Croxen et al., 2013; Köhler and Dobrindt, 2011; Leimbach et al., 2013) and pathotype classifications are very useful in a clinical setting for determining the potential severity of an infection and the best course of treatment. Furthermore, many hybrid strains only rarely cause human infections and therefore seem to be inhibited in their potential to colonize the human host (Robins-Browne et al., 2016). This difference might be a consequence of the gene content of *E. coli* strains in relation to their phylogenetic background.

6.3.2 *Impact of phylogenetic genealogy on gene content*

Interspecies recombination between extant *E. coli* of different phylogroups is restricted. Closely related *E. coli* genomes exchange more genetic material (especially within phylogroups) than distantly related ones (Didelot et al., 2012a; Leopold et al., 2011; McNally et al., 2013). Phylogenetic background of *E. coli*, therefore, has a large impact on gene content, as was shown for the distribution of AT proteins (Section 5.1 on page 79) and bovine-associated *E. coli* (Section 5.3.4 on page 140) in our studies (Leimbach et al., 2017; Zude et al., 2014). Thus, it is important to consider the phylogenetic background of *E. coli* strains to determine if putative functional convergence is not actually of phylogenetic nature. Concordantly, several excellent recent studies with large *E. coli* genomic sample sizes could show that extant typical and atypical EPEC, as well as ETEC emerged in different phylogenetic lineages and a global clonal expansion of these lineages occurred. These groups retain lineage-specific gene content acquired since the last common ancestor – possibly a consequence of restricted recombination between phylogroups (Hazen et al., 2013; Ingle et al., 2016b; von Mentzer et al., 2014). Because *E. coli* phylogroups have different prevalences with animals or humans, gene content associated with phylogeny might be a consequence of niche adaptation (Tenaillon et al., 2010). Consequently, the amount of recombination between strains might be a result of overlapping habitats. As inter- and intragroup recombination is the cohesive force to counteract divergence, *E. coli* strains could undergo disconnected evolution within these habitats.

Several pathotypes are in need of an update in regard to population structure and evolutionary relationships to understand their emergence and epidemiology. Especially those defined by the absence of VFs or with VFs prevalent on MGEs prone to HGT, instability, and deletion. As classical pathotype classifications are incomplete and potentially misleading, a more informative and accurate diagnostic bacteriology is required (Hazen et al., 2013; Ingle et al., 2016b; von Mentzer et al., 2014). Therefore, traditional clinical diagnostics to detect few classically associated VFs via PCR or serotyping needs to be enhanced with modern techniques like WGS that have no detection bias, higher sensitivity, and accuracy (Chapter 7 on page 219) (Robins-Browne et al., 2016). With the change of microbiology (and overall biological sciences) from a hypothesis- to a more data-driven scientific enterprise (van Helden, 2013) one could argue:

> "What better place than here, what better time than now?"[11]

6.4 IMPLICATIONS AND OUTLOOK ON BOVINE *E. COLI* MASTITIS

The *multifactorial etiology* of bovine mastitis is a major challenge for disease prevention and suitable treatment of afflicted animals (Ganda et al., 2016). This is especially true for environmental mastitis pathogens like *E. coli*, where *cow factors* and environmental conditions have been implicated in deciding the course of the infection (Section 1.5.1 on page 50) (Burvenich et al., 2003). There is no "easy" solution for the control of *E. coli* bovine mastitis.

6.4.1 *Current shortcomings of mastitis-associated* E. coli *genomics*

In Leimbach et al. (2017) (Section 5.3.4 on page 140) we could show that there is no evidence for the MPEC pathotype proposed by Bradley and Green (2001) and Shpigel et al. (2008). In our strain panel of 16 MAEC there was no virulence-associated gene that was significantly enriched in comparison to the nine bovine commensal isolates. However, if a particular gene plays an important role in *E. coli* mastitis-mediated disease, then one would expect it to have a wide distribution among MAEC in contrast to commensals. Our results, therefore, support the high geno- and phenotypical diversity of bovine *E. coli* regardless of isolation source (udder or fecal commensal) (Houser et al., 2008), and the opportunistic character of *E. coli* udder infections (Section 1.5.3 on page 55).

Nevertheless, several studies demonstrated an association of specific *E. coli* VFs with MAEC (Table 2 on page 58), albeit none of these

[11] Rage against the machine, "Guerrilla Radio" 1999.

VFs were predominantly found in the respective MAEC strains. Many of these publications have other shortcomings like the lack of an adequate strain sample size, small phylogenetic diversity, or lack of commensal comparator strains. The majority implicated a different set of VFs as putative MAEC factors and the few overlaps between them mostly result from investigations with similar restricted VF panels. However, a putative functional relatedness, like disease-associated VFs, can be clouded by a one-sided phylogenetic background of the strains and its impact on gene content (Section 6.3.2 on page 204). Because our Leimbach et al. (2017) study likewise includes a smaller sample size, we included an extensive analysis on the phylogenetic background of the strains' gene content. Additionally, our analysis resulted in a negative result (no VFs significantly enriched in MAEC), which is less likely to achieve with a small strain panel.

Within the high diversity of bovine *E. coli* there are also strains with a low ability to cause IMI. On the one hand, there is "environmental" *E. coli* isolate K71 that is non-pathogenic in the bovine udder (Blum et al., 2017). This might be a consequence of its LPS biosynthesis impairment, presumably resulting in a higher phagocytosis susceptibility by PMNs and a decrease of the inflammatory reaction. On the other hand, the non-pathogenic, laboratory adapted *E. coli* K-12 MG1655 strain can cause an inflammation reaction in a mastitis model mouse system (Blum et al., 2015). Whether such a result can be attributed to various *E. coli* lineages more adapted to elicit mastitis, as proposed by Kempf et al. (2016) and Goldstone et al. (2016) (Section 1.5.3 on page 55), seems unlikely. After all, *E. coli* MG1655 has a phylogroup A and K71 a phylogroup B1 background.

Mastitis incidence varies with differences in diet, environmental factors, and countries (Bean et al., 2004; Houser et al., 2008). Thus, future studies (genomic and phenotypic ones) should include isolates from different herds, farms, management styles, countries, and with a suitable phylogenetic diversity to properly represent the *E. coli* population and achieve the necessary statistical effect sizes. Considering the possibilities of HTS, I am confident future research will remedy the current shortcomings. These techniques in combination with open data sharing and cooperation between different research groups around the world would greatly advance bovine mastitis *E. coli* research. The Global Enteric Multicenter Study (GEMS)[12] funded by the Bill & Melinda Gates Foundation could serve as an example for such a collaboration (Kotloff et al., 2013). Other studies can then draw from this resource and e. g. analyze isolates in more detail, just as Hazen et al. (2016) and Ingle et al. (2016b) did for hybrid IPEC pathotypes (Section 6.3.1 on page 203).

[12] An epidemiological study drawing fecal samples of children with and without moderate to severe diarrhea.

6.4.2 *Prevention and treatment strategies for* E. coli *mastitis*

The list of potential prevention and treatment strategies for bovine *E. coli* mastitis is long. Despite many measures like vaccinations and extensive antibiotic usage, mastitis is not fully under control and alternative strategies/novel therapeutic approaches are needed that do not affect public health (Bouchard et al., 2015). The focus of modern mastitis control in dairy herds relies on prevention rather than imperfect treatment regimes (Suojala et al., 2013). An important approach to mastitis control is fast and sensitive microbial diagnostics. There have been major advancements in diagnostic techniques in recent years, that, however, have not found their way into veterinary practice yet (Chapter 7 on page 219).

6.4.2.1 *Improving "cow" and "environmental" factors*

Considering the ubiquity of *E. coli* in bovine feces, its opportunistic character of IMIs, and the importance of cow and environmental factors in the development of the disease (Section 1.5.1 on page 50), mastitis management should be approached in a holistic manner. Herd management needs to be scrutinized and farmers adequately informed. Cow factors can be improved by reducing stress, like providing a clean, dry, cool, and comfortable environment, and optimal nutrition. Thorough hygiene procedures during milking and in the farm environment, but also using bedding systems that reduce fecal contamination of the udder, are important elements of environmental mastitis control (Blowey and Edmondson, 2010; Bradley, 2002; Hogan and Larry Smith, 2003). Especially during the periparturient period, where cattle are most susceptible to *E. coli* mastitis (Section 1.5.3 on page 55), factors that maximize the cow's own defenses (physical integrity of the teat and the immune system) need to be supported.

The genetic selection for increased milk yield and composition have resulted in dairy cows with higher susceptibility to mastitis. High milk production subjects the cows to additional physiological stress and dilutes the humoral and cellular immune defense systems in the udder. Variations in mastitis prevalence exists between breeds and individual cows[13], therefore a selection for animals with a higher resilience against mastitis could counteract this unwanted genetic drift (Blowey and Edmondson, 2010; Rainard and Riollet, 2006).

The innate immunity reaction as a consequence of an *E. coli* IMI might also be modulated. Treatment with lactoferrin has been proposed as it not only inhibits bacterial growth, but also neutralizes LPS and may dampen the overshooting inflammatory response during acute mastitis (Rainard and Riollet, 2006). New techniques like the CRISPR-Cas9

[13] For example a single single nucleotide polymorphism (SNP) in a chemokine receptor is associated with impaired neutrophil migration and correlates with frequency of subclinical mastitis in Holstein cows (Rainard and Riollet, 2006).

system[14] can be used to genetically engineer the bovine mammary gland to produce further innate immune system components, like defensins (Section 1.5.2 on page 53) (Hyvönen et al., 2006; Rainard and Riollet, 2006).

Possible future intervention strategies to prevent bovine mastitis might be maintaining or strengthening the natural, endogenous mammary microbiota to prevent or at least mitigate *E. coli* IMI. Commensal strains can have a direct inhibitory effect on mastitis pathogens or result in competitive colonization exclusion. Mastitis is a dysbiosis of the udder microbiota and its balanced relationship with host structures in the mucosal niches. The result is a dramatically reduced bacterial diversity and altered microbial profile. Therefore, measures to regain this natural diversity after clinical or subclinical mastitis occurrences might have a positive effect and support spontaneous recovery (Falentin et al., 2016; Ganda et al., 2016). Bouchard et al. (2015) recently reported on commensal lactic acid bacteria from mammary teat canals that exhibit growth inhibition properties for the three main mastitis pathogens and advantageously modulate the bovine immune response[15]. Thus, these bacteria might be used as a probiotic to compete with pathogens for mammary gland colonization.

Herds with a low bulk SCC have a higher incidence of *E. coli* mastitis infections (Section 1.5.3 on page 55) and a higher SCC has been associated with increased mastitis resilience through a faster immune response and PMN migration (Blowey and Edmondson, 2010; Burvenich et al., 2003; Kornalijnslijper et al., 2004). The natural microbiota of the cow's udder might be needed to maintain leukocytes (and overall SCC) in the bovine udder at a balanced level. Thus, even (minor) mastitis pathogens (e. g. *Streptococcus* spp. and *Staphylococcus* spp.) can be members of the core/healthy udder microbiota and have a beneficial effect on immune capacities and ultimately on udder health (Oikonomou et al., 2012). Successful measures to decrease contagious mastitis (Section 1.5.1 on page 50) (Bradley, 2002; Hogan and Larry Smith, 2003), might have directly contributed to the increase in mastitis caused by environmental pathogens (especially *E. coli*) by ridding the bovine mammary tissue of its natural immune alertness. The unintentional negative effects of excessive antibiotic use in the dairy industry on the natural udder microbiota should not be underestimated (Ganda et al., 2016).

[14] The clustered, regularly interspaced, short palindromic repeat-CRISPR associated proteins (CRISPR-Cas) system is an adaptive immunity system of prokaryotes to detect and cleave foreign phage and plasmid DNA. Because of its unparalleled ease of use, speed, and precision the CRISPR-Cas9 system can be used for genome editing (Doudna and Charpentier, 2014).

[15] These *Lactobacillus* and *Lactococcus* strains showed anti-inflammatory effects, like a decrease in IL-8 secretion of mammary epithelial cells *in vitro*.

6.4.2.2 *Antibiotic treatment and resistances of* E. coli *mastitis*

Many improvements in decreasing the incidence and prevalence of mastitis have been on the back of widespread and unsustainable antibiotic usage, e. g. (prophylactic) whole herd dry period therapy (Section 1.5.1 on page 50) (Bradley, 2002; Hillerton and Berry, 2005). About 80% of the total antibiotics used in the dairy industry are prescribed for prophylaxis and treatment of bovine mastitis, despite varying degrees of effectiveness (Blowey and Edmondson, 2010; Ganda et al., 2016). Broad-spectrum antimicrobial agents are usually administered to treat coliform mastitis, like fluoroquinolones and third- or fourth-generation β-lactam cephalosporins, although current guidelines do not recommend intramammary antibiotic use for Gram-negative mastitis (Fairbrother et al., 2015; Ganda et al., 2016; Hillerton and Berry, 2005; Suojala et al., 2013).

While the success of antibiotics in battling bovine mastitis elicited by contagious pathogens is unquestioned, there is no convincing evidence that they are beneficial for the treatment of *E. coli* IMIs. In fact, intramammary coliform bacteria only have a low response rate to antimicrobial treatment – the effect on shortening the duration of IMI and pathogen clearance rate is minimal. Especially since clinical *E. coli* udder infections are usually transient with a high spontaneous cure rate, antibiotic treatment seems to be pointless (Ganda et al., 2016; Hillerton and Berry, 2005; Hogan and Larry Smith, 2003; Suojala et al., 2013). Furthermore, there is an economic and public health risk of antibiotic residues in bulk milk for human consumption. *E. coli* IMI with mild and moderate signs should be treated with alternative, non-antimicrobial approaches (Section 6.4.2.1 on page 207). Nevertheless, systemic antimicrobial treatment remains irreplaceable during severe infections due to the risk of bacteremia (Hillerton and Berry, 2005; Suojala et al., 2013).

Not only non-responsiveness of coliform mastitis to antibiotic treatment but also the emergence of resistances have become a major concern on dairy farms. Mismanagement and overuse of antibiotics are critical issues in blunting these key weapons against animal and human diseases. Resistances to antibiotics have risen in parallel to antimicrobials commonly used in dairies and there is a risk of introducing resistant bacteria into the food chain via raw milk and raw milk products (Gomes and Henriques, 2016; Suojala et al., 2013). In particular broad-spectrum antibiotics increase the selection pressure on bacteria and promote the emergence of multidrug resistant strains. Because many are critical drugs for human medicine, their use should be limited to specific indications based on conclusive bacteriological diagnostics, at best performed directly on site (Chapter 7 on page 219). Antibiotic treatment could then be predominantly targeted at Gram-positive bacteria (Suojala et al., 2013). Additionally, the routine use of broad-spectrum antibiotics could be avoided by prescribing narrow-spectrum drugs based on the detailed genomic information of the pathogen.

Accordingly, MAEC carrying ESBL genes have been increasingly isolated from cows with mastitis milk[16]. ESBLs confer high levels of resistance to most β-lactams, including last generation cephalosporins. Because several isolates show a close phylogenetic relationship to epidemiological successful ESBL-carriers in humans, they illustrate the emerging dangers in antibiotic resistances in livestock and possible introduction into the food chain (Dahmen et al., 2013; Locatelli et al., 2009). Several of the ESBL genes were found on conjugative plasmids, especially in combination with other resistance markers, which promotes co-selection of the resistances and HGT (Freitag et al., 2016).

Several of the 25 bovine-associated *E. coli* strains analyzed in Leimbach et al. (2017) (Section 5.3.4 on page 140) contain genes conferring antibiotic resistances (Table 7 on the facing page). Most common are resistance markers for aminoglycosides, β-lactams, sulfonamides, and tetracyclines in our strain panel. Several of these are present on the AMR-SSuT (antimicrobial multidrug resistance to streptomycin, sulfonamide, and tetracycline) resistance island described in Leimbach et al. (2017) (see Figure S4 of the publication's supplemental material on page 169). Highlights of multiresistance are six commensal strains RiKo 2299/09, RiKo 2305/09, RiKo 2308/09, RiKo 2340/09, RiKo 2351/09, and W26, which are all ESBL producers (Table 7). Interestingly, most of the resistant strains in this study are commensal fecal isolates and not MAEC, the reason for which is still unknown. Antimicrobial resistances to aminoglycosides, β-lactams, sulphonamides, and tetracycline were also detected in several previous studies on MAEC or bovine commensal isolates (Blum et al., 2008; Freitag et al., 2016; Ibrahim et al., 2016; Liu et al., 2014; Locatelli et al., 2009; Suojala et al., 2011). Although antibiotic resistances are not uncommon in commensal strains, they usually have no specific association with commensal isolates in comparison to MAEC (Blum et al., 2008). Our data and these studies highlight the problems in treatment of bovine mastitis with antibiotics and the tremendous potential for the transfer of multidrug resistance via HGT between commensal and pathogenic *E. coli* strains. There is an urgent need for proper prevention strategies and alternative treatments to further limit the use of antimicrobials in livestock.

[16] Different types of ESBLs have been found like $bla_{\text{CTX-M}}$ and bla_{TEM} (Dahmen et al., 2013; Freitag et al., 2016; Ghatak et al., 2013; Locatelli et al., 2009).

Table 7: Antibiotic resistance genes present in the bovine-associated *E. coli* strain panel of Leimbach et al. (2017). The genes were identified with the web server ResFinder (v2.1) (Zankari et al., 2012) and grouped according to the antibiotic class they confer resistance to.

STRAIN	PATHO*	PHYLO*	AMG*	BLA*	CMP*	MAC*	SUL*	TET*	TMP*	QUI*
1303	MAEC	A	*strAB*				*sul2*	*tetDCBR*		
131/07	MAEC	A								
2772a	MAEC	B1								
3234/A	MAEC	A								
AA86	commensal	B2		*bla*$_{\text{TEM-1}}$						
D6-113.11	MAEC	E								
D6-117.07	MAEC	A								
D6-117.29	MAEC	A	*strAB*				*sul2*	*tetDCBR*		
ECA-727	MAEC	A	*aphA7*, *strAB*				*sul2*	*tetDCBR*		
ECA-O157	MAEC	A	*aadA1*					*tetAR*		
ECC-1470	MAEC	B1								
ECC-Z	MAEC	A								
MPEC4839	MAEC	A								
MPEC4969	MAEC	B1								

Continued on next page.

Table 7: Antibiotic resistance genes present in the bovine-associated *E. coli* strain panel of Leimbach et al. (2017). The genes were identified with the web server ResFinder (v2.1) (Zankari et al., 2012) and grouped according to the antibiotic class they confer resistance to (continued).

STRAIN	PATHO*	PHYLO*	AMG*	BLA*	CMP*	MAC*	SUL*	TET*	TMP*	QUI*
O157:H43 T22	commensal	B1								
O32:H37 P4	MAEC	A								
P4-NR	MAEC	B1								
RiKo 2299/09	commensal	B1	*aadA5*, *aacC2*	$bla_{\text{CTX-M-15}}$, $bla_{\text{OXA-1}}$		*mphRA*	*sul1*	*tetAR*	*dfrA17*	*aac(6')Ib-cr*
RiKo 2305/09	commensal	B1	*aadA1*, *aadA5*, *aadB*	$bla_{\text{CTX-M-15}}$	*floR*	*mphRA*	*sul1*, *sul2*	*tetAR*	*dfrA17*	
RiKo 2308/09	commensal	A	*aadA1*, *aadB*, *aacC2*, *aphA7*, *strAB*	$bla_{\text{CTX-M-1}}$, $bla_{\text{TEM-1}}$	*floR*	*mphA*	*sul1*, *sul2*	*tetDCBR*		
RiKo 2331/09	commensal	B1								
RiKo 2340/09	commensal	A	*aadA1*, *aadB*, *aacC2*, *aphA7*, *strAB*	$bla_{\text{TEM-1}}$	*floR*		*sul1*, *sul2*	*tetDCBR*		

Continued on next page.

Table 7: Antibiotic resistance genes present in the bovine-associated *E. coli* strain panel of Leimbach et al. (2017). The genes were identified with the web server ResFinder (v2.1) (Zankari et al., 2012) and grouped according to the antibiotic class they confer resistance to (continued).

STRAIN	PATHO*	PHYLO*	AMG*	BLA*	CMP*	MAC*	SUL*	TET*	TMP*	QUI*
RiKo 2351/09	commensal	B1	*aadA1*, *strAB*	*bla*$_{\text{TEM-1}}$	*catA1*	*mphB*	*sul1*, *sul2*	*tetAR*	*dhfrI*	
UVM2	MAEC	A								
W26	commensal	B1	*aph(3')-Ia*, *strAB*	*bla*$_{\text{TEM-1}}$			*sul2*	*tetAR*		

* Abbreviations are as follows: Patho = pathotype, Phylo = phylogroup, AMG = aminoglycosides, BLA = β-lactams, CMP = chloramphenicol/florfenicol, MAC = macrolides, SUL = sulfonamides, TET = tetracycline, TMP = trimethoprim, QUI = quinolones.

Because of the extensive increase in multidrug-resistant bacteria one alternative treatment to antibiotics is experiencing a renaissance: bacteriophage therapy. Phage therapy is effective against multi-resistant bacterial infections and can serve as a surrogate in situations where treatment has run out of antibiotic options. However, phages have very specific targets and a suitable phage cocktail needs to be chosen – a task that can be facilitated by WGS of the bacterial pathogen (Section 7.1 on page 219) to detect outer membrane proteins (OMPs) that serve as attachment sites and phage immunity systems (like CRISPR-Cas, see footnote [14] on page 208). Narrow phage host restriction has the advantage of avoiding side effects on the natural microbiota (Matsuzaki et al., 2014). In a mammary setting additional research is required to investigate if there is a putative inhibition by milk constituents or if phage therapy can be effective against a highly diverse MAEC population (Gomes and Henriques, 2016).

6.4.2.3 *Vaccination against MAEC*

The holy grail in preventing infections and circumventing widespread antibiotics use is vaccination. An effective vaccine for bovine *E. coli* mastitis would induce the innate and adaptive immune response to synergistically elicit a sudden and effective reaction to eliminate invading *E. coli* (Rainard et al., 2016). This is a formidable challenge for the large antigenic and genetic diversity of environmental MAEC and so far vaccination strategies to successfully prevent *E. coli* mastitis have not fulfilled this promise.

vaccination for E. coli *bovine mastitis not effective*

A commercial "J5 core antigen" vaccine[17], that is administered subcutaneously, has been in use in North America for many years. However, its effect is only of short duration and it does not prevent clinical *E. coli* mastitis, although it decreases the severity of clinical signs. The short-lived protective immunity in the udder is a result of the body-udder barrier, where immunity acquired in the body is only partial and at a lower level present in the udder and similarly vice versa. The resulting failure in mounting an adequate and long-lasting immune response is another major challenge in the development of effective vaccines for *E. coli* IMI (Hogan and Larry Smith, 2003; Schukken et al., 2011).

The high genome plasticity of MAEC includes an inherent variability of OM antigenic structures like the O-antigen of LPS, flagellin etc. (Section 6.1 on page 195). Vaccination is therefore expected to ever only produce a partial reduction in incidence and protection from a limited range of strains, if at all. Two recent studies on intramammary immunization with ultraviolet-killed MAEC ECC-Z and heat-killed MAEC O32:H37 P4 confirm this conclusion, as a less severe inflammation

[17] This is a whole cell bacterin vaccine from heat-killed O111:B4 *E. coli* J5. *E. coli* J5 is a rough mutant, that lacks the O-antigen moiety of LPS (González et al., 1989; Schukken et al., 2011).

response and partial protection from IMI was only present for the challenge with identical strains (Herry et al., 2017; Pomeroy et al., 2016).

6.4.3 *Outlook on* E. coli *mastitis research*

There are still many open questions regarding the aetiology and especially the vastly different disease pathologies of *E. coli* IMI. Unraveling these should help dairy farmers to focus on the most important areas for mastitis prevention and control.

In this context it is important to consider that a pathogen (and its virulence determinants) is only pathogenic within the background of a susceptible host and thus the host-pathogen interaction always needs to be evaluated simultaneously (Pirofski and Casadevall, 2012). This is especially true for the opportunistic relationship between *E. coli* and bovine mastitis (Section 1.5.3 on page 55).

I want to touch shortly upon what microbial genomics and HTS can do in the future to expand our current knowledge on *E. coli* IMI.

GENOMICS Because, the genomic plasticity of bovine *E. coli* and their association with mastitis was long not appropriately addressed there is still much to learn from WGS of MAEC and commensal *E. coli*. Genomic analysis of the gene content of a large and diverse strain panel with MAEC and fecal bovine *E. coli*, as proposed in Section 6.4.1 on page 205, at best associated with phenotypical metadata of the corresponding disease progression, could finally settle the ambiguity of the proposed MPEC pathotype. It might also contribute to the understanding of mammary epithelial cell invasion properties of *E. coli* strains putatively associated with chronic/persistent IMIs (Section 1.5.3 on page 55) (Almeida et al., 2011; Dogan et al., 2006; Döpfer et al., 1999, 2000). One could also go beyond gene content analysis of MAEC isolates and analyze different alleles of orthologous proteins and intergenic (regulatory) regions in different strains with respect to virulence properties.

Shotgun metagenomics approaches (instead of the simpler descriptive 16S rRNA gene metabarcoding techniques) can be used to elucidate the relationship of pathogenic strains with the present natural microbiota during the colonization process. Prior to that the healthy bovine udder microbiota needs to be characterized further to understand its dynamics, especially in the face of the low genetic diversity of todays dairy cattle. I suppose that metagenomic sequencing with strain-level resolution would also detect some individual udder quarters infected with different *E. coli* strains at the same time, a fact that is often overlooked by "single-colony" analyses (Section 7.3 on page 222) (Scholz et al., 2016). It would also enable the detection of low-abundance organisms, especially in clinical mastitis with negative aerobic culture (Ganda et al., 2016).

TRANSCRIPTOMICS The ability to cause mastitis was shown to not be dependent on a difference in gene content between commensal *E. coli* and MAEC in this thesis (Section 5.3.4 on page 140) (Leimbach et al., 2017). However, *E. coli* strains that naturally inhabit the gastrointestinal tract have to adapt their metabolism and surface-exposed proteins (like adhesins) in order to colonize, persist, and propagate in the udder milieu (Section 1.5.3 on page 55). Many regulons have to be adjusted accordingly and the pathogenicity of MAEC might be a consequence of a more suitable transcriptional response in comparison to commensals without gross changes in gene content. RNA-Seq can be used to analyze genome-wide differential gene expression of bovine-associated *E. coli*, at best from *in vivo* IMI. This could be combined with dual RNA-Seq to analyze the host-pathogen transcriptional response simultaneously (Westermann et al., 2012, 2016). Additionally, transcriptomics is preferably combined with other "-omics" techniques, like proteomics and metabolomics, to get a full picture of the phenotypic adaptability of different bovine *E. coli* isolates. Finally, metagenomics HTS can be supported by metatranscriptomics to profile the transcriptional activity of MAEC in conjunction with the complex udder microbiota and interrogate strain activity *in vivo*.

Recently, several studies that examined clinical severity of ExPEC infections (both bacteremia and UTI) concluded that pathogenicity is not dependent on VF presence or genomic adaptation but, at least in UTI, on differential regulation of bacterial core functions in a subset of "urine-associated *E. coli* (UAEC)" (Landraud et al., 2013; Nielsen et al., 2016; Schreiber et al., 2017). Even small mutations can lead to a pathogenic phenotype by regaining functional transcription of a biosynthesis pathway or changing global transcription, as was described for the non-pathogenic laboratory workhorse *E. coli* K-12[18] (Browning et al., 2013; Koli et al., 2011).

MUTAGENESIS Considering the vast amount of research on bovine *E. coli* mastitis and the proposal of a MPEC pathotype with a restricted gene content (Section 1.5.3 on page 55) it is surprising that traditional microbiological reductionist approaches, like bacterial gene knockouts and complementations in MAEC or competition experiments in bovine or murine model systems, are not used more frequently in experimental mastitis assays. A high-throughput method that is particularly suitable to identify genome-wide bacterial determinants required for pathogenesis is sequencing transposon mutants (Tn-Seq). In this approach single-locus transposon insertion sites in thousands of mutants are tracked via HTS before and after a selective pressure has been put on the mutant library, e. g. growth in milk or infection of a

[18] Both reverting a small genetic lesion to repair LPS O-antigen biosynthesis or introducing a mutant histone-like protein, HU, that drastically changes the transcription profile, lead to a pathogenic lifestyle in *E. coli* K-12 (Browning et al., 2013; Koli et al., 2011).

model system. The fitness contribution of each gene can subsequently be determined by comparing the relative frequency of each mutant in the population after re-isolation (van Opijnen and Camilli, 2013). Such an experiment for MAEC 1303 with its finished-quality genome sequence (Section 5.3.2 on page 134) (Leimbach et al., 2015) has been started during this thesis.

7

THE FUTURE OF DIAGNOSTIC AND PUBLIC HEALTH MICROBIOLOGY

The nucleic acid sequence of clinical isolate genomes is inherently *data rich* and can provide (real-time) information on the isolation time, transmission route/network, selection pressure on the bacterial population, and current/future capacity for antibiotic resistances as well as virulence determinants (Croxen et al., 2013; Didelot et al., 2012b). Furthermore, digital sequence information has an intrinsic standardization and reproducibility (similar to MLST) and is comparable between different laboratories (Maiden, 2006; Pallen and Loman, 2011). Thus, it can be used to act not only upon single-patient level, but also across hospitals, non-nosocomial pathogens, and across countries for a greater public health benefit (as long as data are shared). As a consequence infection control can be better targeted and employed more effectively (Gardy et al., 2015). The technique is also universal and can be applied to monitor hospitals, patients, food, environments, and veterinary practice amongst others (Land et al., 2015). A wealth of information is currently lost in clinical microbiology diagnostic labs – sequencing all isolates and opening up the datasets would be a fantastic resource for population genomics, evolution, local/global epidemiology, and HGT of antibiotic/virulence determinants.

HTS diagnostics is universal

On several occasions during this discussion the current state of diagnostic tools has been questioned and the need for modern high-throughput techniques has been illustrated. Clinical microbiology relies on fast and accurate results for patient treatment, especially in time-sensitive outbreak settings.

7.1 WHY DO WE NEED TO MODERNIZE MICROBIAL DIAGNOSTICS AND EPIDEMIOLOGY?

In this thesis are several examples where traditional microbial diagnostics reached its *limits*:

- The hybrid STEC/EAEC pathotype during the 2011 German outbreak caused initially problems as detection assays in public health laboratories were not testing for the unusual O104:H4 serotype (Section 1.4 on page 48).

- It was shown that AT proteins cannot simply be used as markers for *E. coli* pathotypes without considering the phylogenetic background of the strains (Section 5.1 on page 79) (Zude et al., 2014).

- MAEC strain 1303 was initially wrongly serogrouped as O5 instead of O70 (Section 6.1 on page 195).
- We could not support the hypothesis of a MPEC pathotype for bovine mastitis, thus there are no genetic markers that can be used for detecting pathogenic *E. coli* associated with IMI (Section 5.3.4 on page 140) (Leimbach et al., 2017).

sequence-based diagnostics

Utilizing WGS of bacterial isolates with bioinformatical analyses and high-quality reference databases could resolve these drawbacks. The benefits of modern microbial diagnostics via HTS to rapidly identify, characterize, and monitor the spread of pathogenic *E. coli* are obvious – especially in the case of newly emerging hybrid *E. coli* pathotypes (Section 6.3.1 on page 203) (Robins-Browne et al., 2016).

An area where *sequence-based diagnostics* will have a tremendous impact are antibiotic resistances. Over- and misuse of antibiotics in human and veterinary medicine has lead to a rising threat of antibiotic resistant bacteria in many pathogens, including *E. coli* (Section 6.4.2.2 on page 209). The WHO urgently warned of a 'post-antibiotic era' in its global report on antimicrobial resistance surveillance 2014:

> "A post-antibiotic era – in which common infections and minor injuries can kill – far from being an apocalyptic fantasy, is instead a very real possibility for the 21st century."
> – WHO (2014)

In the report's action plan the WHO urges the scientific community to develop effective, rapid, and low-cost diagnostic tools to guide optimal antibiotic use in human and animal medicine. Evidence-based prescribing and dispensing of antibiotics should be the standard of care, whereas today antimicrobials are rarely prescribed based on a definitive diagnosis, especially in veterinary medicine (O'Neill, 2016; WHO, 2014).

7.2 CURRENT STATE OF CLINICAL MICROBIOLOGY

Sequencing will eventually replace most of the traditional diagnostics tests, as sequence-based *in silico* serotyping, antimicrobial susceptibility testing, and VF typing methods are already more accurate, exhaustive, and have a higher resolution than traditional phenotypic methods for *E. coli* (Figure 11 on the facing page) (Fratamico et al., 2016; Land et al., 2015; Loman and Pallen, 2015). Overall, genotype-to-phenotype relationships will be more easily predicted with the growth of corresponding, high-quality databases and improvements in analysis tools (Section 1.1.2 on page 9) (Didelot et al., 2012b), like the `ecoli_VF_collection` (Chapter 3 on page 65) (Leimbach, 2016b). Nevertheless, there will always be a place for phenotypical resistance diagnostic tests, albeit only in special cases e. g. where new resistances arise that are not present in the reference databases (Schürch and van Schaik, 2017).

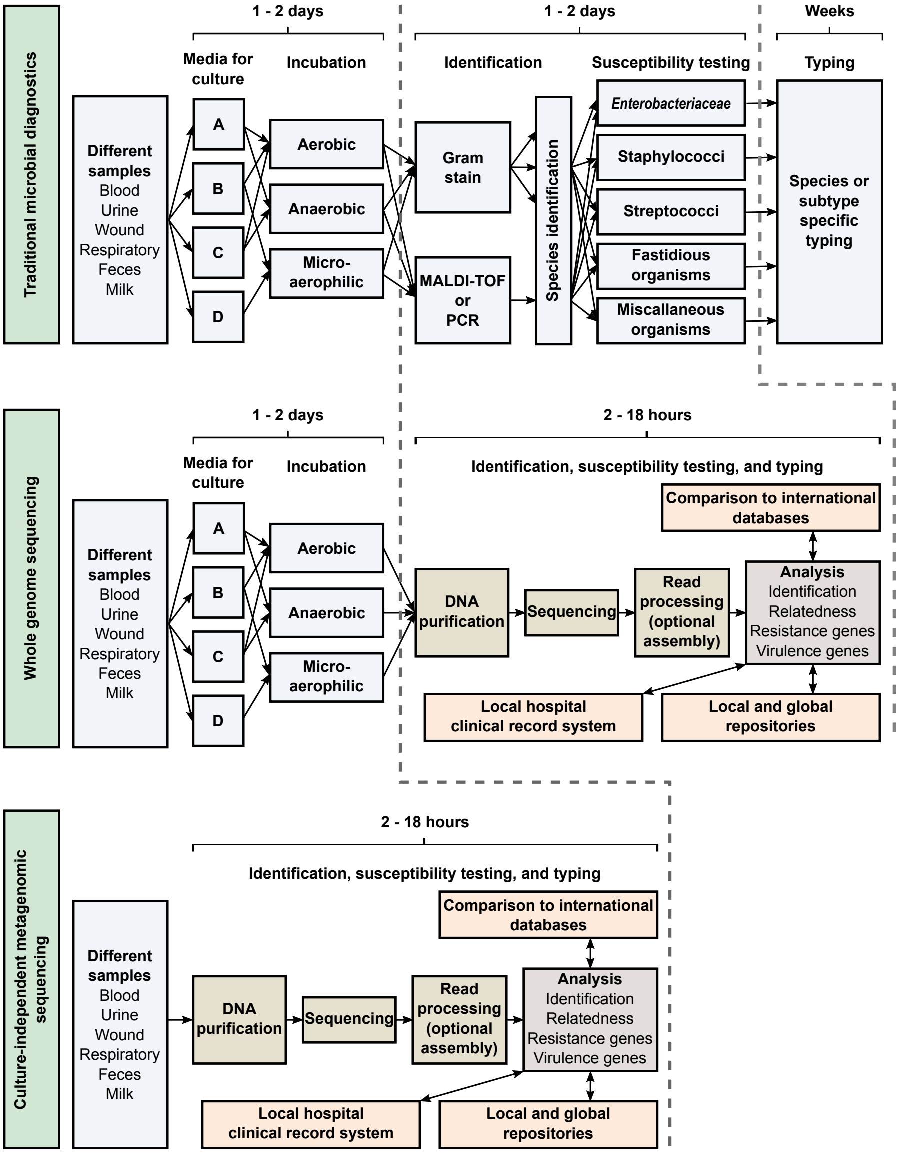


Figure 11: Comparison of conventional, isolate WGS-based, and culture-independent sequence-based microbial diagnostics workflows and estimated time frames. Central data repositories and reference databases will need to be constantly updated on the basis of new results, as indicated by the double-headed arrows. The schematic is an abbreviated form to highlight the central workflow steps. E. g. media for culturing are usually selective, followed by a Gram stain typically supplemented with colony morphology and biochemical tests to identify the pathogenic species. MALDI-TOF = matrix-assisted laser desorption ionization–time-of-flight mass spectrometry. Adapted from Didelot et al. (2012b) and Hasman et al. (2014) using `Inkscape`.

The Gastrointestinal Bacteria Reference Unit of Public Health England has implemented routine use of WGS since April 2014 as the reference laboratory for enteric bacterial pathogens in the UK. Surveillance of food-borne pathogens has been pioneered by the US Food and Drug Administration's (FDA) GenomeTrakr program[1], a multi-agency network which publishes its WGS results in a dedicated NCBI database (Luheshi et al., 2015; Quick et al., 2015; Schürch and van Schaik, 2017). In hospital-settings routine WGS for multi-drug surveillance is likewise feasible and already saves on costs (Mellmann et al., 2016).

real-time outbreak monitoring

The future of bacteriological epidemiology and diagnostics basically started with the 2011 German STEC outbreak and the first WGS analysis of a pathogen's genome in an *ongoing* epidemic (Section 6.2.2 on page 197). Rapid sequencing, the release of data with an *open license*, and utilizing digital age technology for collaboration and information dissemination made the "crowdsourced" analysis during the epidemic possible (Loman and Pallen, 2015).

The biggest transition challenges for public health laboratories are standardization of fast, meaningful, and robust bioinformatical algorithms (Section 1.1.2 on page 9), implementation of quality-control measures, applications to reliably handle the massive amounts of data, and approval from regulatory agencies. Analysis tools will need to inform non-bioinformaticians with a "push-button" fast and accurate diagnosis, and epidemiologists with timely data for tracking outbreaks in geospatial real time (Croxen et al., 2013; Land et al., 2015; Loman and Pallen, 2015; Pallen, 2016). This should work at best with hand-held computers, like the omnipresent smartphones.

There are already a couple successful examples of real-time bacterial isolate WGS diagnostics and epidemics interventions summarized in overview articles by Chan et al. (2012b), Didelot et al. (2012b), Fricke and Rasko (2014), and Hasman et al. (2014). The dropping costs for HTS (Section 1.1.3 on page 11) will make these practices more common in the near future, hopefully also outside academia.

7.3 WHAT CAN WE EXPECT IN THE FUTURE?

culture-independent *shotgun metagenomics diagnostics*

An exciting aspect of sequence-based microbial diagnostics is that it may eliminate the need for culturing bacterial isolates in the future altogether. Treatment strategies in a clinical setting are time-dependent, thus circumventing laboratory bacterial isolation in pure culture leads to faster test results for timely and informed clinical decisions (Hasman et al., 2014; Huang et al., 2017). *Culture-independent* sequence-based diagnostics rely on techniques established by metagenomics to sequence whole populations of bacteria from infection sites at the same time. Metabarcoding techniques of phylogenetically informative microbial

[1] https://www.fda.gov/food/foodscienceresearch/wholegenomesequencingprogramwgs/ucm363134.htm

genes (most prominently 16S rRNA amplicon sequencing[2]) can be used to characterize the microbial diversity within a habitat. The more detailed shotgun metagenomics approach, i. e. direct sequencing of DNA extracted from microbiologically complex samples without culturing or target-specific amplification, is able to infer the gene content and metabolic properties of a microbial community and their relationship towards each other as well as their environment (e. g. human or bovine host) (Pallen, 2014; Relman, 2013). This can be exploited for diagnostic purposes to detect and characterize the bacterial disease agents from a complex mix of organisms (Figure 11 on page 221) (Hasman et al., 2014; Pallen, 2014).

Culture-independent shotgun metagenomic diagnostics has several advantages in comparison to conventional, culture-based, selective, single-colony diagnostics[3] (Hasman et al., 2014; Huang et al., 2017; Loman et al., 2013; Pallen, 2014; Pallen and Loman, 2011; Pirofski and Casadevall, 2012):

- It avoids amplification bias and limited specificity of other culture-independent approaches, like immunoassays, in-situ hybridizations, and PCR.
- It has the potential to be totally unbiased in detecting pathogens (regardless if the pathogen is of bacterial, fungal, or viral origin) with total DNA/RNA sequencing, appropriate sequence coverage, and algorithms that can detect also subdominant pathogens (see below).
- It can elucidate polymicrobial and synergistic co-infections.
- It can detect variants in the overall pathogenic population (e. g. variable antibiotic resistance genotypes).
- It can discover uncultivable and unknown pathogens – "unknown unknowns" (Section 1.1 on page 3).
- Changes in the natural microbiota of the sample can be detected simultaneously, opening new possibilities for *microbiota diagnostics*[4].

A landmark study by Loman et al. (2013) showed that many microbial diagnostic prerequisites can be met via shotgun metagenomics. The authors were able to identify the O104:H4 STEC strain causing the 2011 German epidemic (Section 1.4 on page 48) by sequencing stool samples from patients, without the need of culturing or using any information learned through other means about the microbial diagnosis.

[2] 16S rRNA gene studies, however, have a limited resolution at the species and subspecies level, because of the high conservation of the gene's sequence (Huang et al., 2017).

[3] Where you can find only what you are looking for.

[4] Pathogen-specific signatures in the perturbed gut microbiota could work as diagnostic markers.

By iterative filtering of abundant human host and healthy fecal microbiota DNA, it was possible to obtain a draft assembly of the O104:H4 STEC outbreak strain. Additionally, the majority of positive fecal samples could be identified by mapping the metagenome sequence reads to a O104:H4 STEC reference genome. This approach was also able to subtype pathogens at a level sufficient for outbreak investigations, e. g. MLST STs (Huang et al., 2017).

strain-level resolution within metagenomics sequence data

Since then there have been major algorithmic improvements in *strain-level resolution* of metagenomic samples, concurrent with cheaper and deeper sequencing coverage of samples. Discriminating genomes in metagenome sequence reads has been advanced significantly in recent years, mostly by *binning algorithms*[5] (Brown, 2015). Shotgun metagenomic sequencing and pan-genome-based analysis e. g. successfully identified ExPEC carrying typical UPEC VFs as a risk factor for necrotizing enterocolitis in preterm infants (Scholz et al., 2016).

How far metagenomics can replace conventional culture methods for diagnostics remains to be seen, but improvements in sequencing technologies are continuing. The MinION nanopore sequencer from Oxford Nanopore Technologies has two advantages for diagnostics in comparison to other HTS machines (Section 1.1.1.1 on page 6). First, because the MinION fits in a pocket, it is possible to bring the sequencer to the sample instead of the sample to a laboratory (Erlich, 2015; Loman and Watson, 2015). Second, a unique property of the MinION is that sequence data is streamed to a computer during the run and can be analyzed in *real-time*, in contrast to the competition where a sequencing run first has to be completed. This ability was showcased in a hospital outbreak of *Salmonella enterica* serovar Enteritidis in the UK, where the MinION was able to acquire clinically relevant data within minutes, identify the bacterial species within 20 min, and determine the isolate to be part of the outbreak in less than 2 h after the sequencing run was started (Quick et al., 2015) (Figure 11 on page 221).

For some samples, like blood and urine (Hasman et al., 2014), a microfluidics-based system can be imagined in the future that is integrated with a nanopore machine to monitor macromolecules from both the pathogen and host. DNA, RNA, and proteins can then be assayed to investigate infection and inflammation all in one workflow and in real-time (Loman and Pallen, 2015). In the dairy industry about 40% of milk samples from cows with clinical mastitis have negative results by conventional aerobic culture. Also classical bacterial culture has a time delay of 24 – 48 hours to obtain results (Ganda et al., 2016; Oikonomou et al., 2012). Thus, a culture-independent nanopore-based streaming control of milk samples cow-side, maybe directly integrated in the milking machines on farms, to monitor both the host's inflamma-

[5] Metagenomic binning of reads or contigs uses shared frequencies across samples, e. g. sequence composition (GC content, tetranucleotide frequency, and codon usage . . .), sequence similarity, coverage, or specific marker genes in order to assign these to a species or strain (Marx, 2016; Peabody et al., 2015)

tion response and detect mastitis pathogens, is possible in the future. Such a real-time, point-of-care strain-level diagnosis would be an important component to treat and prevent specific pathogens[6] instead of general treatment, as well as antibiotic stewardship (Section 6.4.2 on page 207).

> Sequencing will not replace every aspect of culture-based clinical bacteriology. However, the sequencing revolution is here, we just need to embrace it!

[6] Especially in time sensitive circumstances like acute and severe *E. coli* mastitis.

Part V

APPENDIX

A

INDIVIDUAL AUTHOR CONTRIBUTIONS

This chapter describes the detailed individual author contributions for each publication included in the introduction on *E. coli* genomics (Section 1.2.1 on page 13) and the publications chapter (Chapter 5 on page 79). The first table lists the contributions for each part of the respective study, the second table for each figure and table (if present).

Furthermore, a statement is included on page 237 that legal second publication rights for the manuscripts were obtained, where necessary.

E. COLI AS AN ALL-ROUNDER: THE THIN LINE BETWEEN COMMENSALISM AND PATHOGENICITY

Author contribution tables for Leimbach et al. (2013) in Section 1.2.1 on page 13.

Table 8: Individual author contributions for each part of Leimbach et al. (2013).

PARTICIPATED IN	AUTHOR INITIALS*	
Study design & methods development	AL, UD	JH
Data collection	AL	
Data analysis & interpretation	AL	UD
Manuscript Writing		
Introduction	AL, UD	
Materials & methods	AL, UD	
Results	AL, UD	
Discussion	AL, UD	
First draft	AL, UD	JH

* Responsibility decreasing from left to right.

Table 9: Individual author contributions for the figures and table of Leimbach et al. (2013).

FIGURE/TABLE	AUTHOR INITIALS*
Figure 1	AL
Figure 2	AL
Table 1	UD

* Responsibility decreasing from left to right.

PREVALENCE OF AUTOTRANSPORTERS IN *ESCHERICHIA COLI*: WHAT IS THE IMPACT OF PHYLOGENY AND PATHOTYPE?

Author contribution tables for Zude et al. (2014) in Section 5.1 on page 79.

Table 10: Individual author contributions for each part of Zude et al. (2014).

PARTICIPATED IN	AUTHOR INITIALS*	
Study design & methods development	AL, IZ, UD	
Data collection	AL, IZ	
Data analysis & interpretation	AL, IZ	UD
Manuscript Writing		
Introduction	IZ	AL, UD
Materials & methods	AL, IZ	UD
Results	AL, IZ	UD
Discussion	AL, IZ	UD
First draft	AL, IZ	UD

* Responsibility decreasing from left to right.

Table 11: Individual author contributions for the figures/tables of Zude et al. (2014).

FIGURE/TABLE	AUTHOR INITIALS*	
Table 1	AL	
Figure 1	AL	
Table 2	AL	
Figure 2	AL	
Figure 3 A/B	IZ	AL
Figure 4	IZ	
Figure 5 A/B/C	IZ	
Figure 6 A/B/C	IZ	
Figure 7	IZ	
Figure 8	IZ	
Figure S1A/B	AL	
Figure S2A/B/C/D	AL	
Table S1	AL	IZ
Table S2	IZ	
Table S3	IZ	
Table S4	AL	IZ

Continued on next page.

Table 11: Individual author contributions for the figures/tables of Zude et al. (2014) (continued).

FIGURE/TABLE	AUTHOR INITIALS*
Table S5	AL

* Responsibility decreasing from left to right.

GENOME SEQUENCE ANALYSES OF TWO ISOLATES FROM THE RECENT *ESCHERICHIA COLI* OUTBREAK IN GERMANY REVEAL THE EMERGENCE OF A NEW PATHOTYPE: ENTERO-AGGREGATIVE-HAEMORRHAGIC *ESCHERICHIA COLI* (EAHEC)

Author contribution tables for Brzuszkiewicz et al. (2011) in Section 5.2 on page 104.

Table 12: Individual author contributions for each part of Brzuszkiewicz et al. (2011).

PARTICIPATED IN	AUTHOR INITIALS*			
Study design & methods development	AL, AT, EB	GG, JS, RD	HP, JB	
Data collection	AT, FM	HP, JB	AL	EB, JS
Data analysis & interpretation	AL, AT, EB, JS	GG, HL, RD		
Manuscript Writing				
Introduction	AL, AT, EB, GG, JS, RD	HL		
Materials & methods	AL, AT, EB, GG, JS, RD	HL		
Results	AL, AT, EB, GG, JS, RD	HL		
Discussion	AL, AT, EB, GG, JS, RD	HL		
First draft	AL, AT, EB, GG, JS, RD	HL	HP, JB	FM

* Responsibility decreasing from left to right.

Table 13: Individual author contributions for the figures/tables of Brzuszkiewicz et al. (2011).

FIGURE/TABLE	AUTHOR INITIALS*
Table 1	AL, AT, EB, JS

Continued on next page.

Table 13: Individual author contributions for the figures/tables of Brzuszkiewicz et al. (2011) (continued).

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	JS	AL
Figure 2	AL	
Figure 3	EB	AL, JS
Figure 4 a/b	EB	AL, HL, JS
Figure 5	EB	AL, AT, GG, JS, RD
Table S1	JS	AL, EB
Table S2	HL	EB
Table S3	EB	JS
Table S4	EB	JS
Figure S1	FM	AT
Figure S2 A/B	JS	HL
Figure S3	JS	AL

* Responsibility decreasing from left to right.

The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type

Author contribution tables for Duda et al. (2011) in Section 5.3.1 on page 119.

Table 14: Individual author contributions for each part of Duda et al. (2011).

PARTICIPATED IN	AUTHOR INITIALS*				
Study design & methods development	KD	OH	AL, BL	UD	HB
Data collection	KD	BL	AL	HB	
Data analysis & interpretation	KD	BL	AL, UD	OH	HB
Manuscript Writing					
Introduction	KD	OH			
Materials & methods	KD	BL	AL		
Results	KD	BL, UD	AL	HB, OH	
Discussion	KD	OH, UD	HB	AL, BL	

Continued on next page.

Table 14: Individual author contributions for each part of Duda et al. (2011) (continued).

FIGURE/TABLE	AUTHOR INITIALS*				
First draft	KD	OH	UD	HB	AL, BL, EB

* Responsibility decreasing from left to right.

Table 15: Individual author contributions for the figures/tables of Duda et al. (2011).

FIGURE/TABLE	AUTHOR INITIALS*	
Figure 1	not applicable	
Table 1	KD	OH
Figure 2 a/b	KD	OH
Figure 3	KD	OH
Figure 4 a/b	AL	UD
Table 2	AL	UD
Figure 5	HB	
Figure 6	KD	BL
Figure 7	KD	OH
Table S1	AL	UD

* Responsibility decreasing from left to right.

COMPLETE GENOME SEQUENCES OF *ESCHERICHIA COLI* STRAINS 1303 AND ECC-1470 ISOLATED FROM BOVINE MASTITIS

Author contribution table for Leimbach et al. (2015) in Section 5.3.2 on page 134.

Table 16: Individual author contributions for each part of Leimbach et al. (2015).

PARTICIPATED IN	AUTHOR INITIALS*	
Study design & methods development	AL	UD
Data collection	AL	
Data analysis & interpretation	AL	
Manuscript Writing		
Introduction	AL	

Continued on next page.

Table 16: Individual author contributions for each part of Leimbach et al. (2015) (continued).

FIGURE / TABLE	AUTHOR INITIALS*				
Materials & methods	AL				
Results	AL				
Discussion	AL				
First draft	AL	UD	AP	RD	AW, FS, YS

[*] Responsibility decreasing from left to right.

WHOLE-GENOME DRAFT SEQUENCES OF SIX COMMENSAL FECAL AND SIX MASTITIS-ASSOCIATED *ESCHERICHIA COLI* STRAINS OF BOVINE ORIGIN

Author contribution tables for Leimbach et al. (2016) in Section 5.3.3 on page 137.

Table 17: Individual author contributions for each part of Leimbach et al. (2016).

PARTICIPATED IN	AUTHOR INITIALS*				
Study design & methods development	AL	UD			
Data collection	AL				
Data analysis & interpretation	AL				
Manuscript Writing					
Introduction	AL				
Materials & methods	AL				
Results	AL				
Discussion	AL				
First draft	AL	UD	AP	RD	AW, HZ, NS, OW, WP

[*] Responsibility decreasing from left to right.

Table 18: Individual author contributions for the table of Leimbach et al. (2016).

TABLE	AUTHOR INITIALS*
Table 1	AL

[*] Responsibility decreasing from left to right.

NO EVIDENCE FOR A BOVINE MASTITIS *ESCHERICHIA COLI* PATHOTYPE

Author contribution tables for Leimbach et al. (2017) in Section 5.3.4 on page 140.

Table 19: Individual author contributions for each part of Leimbach et al. (2017).

PARTICIPATED IN	AUTHOR INITIALS*			
Study design & methods development	AL	UD		
Data collection	AL			
Data analysis & interpretation	AL			
Manuscript Writing				
Introduction	AL			
Materials & methods	AL			
Results	AL			
Discussion	AL			
First draft	AL	UD	JV	AP, RD

* Responsibility decreasing from left to right.

Table 20: Individual author contributions for the datasets/figures/tables of Leimbach et al. (2017).

FIGURE/TABLE	AUTHOR INITIALS*		
Table 1	AL		
Figure 1	AL		
Figure 2	AL		JV
Figure 3 A/B	AL		
Figure 4	AL		
Figure 5 A/B	AL		
Table 2	AL		
Figure 6	AL		
Table S1	AL		
Table S2	AL		
Table S3	AL		AP
Table S4	AL		
Table S5	AL		
Table S6	AL		
Table S7	AL		

Continued on next page.

Table 20: Individual author contributions for the datasets/figures/tables of Leimbach et al. (2017) (continued).

FIGURE/TABLE	AUTHOR INITIALS*	
Figure S1	AL	
Figure S2 A/B/C/D	AL	
Figure S3 A/B/C/D	AL	DG
Figure S4	AL	
Figure S5 A/B	AL	
Figure S6	AL	
Figure S7 A/B	AL	
Dataset S1	AL	DG
Dataset S2	AL	
Dataset S3	AL	
Dataset S4	AL	
Dataset S5	AL	
Dataset S6	AL	
Dataset S7	AL	
Dataset S8	AL	
Dataset S9	AL	
Dataset S10	AL	
Dataset S11	AL	
Dataset S12	AL	
Supplemental Material & Methods	AL	
Dataset S13	AL	
Dataset S14	AL	

* Responsibility decreasing from left to right.

CONFIRMATION

The doctoral researcher confirms that he has obtained permission from both the publishers and the co-authors for legal second publication.

The doctoral researcher and the primary supervisor confirm the correctness of the above mentioned assessment in this chapter.

Andreas Leimbach

Doctoral researcher's name Date, Place Signature

Prof. Dr. Ulrich Dobrindt

Primary supervisor's name Date, Place Signature

B

CURRICULUM VITÆ

BIBLIOGRAPHY

Abraham, S. et al. (2012). "Molecular characterization of commensal *Escherichia coli* adapted to different compartments of the porcine gastrointestinal tract." *Appl. Environ. Microbiol.* 78.19, pp. 6799–6803. DOI: 10.1128/AEM.01688-12.

Abreu, A. G. et al. (2013). "Autotransporter protein-encoding genes of diarrheagenic *Escherichia coli* are found in both typical and atypical enteropathogenic *E. coli* strains." *Appl. Environ. Microbiol.* 79.1, pp. 411–414. DOI: 10.1128/AEM.02635-12.

Ackermann, N. et al. (2008). "Contribution of trimeric autotransporter C-terminal domains of oligomeric coiled-coil adhesin (Oca) family members YadA, UspA1, EibA, and Hia to translocation of the YadA passenger domain and virulence of *Yersinia enterocolitica*." *J. Bacteriol.* 190.14, pp. 5031–5043. DOI: 10.1128/JB.00161-08.

Actis, L. A., M. E. Tolmasky, and J. H. Crosa (1999). "Bacterial plasmids: replication of extrachromosomal genetic elements encoding resistance to antimicrobial compounds." *Front. Biosci.* 4, pp. D43–62.

Adam, E. et al. (2010). "Probiotic *Escherichia coli* Nissle 1917 activates DC and prevents house dust mite allergy through a TLR4-dependent pathway." *Eur. J. Immunol.* 40.7, pp. 1995–2005. DOI: 10.1002/eji.200939913.

Adler, J., G. L. Hazelbauer, and M. M. Dahl (1973). "Chemotaxis toward sugars in *Escherichia coli*." *J. Bacteriol.* 115.3, pp. 824–847.

Ahmed, S. A. et al. (2012). "Genomic comparison of *Escherichia coli* O104:H4 isolates from 2009 and 2011 reveals plasmid, and prophage heterogeneity, including Shiga toxin encoding phage stx2." *PLoS ONE* 7.11, e48228. DOI: 10.1371/journal.pone.0048228.

Alikhan, N.-F. et al. (2011). "BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons." *BMC Genomics* 12, p. 402. DOI: 10.1186/1471-2164-12-402.

Allsopp, L. P. et al. (2010). "UpaH is a newly identified autotransporter protein that contributes to biofilm formation and bladder colonization by uropathogenic *Escherichia coli* CFT073." *Infect. Immun.* 78.4, pp. 1659–1669. DOI: 10.1128/IAI.01010-09.

Allsopp, L. P. et al. (2012). "Molecular characterization of UpaB and UpaC, two new autotransporter proteins of uropathogenic *Escherichia coli* CFT073." *Infect. Immun.* 80.1, pp. 321–332. DOI: 10.1128/IAI.05322-11.

Almeida, R. A. et al. (2011). "Intracellular fate of strains of *Escherichia coli* isolated from dairy cows with acute or chronic mastitis." *Vet. Res. Commun.* 35.2, pp. 89–101. DOI: 10.1007/s11259-010-9455-5.

Alsam, S. et al. (2006). "*Escherichia coli* interactions with *Acanthamoeba*: a symbiosis with environmental and clinical implications." *J. Med. Microbiol.* 55 (Pt 6), pp. 689–694. DOI: 10.1099/jmm.0.46497-0.

Altenhoefer, A. et al. (2004). "The probiotic *Escherichia coli* strain Nissle 1917 interferes with invasion of human intestinal epithelial cells by different enteroinvasive bacterial pathogens." *FEMS Immunol. Med. Microbiol.* 40.3, pp. 223–229. DOI: 10.1016/S0928-8244(03)00368-7.

Altschul, S. F. et al. (1990). "Basic local alignment search tool." *J. Mol. Biol.* 215.3, pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2.

Altschul, S. F. et al. (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* 25.17, pp. 3389–3402. DOI: 10.1093/nar/25.17.3389.

Ambur, O. H. et al. (2009). "Genome dynamics in major bacterial pathogens." *FEMS Microbiol. Rev.* 33.3, pp. 453–470. DOI: 10.1111/j.1574-6976.2009.00173.x.

Amor, K. et al. (2000). "Distribution of core oligosaccharide types in lipopolysaccharides from *Escherichia coli*." *Infect. Immun.* 68.3, pp. 1116–1124. DOI: 10.1128/IAI.68.3.1116-1124.2000.

Andrianopoulos, K., L. Wang, and P. R. Reeves (1998). "Identification of the fucose synthetase gene in the colanic acid gene cluster of *Escherichia coli* K-12." *J. Bacteriol.* 180.4, pp. 998–1001.

Anfora, A. T. and R. A. Welch (2006). "DsdX is the second D-serine transporter in uropathogenic *Escherichia coli* clinical isolate CFT073." *J. Bacteriol.* 188.18, pp. 6622–6628. DOI: 10.1128/JB.00634-06.

Anfora, A. T. et al. (2007). "Roles of serine accumulation and catabolism in the colonization of the murine urinary tract by *Escherichia coli* CFT073." *Infect. Immun.* 75.11, pp. 5298–5304. DOI: 10.1128/IAI.00652-07.

Angiuoli, S. V. and S. L. Salzberg (2011). "Mugsy: fast multiple alignment of closely related whole genomes." *Bioinformatics* 27.3, pp. 334–342. DOI: 10.1093/bioinformatics/btq665.

Archer, C. T. et al. (2011). "The genome sequence of *E. coli* W (ATCC 9637): comparative genome analysis and an improved genome-scale reconstruction of *E. coli*." *BMC Genomics* 12, p. 9. DOI: 10.1186/1471-2164-12-9.

Assefa, S. et al. (2009). "ABACAS: algorithm-based automatic contiguation of assembled sequences." *Bioinformatics* 25.15, pp. 1968–1969. DOI: 10.1093/bioinformatics/btp347.

Baba, T. et al. (2006). "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection." *Mol. Syst. Biol.* 2, p. 2006.0008. DOI: 10.1038/msb4100050.

Bae, W. K. et al. (2006). "A case of hemolytic uremic syndrome caused by *Escherichia coli* O104:H4." *Yonsei Med. J.* 47.3, pp. 437–439. DOI: 10.3349/ymj.2006.47.3.437.

Bankevich, A. et al. (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." *J. Comput. Biol.* 19.5, pp. 455–477. DOI: 10.1089/cmb.2012.0021.

Bannerman, D. D. (2009). "Pathogen-dependent induction of cytokines and other soluble inflammatory mediators during intramammary infection of dairy cows." *J. Anim. Sci.* 87 (13 Suppl), pp. 10–25. DOI: 10.2527/jas.2008-1187.

Barnhart, M. M. and M. R. Chapman (2006). "Curli biogenesis and function." *Annu. Rev. Microbiol.* 60, pp. 131–147. DOI: 10.1146/annurev.micro.60.080805.142106.

Barondess, J. J. and J. Beckwith (1995). "*bor* gene of phage lambda, involved in serum resistance, encodes a widely conserved outer membrane lipoprotein." *J. Bacteriol.* 177.5, pp. 1247–1253. DOI: 10.1128/jb.177.5.1247-1253.1995.

Bateman, A. et al. (2002). "The Pfam protein families database." *Nucleic Acids Res.* 30.1, pp. 276–280. DOI: 10.1093/nar/gkp985.

Bean, A., J. Williamson, and R. T. Cursons (2004). "Virulence genes of *Escherichia coli* strains isolated from mastitic milk." *J. Vet. Med. B Infect. Dis. Vet. Public Health* 51.6, pp. 285–287. DOI: 10.1111/j.1439-0450.2004.00772.x.

Bekal, S. et al. (2003). "Rapid identification of *Escherichia coli* pathotypes by virulence gene detection with DNA microarrays." *J. Clin. Microbiol.* 41.5, pp. 2113–2125. DOI: 10.1128/JCM.41.5.2113-2125.2003.

Bentley, S. D. and J. Parkhill (2015). "Genomic perspectives on the evolution and spread of bacterial pathogens." *Proc. Biol. Sci.* 282.1821, p. 20150488. DOI: 10.1098/rspb.2015.0488.

Benz, I. and M. A. Schmidt (1992). "Isolation and serologic characterization of AIDA-I, the adhesin mediating the diffuse adherence phenotype of the diarrhea-associated *Escherichia coli* strain 2787 (O126:H27)." *Infect. Immun.* 60.1, pp. 13–18.

Benz, I. and M. A. Schmidt (2011). "Structures and functions of autotransporter proteins in microbial pathogens." *Int. J. Med. Microbiol.* 301.6, pp. 461–468. DOI: 10.1016/j.ijmm.2011.03.003.

Berger, C. N. et al. (2010). "Fresh fruit and vegetables as vehicles for the transmission of human pathogens." *Environ. Microbiol.* 12.9, pp. 2385–2397. DOI: 10.1111/j.1462-2920.2010.02297.x.

Bernier-Fébreau, C. et al. (2004). "Use of deoxyribose by intestinal and extraintestinal pathogenic *Escherichia coli* strains: a metabolic adaptation involved in competitiveness." *Infect. Immun.* 72.10, pp. 6151–6156. DOI: 10.1128/IAI.72.10.6151-6156.2004.

Bernier, C., P. Gounon, and C. Le Bouguénec (2002). "Identification of an aggregative adhesion fimbria (AAF) type III-encoding operon in enteroaggregative *Escherichia coli* as a sensitive probe for detecting the AAF-encoding operon family." *Infect. Immun.* 70.8, pp. 4302–4311. DOI: 10.1128/IAI.70.8.4302-4311.2002.

Bernstein, H. D. (2015). "Looks can be deceiving: recent insights into the mechanism of protein secretion by the autotransporter pathway." *Mol. Microbiol.* 97.2, pp. 205–215. DOI: 10.1111/mmi.13031.

Bielaszewska, M. et al. (2005). "Phenotypic and molecular analysis of tellurite resistance among enterohemorrhagic *Escherichia coli* O157:H7 and sorbitol-fermenting O157:NM clinical isolates." *J. Clin. Microbiol.* 43.1, pp. 452–454. DOI: 10.1128/JCM.43.1.452-454.2005.

Bielaszewska, M. et al. (2011a). "Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study." *Lancet Infect. Dis.* 11.9, pp. 671–676. DOI: 10.1016/S1473-3099(11)70165-7.

Bielaszewska, M. et al. (2011b). "Chromosomal instability in enterohaemorrhagic *Escherichia coli* O157:H7: impact on adherence, tellurite resistance and colony phenotype." *Mol. Microbiol.* 79.4, pp. 1024–1044. DOI: 10.1111/j.1365-2958.2010.07499.x.

Bielaszewska, M. et al. (2014). "Heteropathogenic virulence and phylogeny reveal phased pathogenic metamorphosis in *Escherichia coli* O2:H6." *EMBO Mol. Med.* 6.3, pp. 347–357. DOI: 10.1002/emmm.201303133.

Blattner, F. R. et al. (1997). "The complete genome sequence of *Escherichia coli* K-12." *Science* 277.5331, pp. 1453–1462. DOI: 10.1126/science.277.5331.1453.

Blowey, R. and P. Edmondson, eds. (2010). *Mastitis control in dairy herds*. 2nd ed. Wallingford: CABI. DOI: 10.1079/9781845935504.0000. URL: http://www.cabi.org/cabebooks/ebook/20103163287.

Blum, S. et al. (2008). "Identification of a bovine mastitis *Escherichia coli* subset." *Vet. Microbiol.* 132 (1-2), pp. 135–148. DOI: 10.1016/j.vetmic.2008.05.012.

Blum, S. E. and G. Leitner (2013). "Genotyping and virulence factors assessment of bovine mastitis *Escherichia coli*." *Vet. Microbiol.* 163 (3-4), pp. 305–312. DOI: 10.1016/j.vetmic.2012.12.037.

Blum, S. E., E. D. Heller, and G. Leitner (2014). "Long term effects of *Escherichia coli* mastitis." *Vet. J.* 201.1, pp. 72–77. DOI: 10.1016/j.tvjl.2014.04.008.

Blum, S. E. et al. (2015). "Genomic and phenomic study of mammary pathogenic *Escherichia coli*." *PLoS ONE* 10.9, e0136387. DOI: 10.1371/journal.pone.0136387.

Blum, S. E. et al. (2017). "Comparison of the immune responses associated with experimental bovine mastitis caused by different strains of *Escherichia coli*." *J. Dairy Res.* 84.2, pp. 190–197. DOI: 10.1017/S0022029917000206.

Blum, S. et al. (2012). "Genome analysis of bovine-mastitis-associated *Escherichia coli* O32:H37 strain P4." *J. Bacteriol.* 194.14, p. 3732. DOI: 10.1128/JB.00535-12.

Boisen, N. et al. (2008). "New adhesin of enteroaggregative *Escherichia coli* related to the Afa/Dr/AAF family." *Infect. Immun.* 76.7, pp. 3281–3292. DOI: 10.1128/IAI.01646-07.

Boll, E. J. et al. (2013). "Role of enteroaggregative *Escherichia coli* virulence factors in uropathogenesis." *Infect. Immun.* 81.4, pp. 1164–1171. DOI: 10.1128/IAI.01376-12.

Bouchard, D. S. et al. (2015). "Lactic acid bacteria isolated from bovine mammary microbiota: potential allies against bovine mastitis." *PLoS ONE* 10.12, e0144831. DOI: 10.1371/journal.pone.0144831.

Bouckaert, J. et al. (2006). "The affinity of the FimH fimbrial adhesin is receptor-driven and quasi-independent of *Escherichia coli* pathotypes." *Mol. Microbiol.* 61.6, pp. 1556–1568. DOI: 10.1111/j.1365-2958.2006.05352.x.

Boulanger, V. et al. (2001). “Induction of nitric oxide production by bovine mammary epithelial cells and blood leukocytes.” *J. Dairy Sci.* 84.6, pp. 1430–1437. DOI: 10.3168/jds.S0022-0302(01)70175-0.

Boyd, E. F. and D. L. Hartl (1998). “Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution.” *J. Bacteriol.* 180.5, pp. 1159–1165.

Brade, L. et al. (1996). “Specificity of monoclonal antibodies against *Escherichia coli* K-12 lipopolysaccharide.” *J. Endotoxin Res.* 3.1, pp. 39–47. DOI: 10.1177/096805199600300105.

Bradley, A. J. and M. J. Green (2001). “Adaptation of *Escherichia coli* to the bovine mammary gland.” *J. Clin. Microbiol.* 39.5, pp. 1845–1849. DOI: 10.1128/JCM.39.5.1845-1849.2001.

Bradley, A. (2002). “Bovine mastitis: an evolving disease.” *Vet. J.* 164.2, pp. 116–128. DOI: 10.1053/tvjl.2002.0724.

Brockmeyer, J. et al. (2009). “Structure and function relationship of the autotransport and proteolytic activity of EspP from Shiga toxin-producing *Escherichia coli*.” *PLoS ONE* 4.7, e6100. DOI: 10.1371/journal.pone.0006100.

Bronowski, C. et al. (2008). “A subset of mucosa-associated *Escherichia coli* isolates from patients with colon cancer, but not Crohn’s disease, share pathogenicity islands with urinary pathogenic *E. coli*.” *Microbiology (Reading, Engl.)* 154 (Pt 2), pp. 571–583. DOI: 10.1099/mic.0.2007/013086-0.

Brown, C. T. (2015). “Strain recovery from metagenomes.” *Nat. Biotechnol.* 33.10, pp. 1041–1043. DOI: 10.1038/nbt.3375.

Browning, D. F. et al. (2013). “Laboratory adapted *Escherichia coli* K-12 becomes a pathogen of *Caenorhabditis elegans* upon restoration of O antigen biosynthesis.” *Mol. Microbiol.* 87.5, pp. 939–950. DOI: 10.1111/mmi.12144.

Brzuszkiewicz, E. et al. (2006). “How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic *Escherichia coli* strains.” *Proc. Natl. Acad. Sci. U.S.A.* 103.34, pp. 12879–12884. DOI: 10.1073/pnas.0603038103.

Brzuszkiewicz, E. et al. (2011). “Genome sequence analyses of two isolates from the recent *Escherichia coli* outbreak in Germany reveal the emergence of a new pathotype: Entero-Aggregative-Haemorrhagic *Escherichia coli* (EAHEC).” *Arch. Microbiol.* 193.12, pp. 883–891. DOI: 10.1007/s00203-011-0725-6.

Buchholz, U. et al. (2011). “German outbreak of *Escherichia coli* O104:H4 associated with sprouts.” *N. Engl. J. Med.* 365.19, pp. 1763–1770. DOI: 10.1056/NEJMoa1106482.

Burgos, Y. and L. Beutin (2010). “Common origin of plasmid encoded alpha-hemolysin genes in *Escherichia coli*.” *BMC Microbiol.* 10, p. 193. DOI: 10.1186/1471-2180-10-193.

Burvenich, C. et al. (2007). “Cumulative physiological events influence the inflammatory response of the bovine udder to *Escherichia coli* infections during the transition period.” *J. Dairy Sci.* 90 Suppl 1, E39–54. DOI: 10.3168/jds.2006-696.

Burvenich, C. et al. (2003). “Severity of *E. coli* mastitis is mainly determined by cow factors.” *Vet. Res.* 34.5, pp. 521–564. DOI: 10.1051/vetres:2003023.

Camacho, C. et al. (2009). “BLAST+: architecture and applications.” *BMC Bioinformatics* 10, p. 421. DOI: 10.1186/1471-2105-10-421.

Carver, T. J. et al. (2005). “ACT: the Artemis Comparison Tool.” *Bioinformatics* 21.16, pp. 3422–3423. DOI: 10.1093/bioinformatics/bti553.

Carver, T. et al. (2009). “DNAPlotter: circular and linear interactive genome visualization.” *Bioinformatics* 25.1, pp. 119–120. DOI: 10.1093/bioinformatics/btn578.

Castillo, A., L. E. Eguiarte, and V. Souza (2005). “A genomic population genetics analysis of the pathogenic enterocyte effacement island in *Escherichia coli*: the search for the unit of selection.” *Proc. Natl. Acad. Sci. U.S.A.* 102.5, pp. 1542–1547. DOI: 10.1073/pnas.0408633102.

Caugant, D. A., B. R. Levin, and R. K. Selander (1981). “Genetic diversity and temporal variation in the *E. coli* population of a human host.” *Genetics* 98.3, pp. 467–490.

Celik, N. et al. (2012). “A bioinformatic strategy for the detection, classification and analysis of bacterial autotransporters.” *PLoS ONE* 7.8, e43245. DOI: 10.1371/journal.pone.0043245.

Chain, P. S. et al. (2009). "Genome project standards in a new era of sequencing." *Science* 326.5950, pp. 236–237. DOI: 10.1126/science.1180614.

Chan, J. Z.-M. et al. (2012a). "Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*." *BMC Microbiol.* 12, p. 302. DOI: 10.1186/1471-2180-12-302.

Chan, J. Z.-M. et al. (2012b). "Genome sequencing in clinical microbiology." *Nat. Biotechnol.* 30.11, pp. 1068–1071. DOI: 10.1038/nbt.2410.

Chandras, C. et al. (2009). "Models for financial sustainability of biological databases and resources." *Database (Oxford)* 2009. DOI: 10.1093/database/bap017.

Chattaway, M. A. et al. (2011). "Enteroaggregative *E. coli* O104 from an outbreak of HUS in Germany 2011, could it happen again?" *J. Infect. Dev. Ctries.* 5.6, pp. 425–436. DOI: 10.3855/jidc.2166.

Chaudhuri, R. R. and I. R. Henderson (2012). "The evolution of the *Escherichia coli* phylogeny." *Infect. Genet. Evol.* 12.2, pp. 214–226. DOI: 10.1016/j.meegid.2012.01.005.

Chaudhuri, R. R. et al. (2010). "Complete genome sequence and comparative metabolic profiling of the prototypical enteroaggregative *Escherichia coli* strain 042." *PLoS ONE* 5.1, e8801. DOI: 10.1371/journal.pone.0008801.

Chen, L. et al. (2005). "VFDB: a reference database for bacterial virulence factors." *Nucleic Acids Res.* 33 (Database issue), pp. D325–328. DOI: 10.1093/nar/gki008.

Chen, L. et al. (2012a). "VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors." *Nucleic Acids Res.* 40 (Database issue), pp. D641–645. DOI: 10.1093/nar/gkr989.

Chen, L. et al. (2016). "VFDB 2016: hierarchical and refined dataset for big data analysis–10 years on." *Nucleic Acids Res.* 44 (D1), pp. D694–697. DOI: 10.1093/nar/gkv1239.

Chen, S. L. et al. (2006). "Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach." *Proc. Natl. Acad. Sci. U.S.A.* 103.15, pp. 5977–5982. DOI: 10.1073/pnas.0600938103.

Chen, W.-H. et al. (2012b). "OGEE: an online gene essentiality database." *Nucleic Acids Res.* 40 (Database issue), pp. D901–906. DOI: 10.1093/nar/gkr986.

Cheng, D. et al. (2012). "Prevalence and isoforms of the pathogenicity island ETT2 among *Escherichia coli* isolates from colibacillosis in pigs and mastitis in cows." *Curr. Microbiol.* 64.1, pp. 43–49. DOI: 10.1007/s00284-011-0032-0.

Chevreux, B. (2005). "MIRA: an automated genome and EST assembler." Ph.D. thesis. Ruprecht-Karls-University, Heidelberg, Germany. URL: http://www.chevreux.org/thesis/.

Chevreux, B. et al. (1999). "Genome sequence assembly using trace signals and additional sequence information." In: *German conference on bioinformatics*. Vol. 99. Heidelberg, pp. 45–56. URL: http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html.

Ciucanu, I. and F. Kerek (1984). "A simple and rapid method for the permethylation of carbohydrates." *Carbohydr. Res.* 131.2, pp. 209–217. DOI: 10.1016/0008-6215(84)85242-8.

Clarke, D. J. et al. (2011). "Complete genome sequence of the Crohn's disease-associated adherent-invasive *Escherichia coli* strain HM605." *J. Bacteriol.* 193.17, p. 4540. DOI: 10.1128/JB.05374-11.

Clermont, O., S. Bonacorsi, and E. Bingen (2000). "Rapid and simple determination of the *Escherichia coli* phylogenetic group." *Appl. Environ. Microbiol.* 66.10, pp. 4555–4558. DOI: 10.1128/AEM.66.10.4555-4558.2000.

Clermont, O., S. Bonacorsi, and E. Bingen (2001). "The *Yersinia* high-pathogenicity island is highly predominant in virulence-associated phylogenetic groups of *Escherichia coli*." *FEMS Microbiol. Lett.* 196.2, pp. 153–157. DOI: 10.1111/j.1574-6968.2001.tb10557.x.

Clermont, O. et al. (2011). "Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds." *Infect. Genet. Evol.* 11.3, pp. 654–662. DOI: 10.1016/j.meegid.2011.02.005.

Cock, P. J. et al. (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." *Bioinformatics* 25.11, pp. 1422–1423. DOI: 10.1093/bioinformatics/btp163.

Cooper, K. K. et al. (2014). "Comparative genomics of enterohemorrhagic *Escherichia coli* O145:H28 demonstrates a common evolutionary lineage with *Escherichia coli* O157:H7." *BMC Genomics* 15, p. 17. DOI: 10.1186/1471-2164-15-17.

Corander, J. et al. (2012). "Population structure in the *Neisseria*, and the biological significance of fuzzy species." *J. R. Soc. Interface* 9.71, pp. 1208–1215. DOI: 10.1098/rsif.2011.0601.

Cotter, S. E. et al. (2005). "Architecture and adhesive activity of the *Haemophilus influenzae* Hsf adhesin." *J. Bacteriol.* 187.13, pp. 4656–4664. DOI: 10.1128/JB.187.13.4656-4664.2005.

Crémet, L. et al. (2013). "Comparison of three methods to study biofilm formation by clinical strains of *Escherichia coli*." *Diagn. Microbiol. Infect. Dis.* 75.3, pp. 252–255. DOI: 10.1016/j.diagmicrobio.2012.11.019.

Crossman, L. C. et al. (2010). "A commensal gone bad: complete genome sequence of the prototypical enterotoxigenic *Escherichia coli* strain H10407." *J. Bacteriol.* 192.21, pp. 5822–5831. DOI: 10.1128/JB.00710-10.

Croxen, M. A. and B. B. Finlay (2010). "Molecular mechanisms of *Escherichia coli* pathogenicity." *Nat. Rev. Microbiol.* 8.1, pp. 26–38. DOI: 10.1038/nrmicro2265.

Croxen, M. A. et al. (2013). "Recent advances in understanding enteric pathogenic *Escherichia coli*." *Clin. Microbiol. Rev.* 26.4, pp. 822–880. DOI: 10.1128/CMR.00022-13.

Dahmen, S. et al. (2013). "Characterization of extended-spectrum beta-lactamase (ESBL)-carrying plasmids and clones of *Enterobacteriaceae* causing cattle mastitis in France." *Vet. Microbiol.* 162 (2-4), pp. 793–799. DOI: 10.1016/j.vetmic.2012.10.015.

Daniels, C., C. Vindurampulle, and R. Morona (1998). "Overexpression and topology of the *Shigella flexneri* O-antigen polymerase (Rfc/Wzy)." *Mol. Microbiol.* 28.6, pp. 1211–1222. DOI: 10.1046/j.1365-2958.1998.00884.x.

Darling, A. E., B. Mau, and N. T. Perna (2010). "progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement." *PLoS ONE* 5.6, e11147. DOI: 10.1371/journal.pone.0011147.

Dautin, N. (2010). "Serine protease autotransporters of *Enterobacteriaceae* (SPATEs): biogenesis and function." *Toxins (Basel)* 2.6, pp. 1179–1206. DOI: 10.3390/toxins2061179.

DebRoy, C. et al. (2016). "Comparison of O-antigen gene clusters of all O-serogroups of *Escherichia coli* and proposal for adopting a new nomenclature for O-typing." *PLOS ONE* 11.1, e0147434. DOI: 10.1371/journal.pone.0147434.

Delcher, A. L. et al. (2007). "Identifying bacterial genes and endosymbiont DNA with Glimmer." *Bioinformatics* 23.6, pp. 673–679. DOI: 10.1093/bioinformatics/btm009.

Denamur, E. (2011). "The 2011 Shiga toxin-producing *Escherichia coli* O104:H4 German outbreak: a lesson in genomic plasticity." *Clin. Microbiol. Infect.* 17.8, pp. 1124–1125. DOI: 10.1111/j.1469-0691.2011.03620.x.

Desvaux, M., N. J. Parham, and I. R. Henderson (2004). "The autotransporter secretion system." *Res. Microbiol.* 155.2, pp. 53–60. DOI: 10.1016/j.resmic.2003.10.002.

Desvaux, M. et al. (2009). "Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue." *Trends Microbiol.* 17.4, pp. 139–145. DOI: 10.1016/j.tim.2009.01.004.

Dhillon, B. K. et al. (2015). "IslandViewer 3: more flexible, interactive genomic island discovery, visualization and analysis." *Nucleic Acids Res.* 43 (W1), W104–108. DOI: 10.1093/nar/gkv401.

Di Padova, F. E. et al. (1993). "A broadly cross-protective monoclonal antibody binding to *Escherichia coli* and *Salmonella* lipopolysaccharides." *Infect. Immun.* 61.9, pp. 3863–3872.

Diard, M. et al. (2007). “*Caenorhabditis elegans* as a simple model to study phenotypic and genetic virulence determinants of extraintestinal pathogenic *Escherichia coli*.” *Microbes Infect.* 9.2, pp. 214–223. DOI: 10.1016/j.micinf.2006.11.009.

Diard, M. et al. (2010). “Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization.” *J. Bacteriol.* 192.19, pp. 4885–4893. DOI: 10.1128/JB.00804-10.

Díaz, E. et al. (2001). “Biodegradation of aromatic compounds by *Escherichia coli*.” *Microbiol. Mol. Biol. Rev.* 65.4, pp. 523–569. DOI: 10.1128/MMBR.65.4.523-569.2001.

Didelot, X. et al. (2012a). “Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*.” *BMC Genomics* 13, p. 256. DOI: 10.1186/1471-2164-13-256.

Didelot, X. et al. (2012b). “Transforming clinical microbiology with bacterial genome sequencing.” *Nat. Rev. Genet.* 13.9, pp. 601–612. DOI: 10.1038/nrg3226.

Dixit, S. M. et al. (2004). “Diversity analysis of commensal porcine *Escherichia coli* - associations between genotypes and habitat in the porcine gastrointestinal tract.” *Microbiology (Reading, Engl.)* 150 (Pt 6), pp. 1735–1740. DOI: 10.1099/mic.0.26733-0.

Djukic, M. et al. (2015). “High quality draft genome of *Lactobacillus kunkeei* EFB6, isolated from a German European foulbrood outbreak of honeybees.” *Stand. Genomic Sci.* 10, p. 16. DOI: 10.1186/1944-3277-10-16.

Dobrindt, U. (2005). “(Patho-)Genomics of *Escherichia coli*.” *Int. J. Med. Microbiol.* 295 (6-7), pp. 357–371. DOI: 10.1016/j.ijmm.2005.07.009.

Dobrindt, U. et al. (2003). “Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays.” *J. Bacteriol.* 185.6, pp. 1831–1840. DOI: 10.1128/JB.185.6.1831-1840.2003.

Dobrindt, U. et al. (2004). “Genomic islands in pathogenic and environmental microorganisms.” *Nat. Rev. Microbiol.* 2.5, pp. 414–424. DOI: 10.1038/nrmicro884.

Dogan, B. et al. (2006). “Adherent and invasive *Escherichia coli* are associated with persistent bovine mastitis.” *Vet. Microbiol.* 116.4, pp. 270–282. DOI: 10.1016/j.vetmic.2006.04.023.

Dogan, B. et al. (2012). “Phylogroup and *lpfA* influence epithelial invasion by mastitis associated *Escherichia coli*.” *Vet. Microbiol.* 159 (1-2), pp. 163–170. DOI: 10.1016/j.vetmic.2012.03.033.

Doolittle, W. F. and O. Zhaxybayeva (2009). “On the origin of prokaryotic species.” *Genome Res.* 19.5, pp. 744–756. DOI: 10.1101/gr.086645.108.

Döpfer, D. et al. (1999). “Recurrent clinical mastitis caused by *Escherichia coli* in dairy cows.” *J. Dairy Sci.* 82.1, pp. 80–85. DOI: 10.3168/jds.S0022-0302(99)75211-2.

Döpfer, D. et al. (2000). “Adhesion and invasion of *Escherichia coli* from single and recurrent clinical cases of bovine mastitis *in vitro*.” *Vet. Microbiol.* 74.4, pp. 331–343. DOI: 10.1016/S0378-1135(00)00191-7.

Doudna, J. A. and E. Charpentier (2014). “Genome editing. The new frontier of genome engineering with CRISPR-Cas9.” *Science* 346.6213, p. 1258096. DOI: 10.1126/science.1258096.

Drobnak, I. et al. (2015). “Of linkers and autochaperones: an unambiguous nomenclature to identify common and uncommon themes for autotransporter secretion.” *Mol. Microbiol.* 95.1, pp. 1–16. DOI: 10.1111/mmi.12838.

D’Souza, J. M., L. Wang, and P. Reeves (2002). “Sequence of the *Escherichia coli* O26 O antigen gene cluster and identification of O26 specific genes.” *Gene* 297 (1-2), pp. 123–127. DOI: 10.1016/S0378-1119(02)00876-4.

Duda, K. A. et al. (2011). “The lipopolysaccharide of the mastitis isolate *Escherichia coli* strain 1303 comprises a novel O-antigen and the rare K-12 core type.” *Microbiology (Reading, Engl.)* 157 (Pt 6), pp. 1750–1760. DOI: 10.1099/mic.0.046912-0.

Dufour, D. et al. (2011). “First evidence of the presence of genomic islands in *Escherichia coli* P4, a mammary pathogen frequently used to induce experimental mastitis.” *J. Dairy Sci.* 94.6, pp. 2779–2793. DOI: 10.3168/jds.2010-3446.

Easton, D. M. et al. (2011). “Characterization of EhaJ, a new autotransporter protein from enterohemorrhagic and enteropathogenic *Escherichia coli*.” *Front. Microbiol.* 2, p. 120. DOI: 10.3389/fmicb.2011.00120.

Eidam, C. et al. (2015). “Analysis and comparative genomics of ICE*Mh1*, a novel integrative and conjugative element (ICE) of *Mannheimia haemolytica*.” *J. Antimicrob. Chemother.* 70.1, pp. 93–97. DOI: 10.1093/jac/dku361.

Erlich, Y. (2015). “A vision for ubiquitous sequencing.” *Genome Res.* 25.10, pp. 1411–1416. DOI: 10.1101/gr.191692.115.

Escobar-Páramo, P. et al. (2004a). “A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*.” *Mol. Biol. Evol.* 21.6, pp. 1085–1094. DOI: 10.1093/molbev/msh118.

Escobar-Páramo, P. et al. (2004b). “Large-scale population structure of human commensal *Escherichia coli* isolates.” *Appl. Environ. Microbiol.* 70.9, pp. 5698–5700. DOI: 10.1128/AEM.70.9.5698-5700.2004.

Escobar-Páramo, P. et al. (2006). “Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates.” *Environ. Microbiol.* 8.11, pp. 1975–1984. DOI: 10.1111/j.1462-2920.2006.01077.x.

Etchuuya, R. et al. (2011). “Cell-to-cell transformation in *Escherichia coli*: a novel type of natural transformation involving cell-derived DNA and a putative promoting pheromone.” *PLoS ONE* 6.1, e16355. DOI: 10.1371/journal.pone.0016355.

Fabich, A. J. et al. (2008). “Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine.” *Infect. Immun.* 76.3, pp. 1143–1152. DOI: 10.1128/IAI.01386-07.

Fairbrother, J.-H. et al. (2015). “Characterization of persistent and transient *Escherichia coli* isolates recovered from clinical mastitis episodes in dairy cows.” *Vet. Microbiol.* 176 (1-2), pp. 126–133. DOI: 10.1016/j.vetmic.2014.12.025.

Falentin, H. et al. (2016). “Bovine teat microbiome analysis revealed reduced alpha diversity and significant changes in taxonomic profiles in quarters with a history of mastitis.” *Front. Microbiol.* 7, p. 480. DOI: 10.3389/fmicb.2016.00480.

Feldman, M. F. et al. (1999). “The activity of a putative polyisoprenol-linked sugar translocase (Wzx) involved in *Escherichia coli* O antigen assembly is independent of the chemical structure of the O repeat.” *J. Biol. Chem.* 274.49, pp. 35129–35138. DOI: 10.1074/jbc.274.49.35129.

Feldmann, F. et al. (2007). “The salmochelin siderophore receptor IroN contributes to invasion of urothelial cells by extraintestinal pathogenic *Escherichia coli in vitro*.” *Infect. Immun.* 75.6, pp. 3183–3187. DOI: 10.1128/IAI.00656-06.

Feng, L. et al. (2008). “A genomic islet mediates flagellar phase variation in *Escherichia coli* strains carrying the flagellin-specifying locus *flk*.” *J. Bacteriol.* 190.13, pp. 4470–4477. DOI: 10.1128/JB.01937-07.

Fernandes, J. B. C. et al. (2011). “*Escherichia coli* from clinical mastitis: serotypes and virulence factors.” *J. Vet. Diagn. Invest.* 23.6, pp. 1146–1152. DOI: 10.1177/1040638711425581.

Fleischmann, R. D. et al. (1995). “Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.” *Science* 269.5223, pp. 496–512. DOI: 10.1126/science.7542800.

Francisco, A. P. et al. (2009). “Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach.” *BMC Bioinformatics* 10, p. 152. DOI: 10.1186/1471-2105-10-152.

Francisco, A. P. et al. (2012). “PHYLOViZ: phylogenetic inference and data visualization for sequence based typing methods.” *BMC Bioinformatics* 13, p. 87. DOI: 10.1186/1471-2105-13-87.

Frank, C. et al. (2011). “Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany.” *N. Engl. J. Med.* 365.19, pp. 1771–1780. DOI: 10.1056/NEJMoa1106483.

Fratamico, P. M., C. DebRoy, and D. S. Needleman (2016). “Editorial: Emerging approaches for typing, detection, characterization, and traceback of *Escherichia coli*.” *Front. Microbiol.* 7, p. 2089. DOI: 10.3389/fmicb.2016.02089.

Freitag, C. et al. (2016). “Detection of plasmid-borne extended-spectrum β-lactamase (ESBL) genes in *Escherichia coli* isolates from bovine mastitis.” *Vet. Microbiol.* DOI: 10.1016/j.vetmic.2016.08.010.

Fricke, W. F. and D. A. Rasko (2014). “Bacterial genome sequencing in the clinic: bioinformatic challenges and solutions.” *Nat. Rev. Genet.* 15.1, pp. 49–55. DOI: 10.1038/nrg3624.

Fricke, W. F. et al. (2008). “Insights into the environmental resistance gene pool from the genome sequence of the multidrug-resistant environmental isolate *Escherichia coli* SMS-3-5.” *J. Bacteriol.* 190.20, pp. 6779–6794. DOI: 10.1128/JB.00661-08.

Fricke, W. F. et al. (2009). “Antimicrobial resistance-conferring plasmids with similarity to virulence plasmids from avian pathogenic *Escherichia coli* strains in *Salmonella enterica* serovar Kentucky isolates from poultry.” *Appl. Environ. Microbiol.* 75.18, pp. 5963–5971. DOI: 10.1128/AEM.00786-09.

Frirdich, E. et al. (2003). “Overexpression of the *waaZ* gene leads to modification of the structure of the inner core region of *Escherichia coli* lipopolysaccharide, truncation of the outer core, and reduction of the amount of O polysaccharide on the cell surface.” *J. Bacteriol.* 185.5, pp. 1659–1671. DOI: 10.1128/JB.185.5.1659-1671.2003.

Ganda, E. K. et al. (2016). “Longitudinal metagenomic profiling of bovine milk to assess the impact of intramammary treatment using a third-generation cephalosporin.” *Sci. Rep.* 6, p. 37565. DOI: 10.1038/srep37565.

Garcia, E. C., A. R. Brumbaugh, and H. L. T. Mobley (2011). “Redundancy and specificity of *Escherichia coli* iron acquisition systems during urinary tract infection.” *Infect. Immun.* 79.3, pp. 1225–1235. DOI: 10.1128/IAI.01222-10.

Gardy, J., N. J. Loman, and A. Rambaut (2015). “Real-time digital pathogen surveillance - the time is now.” *Genome Biol.* 16.1, p. 155. DOI: 10.1186/s13059-015-0726-x.

Gawarzewski, I. et al. (2013). “Structural comparison of the transport units of type V secretion systems.” *Biol. Chem.* 394.11, pp. 1385–1398. DOI: 10.1515/hsz-2013-0162.

Gerwig, G. J., J. P. Kamerling, and J. F. Vliegenthart (1979). “Determination of the absolute configuration of mono-saccharides in complex carbohydrates by capillary G.L.C.” *Carbohydr. Res.* 77, pp. 10–17.

Ghanbarpour, R. and E. Oswald (2010). “Phylogenetic distribution of virulence genes in *Escherichia coli* isolated from bovine mastitis in Iran.” *Res. Vet. Sci.* 88.1, pp. 6–10. DOI: 10.1016/j.rvsc.2009.06.003.

Ghatak, S. et al. (2013). “Detection of New Delhi metallo-beta-lactamase and extended-spectrum beta-lactamase genes in *Escherichia coli* isolated from mastitic milk samples.” *Transbound Emerg. Dis.* 60.5, pp. 385–389. DOI: 10.1111/tbed.12119.

Gibbs, R. J., J. Stewart, and I. R. Poxton (2004). “The distribution of, and antibody response to, the core lipopolysaccharide region of *Escherichia coli* isolated from the faeces of healthy humans and cattle.” *J. Med. Microbiol.* 53 (Pt 10), pp. 959–964. DOI: 10.1099/jmm.0.45674-0.

Ginsburg, V. (1961). “Studies on the biosynthesis of guanosine diphosphate L-fucose.” *J. Biol. Chem.* 236, pp. 2389–2393.

Goldenfeld, N. and C. Woese (2007). “Biology’s next revolution.” *Nature* 445.7126, p. 369. DOI: 10.1038/445369a.

Goldstone, R. J., S. Harris, and D. G. E. Smith (2016). “Genomic content typifying a prevalent clade of bovine mastitis-associated *Escherichia coli*.” *Sci. Rep.* 6, p. 30115. DOI: 10.1038/srep30115.

Gomes, F. and M. Henriques (2016). “Control of bovine mastitis: old and recent therapeutic approaches.” *Curr. Microbiol.* 72.4, pp. 377–382. DOI: 10.1007/s00284-015-0958-8.

Gomes, F., M. J. Saavedra, and M. Henriques (2016). “Bovine mastitis disease/pathogenicity: evidence of the potential role of microbial biofilms.” *Pathog. Dis.* 74.3, ftw006. DOI: 10.1093/femspd/ftw006.

González, R. N. et al. (1989). “Prevention of clinical coliform mastitis in dairy cows by a mutant *Escherichia coli* vaccine.” *Can. J. Vet. Res.* 53.3, pp. 301–305.

Goodwin, S., J. D. McPherson, and W. R. McCombie (2016). "Coming of age: ten years of next-generation sequencing technologies." *Nat. Rev. Genet.* 17.6, pp. 333–351. DOI: 10.1038/nrg.2016.49.

Gordon, D. M. and A. Cowling (2003). "The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects." *Microbiology (Reading, Engl.)* 149 (Pt 12), pp. 3575–3586. DOI: 10.1099/mic.0.26486-0.

Grad, Y. H. et al. (2012). "Genomic epidemiology of the *Escherichia coli* O104:H4 outbreaks in Europe, 2011." *Proc. Natl. Acad. Sci. U.S.A.* 109.8, pp. 3065–3070. DOI: 10.1073/pnas.1121491109.

Grad, Y. H. et al. (2013). "Comparative genomics of recent Shiga toxin-producing *Escherichia coli* O104:H4: short-term evolution of an emerging pathogen." *MBio* 4.1, e00452–00412. DOI: 10.1128/mBio.00452-12.

Graninger, M. et al. (2002). "Homologs of the Rml enzymes from *Salmonella enterica* are responsible for dTDP-beta-L-rhamnose biosynthesis in the gram-positive thermophile *Aneurinibacillus thermoaerophilus* DSM 10155." *Appl. Environ. Microbiol.* 68.8, pp. 3708–3715. DOI: 10.1128/AEM.68.8.3708-3715.2002.

Grasselli, E. et al. (2008). "Evidence of horizontal gene transfer between human and animal commensal *Escherichia coli* strains identified by microarray." *FEMS Immunol. Med. Microbiol.* 53.3, pp. 351–358. DOI: 10.1111/j.1574-695X.2008.00434.x.

Gray, C. H. and E. L. Tatum (1944). "X-Ray induced growth factor requirements in bacteria." *Proc. Natl. Acad. Sci. U.S.A.* 30.12, pp. 404–410.

Greinacher, A. et al. (2011). "Treatment of severe neurological deficits with IgG depletion through immunoadsorption in patients with *Escherichia coli* O104:H4-associated haemolytic uraemic syndrome: a prospective trial." *Lancet* 378.9797, pp. 1166–1173. DOI: 10.1016/S0140-6736(11)61253-1.

Grijpstra, J. et al. (2013). "Autotransporter secretion: varying on a theme." *Res. Microbiol.* 164.6, pp. 562–582. DOI: 10.1016/j.resmic.2013.03.010.

Grozdanov, L. et al. (2002). "A single nucleotide exchange in the *wzy* gene is responsible for the semirough O6 lipopolysaccharide phenotype and serum sensitivity of *Escherichia coli* strain Nissle 1917." *J. Bacteriol.* 184.21, pp. 5912–5925. DOI: 10.1128/JB.184.21.5912-5925.2002.

Grozdanov, L. et al. (2004). "Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917." *J. Bacteriol.* 186.16, pp. 5432–5441. DOI: 10.1128/JB.186.16.5432-5441.2004.

Günther, J. et al. (2016). "Comparison of the pathogen species-specific immune response in udder derived cell types and their models." *Vet. Res.* 47.1, p. 22. DOI: 10.1186/s13567-016-0307-3.

Gunther, N. W. et al. (2002). "Assessment of virulence of uropathogenic *Escherichia coli* type 1 fimbrial mutants in which the invertible element is phase-locked on or off." *Infect. Immun.* 70.7, pp. 3344–3354. DOI: 10.1128/IAI.70.7.3344-3354.2002.

Gurevich, A. et al. (2013). "QUAST: quality assessment tool for genome assemblies." *Bioinformatics* 29.8, pp. 1072–1075. DOI: 10.1093/bioinformatics/btt086.

Güttsches, A.-K. et al. (2012). "Anti-inflammatory modulation of immune response by probiotic *Escherichia coli* Nissle 1917 in human blood mononuclear cells." *Innate Immun.* 18.2, pp. 204–216. DOI: 10.1177/1753425910396251.

Guy, L. et al. (2013). "Adaptive mutations and replacements of virulence traits in the *Escherichia coli* O104:H4 outbreak population." *PLoS ONE* 8.5, e63027. DOI: 10.1371/journal.pone.0063027.

Hacker, J. and U. Dobrindt (2006). *Pathogenomics: Genome analysis of pathogenic microbes*. Wiley-VCH Verlag GmbH & Co. KGaA. 568 pp. URL: http://dx.doi.org/10.1002/352760801X.

Hacker, J., U. Hentschel, and U. Dobrindt (2003). "Prokaryotic chromosomes and disease." *Science* 301.5634, pp. 790–793. DOI: 10.1126/science.1086802.

Hafez, M. et al. (2009). "The K5 capsule of *Escherichia coli* strain Nissle 1917 is important in mediating interactions with intestinal epithelial cells and chemokine induction." *Infect. Immun.* 77.7, pp. 2995–3003. DOI: 10.1128/IAI.00040-09.

Haishima, Y., O. Holst, and H. Brade (1992). "Structural investigation on the lipopolysaccharide of *Escherichia coli* rough mutant F653 representing the R3 core type." *Eur. J. Biochem.* 203 (1-2), pp. 127–134. DOI: 10.1111/j.1432-1033.1992.tb19837.x.

Halachev, M. R., N. J. Loman, and M. J. Pallen (2011). "Calculating orthologs in bacteria and archaea: a divide and conquer approach." *PLoS ONE* 6.12, e28388. DOI: 10.1371/journal.pone.0028388.

Hammer, Ø., D. Harper, and P. Ryan (2001). "PAST: paleontological statistics software package for education and data analysis." *Palaeontol. Electron.* 4.1, 9pp. URL: http://palaeo-electronica.org/2001_1/past/issue1_01.htm.

Hanage, W. P. (2016). "Not so simple after all: bacteria, their population genetics, and recombination." *Cold Spring Harb. Perspect. Biol.* 8.7. DOI: 10.1101/cshperspect.a018069.

Hancock, V., L. Ferrières, and P. Klemm (2008). "The ferric yersiniabactin uptake receptor FyuA is required for efficient biofilm formation by urinary tract infectious *Escherichia coli* in human urine." *Microbiology (Reading, Engl.)* 154 (Pt 1), pp. 167–175. DOI: 10.1099/mic.0.2007/011981-0.

Hancock, V., R. M. Vejborg, and P. Klemm (2010a). "Functional genomics of probiotic *Escherichia coli* Nissle 1917 and 83972, and UPEC strain CFT073: comparison of transcriptomes, growth and biofilm formation." *Mol. Genet. Genomics* 284.6, pp. 437–454. DOI: 10.1007/s00438-010-0578-8.

Hancock, V., M. Dahl, and P. Klemm (2010b). "Probiotic *Escherichia coli* strain Nissle 1917 outcompetes intestinal pathogens during biofilm formation." *J. Med. Microbiol.* 59 (Pt 4), pp. 392–399. DOI: 10.1099/jmm.0.008672-0.

Harrington, S. M., E. G. Dudley, and J. P. Nataro (2006). "Pathogenesis of enteroaggregative *Escherichia coli* infection." *FEMS Microbiol. Lett.* 254.1, pp. 12–18. DOI: 10.1111/j.1574-6968.2005.00005.x.

Hashimoto, M. et al. (2005). "Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome." *Mol. Microbiol.* 55.1, pp. 137–149. DOI: 10.1111/j.1365-2958.2004.04386.x.

Hasman, H., T. Chakraborty, and P. Klemm (1999). "Antigen-43-mediated autoaggregation of *Escherichia coli* is blocked by fimbriation." *J. Bacteriol.* 181.16, pp. 4834–4841.

Hasman, H. et al. (2014). "Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples." *J. Clin. Microbiol.* 52.1, pp. 139–146. DOI: 10.1128/JCM.02452-13.

Hayashi, T. et al. (2001). "Complete genome sequence of enterohemorrhagic *Escherichia coli* O157:H7 and genomic comparison with a laboratory strain K-12." *DNA Res.* 8.1, pp. 11–22. DOI: 10.1093/dnares/8.1.11.

Hazen, T. H. et al. (2013). "Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*." *Proc. Natl. Acad. Sci. U.S.A.* 110.31, pp. 12810–12815. DOI: 10.1073/pnas.1306836110.

Hazen, T. H. et al. (2016). "Genomic diversity of EPEC associated with clinical presentations of differing severity." *Nat. Microbiol.* 1, p. 15014. DOI: 10.1038/nmicrobiol.2015.14.

Heinrichs, D. E., J. A. Yethon, and C. Whitfield (1998). "Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*." *Mol. Microbiol.* 30.2, pp. 221–232. DOI: 10.1046/j.1365-2958.1998.01063.x.

Hejnova, J. et al. (2005). "Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83 : K24 : H31)." *Microbiology (Reading, Engl.)* 151 (Pt 2), pp. 385–398. DOI: 10.1099/mic.0.27469-0.

Helmy, M., A. Crits-Christoph, and G. D. Bader (2016). "Ten simple rules for developing public biological databases." *PLOS Comput. Biol.* 12.11, e1005128. DOI: 10.1371/journal.pcbi.1005128.

Henderson, I. R. and J. P. Nataro (2001). "Virulence functions of autotransporter proteins." *Infect. Immun.* 69.3, pp. 1231–1243. DOI: 10.1128/IAI.69.3.1231-1243.2001.

Henderson, I. R., M. Meehan, and P. Owen (1997). “Antigen 43, a phase-variable bipartite outer membrane protein, determines colony morphology and autoaggregation in *Escherichia coli* K-12.” *FEMS Microbiol. Lett.* 149.1, pp. 115–120.

Henderson, I. R. et al. (2004). “Type V protein secretion pathway: the autotransporter story.” *Microbiol. Mol. Biol. Rev.* 68.4, pp. 692–744. DOI: 10.1128/MMBR.68.4.692-744.2004.

Hendrickson, H. (2009). “Order and disorder during *Escherichia coli* divergence.” *PLoS Genet.* 5.1, e1000335. DOI: 10.1371/journal.pgen.1000335.

Herry, V. et al. (2017). “Local immunization impacts the response of dairy cows to *Escherichia coli* mastitis.” *Sci. Rep.* 7.1, p. 3441. DOI: 10.1038/s41598-017-03724-7.

Hertel, R. et al. (2015). “Genome-based identification of active prophage regions by next generation sequencing in *Bacillus licheniformis* DSM13.” *PLoS ONE* 10.3, e0120759. DOI: 10.1371/journal.pone.0120759.

Hill, C. W. et al. (1995). “Correlation of Rhs elements with *Escherichia coli* population structure.” *Genetics* 141.1, pp. 15–24.

Hillerton, J. E. and E. A. Berry (2005). “Treating mastitis in the cow–a tradition or an archaism.” *J. Appl. Microbiol.* 98.6, pp. 1250–1255. DOI: 10.1111/j.1365-2672.2005.02649.x.

Ho Sui, S. J. et al. (2009). “The association of virulence factors with genomic islands.” *PLoS ONE* 4.12, e8094. DOI: 10.1371/journal.pone.0008094.

Hogan, J. and K. Larry Smith (2003). “Coliform mastitis.” *Vet. Res.* 34.5, pp. 507–519. DOI: 10.1051/vetres:2003022.

Hogeveen, H., K. Huijps, and T. J. Lam (2011). “Economic aspects of mastitis: new developments.” *N. Z. Vet. J.* 59.1, pp. 16–23. DOI: 10.1080/00480169.2011.547165.

Holden, N., L. Pritchard, and I. Toth (2009). “Colonization outwith the colon: plants as an alternative environmental reservoir for human pathogenic enterobacteria.” *FEMS Microbiol. Rev.* 33.4, pp. 689–703. DOI: 10.1111/j.1574-6976.2008.00153.x.

Holland, I. B. (2010). “The extraordinary diversity of bacterial protein secretion mechanisms.” *Methods Mol. Biol.* 619, pp. 1–20. DOI: 10.1007/978-1-60327-412-8_1.

Holst, O. et al. (1991). “Structural analysis of the heptose/hexose region of the lipopolysaccharide from *Escherichia coli* K-12 strain W3100.” *Carbohydr. Res.* 215.2, pp. 323–335. DOI: 10.1016/0008-6215(91)84031-9.

Holst, O. and H Brade (1999). “Chemical structure of the core region of lipopolysaccharides.” In: *Endotoxin in Health and Disease*. Vol. 1. Marcel Dekker: New York, NY, USA, pp. 115–154.

Holst, O., A. P. Moran, and P. J. Brennan (2009). “Overview of the glycosylated components of the bacterial cell envelope.” In: *Microbial glycobiology: structures, relevance and applications*. Amsterdam: Academic Press, pp. 3–13.

Houser, B. A. et al. (2008). “Assessment of phenotypic and genotypic diversity of *Escherichia coli* shed by healthy lactating dairy cattle.” *Foodborne Pathog. Dis.* 5.1, pp. 41–51. DOI: 10.1089/fpd.2007.0036.

Huang, A. D. et al. (2017). “Metagenomics of two severe foodborne outbreaks provides diagnostic signatures and signs of coinfection not attainable by traditional methods.” *Appl. Environ. Microbiol.* 83.3. DOI: 10.1128/AEM.02577-16.

Huang, D. B. et al. (2006). “A review of an emerging enteric pathogen: enteroaggregative *Escherichia coli*.” *J. Med. Microbiol.* 55 (Pt 10), pp. 1303–1311. DOI: 10.1099/jmm.0.46674-0.

Huebner, C. et al. (2011). “The probiotic *Escherichia coli* Nissle 1917 reduces pathogen invasion and modulates cytokine expression in Caco-2 cells infected with Crohn’s disease-associated *E. coli* LF82.” *Appl. Environ. Microbiol.* 77.7, pp. 2541–2544. DOI: 10.1128/AEM.01601-10.

Hug, L. A. et al. (2016). “A new view of the tree of life.” *Nat. Microbiol.* 1.5, p. 16048. DOI: 10.1038/nmicrobiol.2016.48.

Huja, S. et al. (2015). "Genomic avenue to avian colisepticemia." *MBio* 6.1. DOI: 10.1128/mBio.01681-14.

Hung, C.-S. et al. (2002). "Structural basis of tropism of *Escherichia coli* to the bladder during urinary tract infection." *Mol. Microbiol.* 44.4, pp. 903–915.

Huson, D. H. and C. Scornavacca (2012). "Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks." *Syst. Biol.* 61.6, pp. 1061–1067. DOI: 10.1093/sysbio/sys062.

Hwang, B.-Y. et al. (2007). "Substrate specificity of the *Escherichia coli* outer membrane protease OmpP." *J. Bacteriol.* 189.2, pp. 522–530. DOI: 10.1128/JB.01493-06.

Hyatt, D. et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." *BMC Bioinformatics* 11, p. 119. DOI: 10.1186/1471-2105-11-119.

Hyvönen, P. et al. (2006). "Transgenic cows that produce recombinant human lactoferrin in milk are not protected from experimental *Escherichia coli* intramammary infection." *Infect. Immun.* 74.11, pp. 6206–6212. DOI: 10.1128/IAI.00238-06.

Ibrahim, D. R. et al. (2016). "Multidrug resistant, extended spectrum β-lactamase (ESBL)-producing *Escherichia coli* isolated from a dairy farm." *FEMS Microbiol. Ecol.* 92.4, fiw013. DOI: 10.1093/femsec/fiw013.

Ideses, D. et al. (2005). "A degenerate type III secretion system from septicemic *Escherichia coli* contributes to pathogenesis." *J. Bacteriol.* 187.23, pp. 8164–8171. DOI: 10.1128/JB.187.23.8164-8171.2005.

Iguchi, A. et al. (2009). "Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69." *J. Bacteriol.* 191.1, pp. 347–354. DOI: 10.1128/JB.01238-08.

Ingle, D. J. et al. (2016a). "*In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages." *Microb. Genom.* 2.7. DOI: 10.1099/mgen.0.000064.

Ingle, D. J. et al. (2016b). "Evolution of atypical enteropathogenic *E. coli* by repeated acquisition of LEE pathogenicity island variants." *Nat. Microbiol.* 1.2, p. 15010. DOI: 10.1038/nmicrobiol.2015.10.

Isobe, N. et al. (2009). "Existence of functional lingual antimicrobial peptide in bovine milk." *J. Dairy Sci.* 92.6, pp. 2691–2695. DOI: 10.3168/jds.2008-1940.

Jackson, R. W. et al. (2011). "The influence of the accessory genome on bacterial pathogen evolution." *Mob. Genet. Elements.* 1.1, pp. 55–65. DOI: 10.4161/mge.1.1.16432.

Jain, E. et al. (2009). "Infrastructure for the life sciences: design and implementation of the UniProt website." *BMC Bioinformatics* 10, p. 136. DOI: 10.1186/1471-2105-10-136.

Jain, M. et al. (2016). "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community." *Genome Biol.* 17.1, p. 239. DOI: 10.1186/s13059-016-1103-0.

Jain, M. et al. (2017). "MinION Analysis and Reference Consortium: phase 2 data release and analysis of R9.0 chemistry." *F1000Research* 6, p. 760. DOI: 10.12688/f1000research.11354.1.

Jann, B. et al. (1994). "Structure of the O16 polysaccharide from *Escherichia coli* O16:K1: an NMR investigation." *Carbohydr. Res.* 264.2, pp. 305–311.

Jansson, P. E. et al. (1981). "Structural studies on the hexose region of the core in lipopolysaccharides from *Enterobacteriaceae*." *Eur. J. Biochem.* 115.3, pp. 571–577. DOI: 10.1111/j.1432-1033.1981.tb06241.x.

Joensen, K. G. et al. (2015). "Rapid and easy *in silico* serotyping of *Escherichia coli* using whole genome sequencing (WGS) data." *J. Clin. Microbiol.* 53.8, pp. 2410–26. DOI: 10.1128/JCM.00008-15.

Joensen, K. G. et al. (2014). "Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic *Escherichia coli*." *J. Clin. Microbiol.* 52.5, pp. 1501–1510. DOI: 10.1128/JCM.03617-13.

Johnson, J. R. and A. L. Stell (2000). "Extended virulence genotypes of *Escherichia coli* strains from patients with urosepsis in relation to phylogeny and host compromise." *J. Infect. Dis.* 181.1, pp. 261–272. DOI: 10.1086/315217.

Johnson, J. R. et al. (2001). "Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*." *J. Infect. Dis.* 183.1, pp. 78–88. DOI: 10.1086/317656.

Johnson, J. R. et al. (2008a). "Molecular epidemiology and phylogenetic distribution of the *Escherichia coli pks* genomic island." *J. Clin. Microbiol.* 46.12, pp. 3906–3911. DOI: 10.1128/JCM.00949-08.

Johnson, T. J., S. J. Johnson, and L. K. Nolan (2006a). "Complete DNA sequence of a ColBM plasmid from avian pathogenic *Escherichia coli* suggests that it evolved from closely related ColV virulence plasmids." *J. Bacteriol.* 188.16, pp. 5975–5983. DOI: 10.1128/JB.00204-06.

Johnson, T. J. et al. (2006b). "DNA sequence of a ColV plasmid and prevalence of selected plasmid-encoded virulence genes among avian *Escherichia coli* strains." *J. Bacteriol.* 188.2, pp. 745–758. DOI: 10.1128/JB.188.2.745-758.2006.

Johnson, T. J. et al. (2007). "The genome sequence of avian pathogenic *Escherichia coli* strain O1:K1:H7 shares strong similarities with human extraintestinal pathogenic *E. coli* genomes." *J. Bacteriol.* 189.8, pp. 3228–3236. DOI: 10.1128/JB.01726-06.

Johnson, T. J. et al. (2008b). "Comparison of extraintestinal pathogenic *Escherichia coli* strains from human and avian sources reveals a mixed subset representing potential zoonotic pathogens." *Appl. Environ. Microbiol.* 74.22, pp. 7043–7050. DOI: 10.1128/AEM.01395-08.

Johnson, T. J., Y. M. Wannemuehler, and L. K. Nolan (2008c). "Evolution of the *iss* gene in *Escherichia coli*." *Appl. Environ. Microbiol.* 74.8, pp. 2360–2369. DOI: 10.1128/AEM.02634-07.

Jores, J., L. Rumer, and L. H. Wieler (2004). "Impact of the locus of enterocyte effacement pathogenicity island on the evolution of pathogenic *Escherichia coli*." *Int. J. Med. Microbiol.* 294 (2-3), pp. 103–113. DOI: 10.1016/j.ijmm.2004.06.024.

Jorgensen, R. A. (2011). "We're all computational biologists now. . . Next stop, the global brain?" *Front. Genet.* 2, p. 68. DOI: 10.3389/fgene.2011.00068.

Journet, L. and E. Cascales (2016). "The type VI secretion system in *Escherichia coli* and related species." *EcoSal Plus* 7.1, ESP–0009–2015. DOI: 10.1128/ecosalplus.ESP-0009-2015.

Juhas, M. et al. (2009). "Genomic islands: tools of bacterial horizontal gene transfer and evolution." *FEMS Microbiol. Rev.* 33.2, pp. 376–393. DOI: 10.1111/j.1574-6976.2008.00136.x.

Kaas, R. S. et al. (2012). "Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse *Escherichia coli* genomes." *BMC Genomics* 13, p. 577. DOI: 10.1186/1471-2164-13-577.

Kaipainen, T. et al. (2002). "Virulence factors of *Escherichia coli* isolated from bovine clinical mastitis." *Vet. Microbiol.* 85.1, pp. 37–46. DOI: 10.1016/S0378-1135(01)00483-7.

Kaper, J. B., J. P. Nataro, and H. L. Mobley (2004). "Pathogenic *Escherichia coli*." *Nat. Rev. Microbiol.* 2.2, pp. 123–140. DOI: 10.1038/nrmicro818.

Karch, H., P. I. Tarr, and M. Bielaszewska (2005). "Enterohaemorrhagic *Escherichia coli* in human medicine." *Int. J. Med. Microbiol.* 295 (6-7), pp. 405–418. DOI: 10.1016/j.ijmm.2005.06.009.

Karch, H. et al. (2012). "The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak." *EMBO Mol. Med.* 4.9, pp. 841–848. DOI: 10.1002/emmm.201201662.

Karmali, M. A. et al. (1983). "Sporadic cases of haemolytic-uraemic syndrome associated with faecal cytotoxin and cytotoxin-producing *Escherichia coli* in stools." *Lancet* 1.8325, pp. 619–620. DOI: 10.1016/S0140-6736(83)91795-6.

Karmali, M. A. et al. (1985). "The association between idiopathic hemolytic uremic syndrome and infection by verotoxin-producing *Escherichia coli*." *J. Infect. Dis.* 151.5, pp. 775–782. DOI: 10.1086/jid/189.3.566.

Kato, J.-i. and M. Hashimoto (2007). “Construction of consecutive deletions of the *Escherichia coli* chromosome.” *Mol. Syst. Biol.* 3, p. 132. DOI: 10.1038/msb4100174.

Kempf, F., V. Loux, and P. Germon (2015). “Genome sequences of two bovine mastitis-causing *Escherichia coli* strains.” *Genome Announc.* 3.2, e00259–15. DOI: 10.1128/genomeA.00259-15.

Kempf, F. et al. (2016). “Genomic comparative study of bovine mastitis *Escherichia coli*.” *PLoS ONE* 11.1, e0147954. DOI: 10.1371/journal.pone.0147954.

Keseler, I. M. et al. (2013). “EcoCyc: fusing model organism databases with systems biology.” *Nucleic Acids Res.* 41 (Database issue), pp. D605–612. DOI: 10.1093/nar/gks1027.

Khachatryan, A. R., T. E. Besser, and D. R. Call (2008). “The streptomycin-sulfadiazine-tetracycline antimicrobial resistance element of calf-adapted *Escherichia coli* is widely distributed among isolates from Washington state cattle.” *Appl. Environ. Microbiol.* 74.2, pp. 391–395. DOI: 10.1128/AEM.01534-07.

Kim, M. et al. (2012). “Draft genome sequence of *Escherichia coli* W26, an enteric strain isolated from cow feces.” *J. Bacteriol.* 194.18, pp. 5149–5150. DOI: 10.1128/JB.01180-12.

Klemm, E. and G. Dougan (2016). “Advances in understanding bacterial pathogenesis gained from whole-genome sequencing and phylogenetics.” *Cell Host & Microbe* 19.5, pp. 599–610. DOI: 10.1016/j.chom.2016.04.015.

Klemm, P. et al. (2006). “Molecular characterization of the *Escherichia coli* asymptomatic bacteriuria strain 83972: the taming of a pathogen.” *Infect. Immun.* 74.1, pp. 781–785. DOI: 10.1128/IAI.74.1.781-785.2006.

Köhler, C.-D. and U. Dobrindt (2011). “What defines extraintestinal pathogenic *Escherichia coli*?” *Int. J. Med. Microbiol.* 301.8, pp. 642–647. DOI: 10.1016/j.ijmm.2011.09.006.

Koli, P. et al. (2011). “Conversion of commensal *Escherichia coli* K-12 to an invasive form via expression of a mutant histone-like protein.” *MBio* 2.5. DOI: 10.1128/mBio.00182-11.

Kondo, I., T. Ishikawa, and H. Nakahara (1974). “Mercury and cadmium resistances mediated by the penicillinase plasmid in *Staphylococcus aureus*.” *J. Bacteriol.* 117.1, pp. 1–7.

Korea, C.-G. et al. (2010). “*Escherichia coli* K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities.” *Environ. Microbiol.* 12.7, pp. 1957–1977. DOI: 10.1111/j.1462-2920.2010.02202.x.

Koren, S. and A. M. Phillippy (2015). “One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly.” *Curr. Opin. Microbiol.* 23C, pp. 110–120. DOI: 10.1016/j.mib.2014.11.014.

Koren, S. et al. (2013). “Reducing assembly complexity of microbial genomes with single-molecule sequencing.” *Genome Biol.* 14.9, R101. DOI: 10.1186/gb-2013-14-9-r101.

Kornalijnslijper, J. E. et al. (2004). “Bacterial growth during the early phase of infection determines the severity of experimental *Escherichia coli* mastitis in dairy cows.” *Vet. Microbiol.* 101.3, pp. 177–186. DOI: 10.1016/j.vetmic.2004.04.005.

Kotloff, K. L. et al. (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, pp. 209–222. DOI: 10.1016/S0140-6736(13)60844-2.

Krause, D. O. et al. (2011). “Complete genome sequence of adherent invasive *Escherichia coli* UM146 isolated from Ileal Crohn’s disease biopsy tissue.” *J. Bacteriol.* 193.2, p. 583. DOI: 10.1128/JB.01290-10.

Krieger, J. N. et al. (2011). “Acute *Escherichia coli* prostatitis in previously health young men: bacterial virulence factors, antimicrobial resistance, and clinical outcomes.” *Urology* 77.6, pp. 1420–1425. DOI: 10.1016/j.urology.2010.12.059.

Kruis, W. et al. (2012). “A double-blind placebo-controlled trial to study therapeutic effects of probiotic *Escherichia coli* Nissle 1917 in subgroups of patients with irritable bowel syndrome." *Int. J. Colorectal Dis.* 27.4, pp. 467–474. DOI: 10.1007/s00384-011-1363-9.

Kurtz, S. et al. (2004). "Versatile and open software for comparing large genomes." *Genome Biol.* 5.2, R12. DOI: 10.1186/gb-2004-5-2-r12.

Laemmli, U. K. (1970). "Cleavage of structural proteins during the assembly of the head of bacteriophage T4." *Nature* 227.5259, pp. 680–685.

Laing, C. R. et al. (2009). "*In silico* genomic analyses reveal three distinct lineages of *Escherichia coli* O157:H7, one of which is associated with hyper-virulence." *BMC Genomics* 10, p. 287. DOI: 10.1186/1471-2164-10-287.

Lan, R. and P. R. Reeves (2000). "Intraspecies variation in bacterial genomes: the need for a species genome concept." *Trends Microbiol.* 8.9, pp. 396–401. DOI: 10.1016/S0966-842X(00)01791-1.

Land, M. et al. (2015). "Insights from 20 years of bacterial genome sequencing." *Funct. Integr. Genomics* 15.2, pp. 141–161. DOI: 10.1007/s10142-015-0433-4.

Landraud, L. et al. (2013). "Severity of *Escherichia coli* bacteraemia is independent of the intrinsic virulence of the strains assessed in a mouse model." *Clin. Microbiol. Infect.* 19.1, pp. 85–90. DOI: 10.1111/j.1469-0691.2011.03750.x.

Lane, M. C. et al. (2005). "Role of motility in the colonization of uropathogenic *Escherichia coli* in the urinary tract." *Infect. Immun.* 73.11, pp. 7644–7656. DOI: 10.1128/IAI.73.11.7644-7656.2005.

Lane, M. C. et al. (2007). "Expression of flagella is coincident with uropathogenic *Escherichia coli* ascension to the upper urinary tract." *Proc. Natl. Acad. Sci. U.S.A.* 104.42, pp. 16669–16674. DOI: 10.1073/pnas.0607898104.

Langille, M. G., W. W. Hsiao, and F. S. Brinkman (2008). "Evaluation of genomic island predictors using a comparative genomics approach." *BMC Bioinformatics* 9, p. 329. DOI: 10.1186/1471-2105-9-329.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." *Nat. Methods* 9.4, pp. 357–359. DOI: 10.1038/nmeth.1923.

Lannoy, C. V. de, D. de Ridder, and J. Risse (2017). "A sequencer coming of age: *de novo* genome assembly using MinION reads." *bioRxiv*, p. 142711. DOI: 10.1101/142711.

Larkin, M. A. et al. (2007). "Clustal W and Clustal X version 2.0." *Bioinformatics* 23.21, pp. 2947–2948. DOI: 10.1093/bioinformatics/btm404.

Law, D. et al. (1992). "Detection by ELISA of low numbers of Shiga-like toxin-producing *Escherichia coli* in mixed cultures after growth in the presence of mitomycin C." *J. Med. Microbiol.* 36.3, pp. 198–202. DOI: 10.1099/00222615-36-3-198.

Le Bouguénec, C. and C. Schouler (2011). "Sugar metabolism, an additional virulence factor in enterobacteria." *Int. J. Med. Microbiol.* 301.1, pp. 1–6. DOI: 10.1016/j.ijmm.2010.04.021.

Le Gall, T. et al. (2007). "Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains." *Mol. Biol. Evol.* 24.11, pp. 2373–2384. DOI: 10.1093/molbev/msm172.

Lechner, M. et al. (2011). "Proteinortho: detection of (co-)orthologs in large-scale analysis." *BMC Bioinformatics* 12, p. 124. DOI: 10.1186/1471-2105-12-124.

Lechner, M. et al. (2014). "Orthology detection combining clustering and synteny for very large datasets." *PLoS ONE* 9.8, e105015. DOI: 10.1371/journal.pone.0105015.

Lee, S. et al. (2010). "Phylogenetic groups and virulence factors in pathogenic and commensal strains of *Escherichia coli* and their association with *bla*CTX-M." *Ann. Clin. Lab. Sci.* 40.4, pp. 361–367.

Legendre, P and L Legendre (1998). *Numerical ecology*. 2nd. Amsterdam: Elsevier Science BV. 852 pp.

Lehtolainen, T. et al. (2003). "Association between virulence factors and clinical course of *Escherichia coli* mastitis." *Acta Vet. Scand.* 44 (3-4), pp. 203–205. DOI: 10.1186/1751-0147-44-203.

Leimbach, A. (2016a). "bac-genomics-scripts: Bovine *E. coli* mastitis comparative genomics edition." *Zenodo*. DOI: 10.5281/zenodo.215824.

Leimbach, A. (2016b). "ecoli_VF_collection: v0.1." *Zenodo*. DOI: 10.5281/zenodo.56686.

Leimbach, A., J. Hacker, and U. Dobrindt (2013). "*E. coli* as an all-rounder: the thin line between commensalism and pathogenicity." *Curr. Top. Microbiol. Immunol.* 358, pp. 3–32. DOI: 10.1007/82_2012_303.

Leimbach, A. et al. (2015). "Complete genome sequences of *Escherichia coli* strains 1303 and ECC-1470 isolated from bovine mastitis." *Genome Announc.* 3.2, e00182–15. DOI: 10.1128/genomeA.00182-15.

Leimbach, A. et al. (2016). "Whole-genome draft sequences of six commensal fecal and six mastitis-associated *Escherichia coli* strains of bovine origin." *Genome Announc.* 4.4, e00753–16. DOI: 10.1128/genomeA.00753-16.

Leimbach, A. et al. (2017). "No evidence for a bovine mastitis *Escherichia coli* pathotype." *BMC Genomics* 18.1, p. 359. DOI: 10.1186/s12864-017-3739-x.

Leonard, S. R. et al. (2016). "Hybrid Shiga toxin-producing and enterotoxigenic *Escherichia* sp. cryptic lineage 1 strain 7v harbors a hybrid plasmid." *Appl. Environ. Microbiol.* 82.14, pp. 4309–4319. DOI: 10.1128/AEM.01129-16.

Leopold, S. R. et al. (2011). "Obscured phylogeny and possible recombinational dormancy in *Escherichia coli*." *BMC Evol. Biol.* 11, p. 183. DOI: 10.1186/1471-2148-11-183.

Letunic, I. and P. Bork (2011). "Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy." *Nucleic Acids Res.* 39 (Web Server issue), W475–478. DOI: 10.1093/nar/gkr201.

Léveillé, S. et al. (2006). "Iha from an *Escherichia coli* urinary tract infection outbreak clonal group A strain is expressed *in vivo* in the mouse urinary tract and functions as a catecholate siderophore receptor." *Infect. Immun.* 74.6, pp. 3427–3436. DOI: 10.1128/IAI.00107-06.

Li, B. et al. (2010). "Phylogenetic groups and pathogenicity island markers in fecal *Escherichia coli* isolates from asymptomatic humans in China." *Appl. Environ. Microbiol.* 76.19, pp. 6698–6700. DOI: 10.1128/AEM.00707-10.

Li, D et al. (2011). "Genomic data from *Escherichia coli* O104:H4 isolate TY-2482." *GigaDB*. DOI: 10.5524/100001.

Li, H. et al. (2009). "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 25.16, pp. 2078–2079. DOI: 10.1093/bioinformatics/btp352.

Li, J. et al. (2015). "SecReT6: a web-based resource for type VI secretion systems found in bacteria." *Environ. Microbiol.* 17.7, pp. 2196–2202. DOI: 10.1111/1462-2920.12794.

Linke, D. et al. (2006). "Trimeric autotransporter adhesins: variable structure, common function." *Trends Microbiol.* 14.6, pp. 264–270. DOI: 10.1016/j.tim.2006.04.005.

Lippolis, J. D., D. O. Bayles, and T. A. Reinhardt (2009). "Proteomic changes in *Escherichia coli* when grown in fresh milk versus laboratory media." *J. Proteome Res.* 8.1, pp. 149–158. DOI: 10.1021/pr800458v.

Lira, W. M., C. Macedo, and J. M. Marin (2004). "The incidence of Shiga toxin-producing *Escherichia coli* in cattle with mastitis in Brazil." *J. Appl. Microbiol.* 97.4, pp. 861–866. DOI: 10.1111/j.1365-2672.2004.02384.x.

Liu, B. et al. (2012). "A novel non-homologous recombination-mediated mechanism for *Escherichia coli* unilateral flagellar phase variation." *Nucleic Acids Res.* 40.10, pp. 4530–4538. DOI: 10.1093/nar/gks040.

Liu, D. and P. R. Reeves (1994). "*Escherichia coli* K12 regains its O antigen." *Microbiology (Reading, Engl.)* 140 (Pt 1), pp. 49–57. DOI: 10.1099/13500872-140-1-49.

Liu, Y. et al. (2014). "Phylogenetic group, virulence factors and antimicrobial resistance of *Escherichia coli* associated with bovine mastitis." *Res. Microbiol.* 165.4, pp. 273–277. DOI: 10.1016/j.resmic.2014.03.007.

Livermore, D. M. (1995). "beta-Lactamases in laboratory and clinical resistance." *Clin. Microbiol. Rev.* 8.4, pp. 557–584.

Lloyd, A. L. et al. (2009). "Genomic islands of uropathogenic *Escherichia coli* contribute to virulence." *J. Bacteriol.* 191.11, pp. 3469–3481. DOI: 10.1128/JB.01717-08.

Łobocka, M. B. et al. (2004). “Genome of bacteriophage P1.” *J. Bacteriol.* 186.21, pp. 7032–7068. DOI: 10.1128/JB.186.21.7032-7068.2004.

Locatelli, C. et al. (2009). “Extended-spectrum β-lactamase production in *E. coli* strains isolated from clinical bovine mastitis.” *Vet. Res. Commun.* 33 Suppl 1, pp. 141–144. DOI: 10.1007/s11259-009-9263-y.

Loman, N. J. and M. J. Pallen (2015). “Twenty years of bacterial genome sequencing.” *Nat. Rev. Microbiol.* 13.12, pp. 787–794. DOI: 10.1038/nrmicro3565.

Loman, N. J. and M. Watson (2015). “Successful test launch for nanopore sequencing.” *Nat. Methods* 12.4, pp. 303–304. DOI: 10.1038/nmeth.3327.

Loman, N. J. et al. (2012a). “High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity.” *Nat. Rev. Microbiol.* 10.9, pp. 599–606. DOI: 10.1038/nrmicro2850.

Loman, N. J. et al. (2012b). “Performance comparison of benchtop high-throughput sequencing platforms.” *Nat. Biotechnol.* 30.5, pp. 434–439. DOI: 10.1038/nbt.2198.

Loman, N. J. et al. (2013). “A culture-independent sequence-based metagenomics approach to the investigation of an outbreak of Shiga-toxigenic *Escherichia coli* O104:H4.” *JAMA* 309.14, pp. 1502–1510. DOI: 10.1001/jama.2013.3231.

Loman, N. J., J. Quick, and J. T. Simpson (2015). “A complete bacterial genome assembled de novo using only nanopore sequencing data.” *Nat. Methods* 12.8, pp. 733–735. DOI: 10.1038/nmeth.3444.

Long, E. et al. (2001). “*Escherichia coli* induces apoptosis and proliferation of mammary cells.” *Cell Death Differ.* 8.8, pp. 808–816. DOI: 10.1038/sj.cdd.4400878.

Lowe, T. M. and S. R. Eddy (1997). “tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.” *Nucleic Acids Res.* 25.5, pp. 955–964. DOI: 10.1093/nar/25.5.0955.

Lu, S. et al. (2011). “Complete genome sequence of the neonatal-meningitis-associated *Escherichia coli* strain CE10.” *J. Bacteriol.* 193.24, p. 7005. DOI: 10.1128/JB.06284-11.

Lügering, A. et al. (2003). “The Pix pilus adhesin of the uropathogenic *Escherichia coli* strain X2194 (O2 : K(-): H6) is related to Pap pili but exhibits a truncated regulatory region.” *Microbiology (Reading, Engl.)* 149 (Pt 6), pp. 1387–1397. DOI: 10.1099/mic.0.26266-0.

Luheshi, L. M., S. Raza, and S. J. Peacock (2015). “Moving pathogen genomics out of the lab and into the clinic: what will it take?” *Genome Med.* 7.1, p. 132. DOI: 10.1186/s13073-015-0254-z.

Lukjancenko, O., T. M. Wassenaar, and D. W. Ussery (2010). “Comparison of 61 sequenced *Escherichia coli* genomes.” *Microb. Ecol.* 60.4, pp. 708–720. DOI: 10.1007/s00248-010-9717-3.

Lum, M. and R. Morona (2012). “IcsA autotransporter passenger promotes increased fusion protein expression on the cell surface.” *Microb. Cell Fact.* 11, p. 20. DOI: 10.1186/1475-2859-11-20.

Luo, C. et al. (2011). “Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species.” *Proc. Natl. Acad. Sci. U.S.A.* 108.17, pp. 7200–7205. DOI: 10.1073/pnas.1015622108.

Ma, J. et al. (2013). “Genetic diversity and features analysis of type VI secretion systems loci in avian pathogenic *Escherichia coli* by wide genomic scanning.” *Infect. Genet. Evol.* 20, pp. 454–464. DOI: 10.1016/j.meegid.2013.09.031.

MacLean, D., J. D. G. Jones, and D. J. Studholme (2009). “Application of 'next-generation' sequencing technologies to microbial genetics.” *Nat. Rev. Microbiol.* 7.4, pp. 287–296. DOI: 10.1038/nrmicro2122.

MacLean, L. L. and M. B. Perry (1997). “Structural characterization of the serotype O:5 O-polysaccharide antigen of the lipopolysaccharide of *Escherichia coli* O:5.” *Biochem. Cell Biol.* 75.3, pp. 199–205.

Maiden, M. C. J. (2006). “Multilocus sequence typing of bacteria.” *Annu. Rev. Microbiol.* 60, pp. 561–588. DOI: 10.1146/annurev.micro.59.030804.121325.

Manrique, M. et al. (2011). “Escherichia coli EHEC Germany outbreak preliminary functional annotation using BG7 system.” *Nature Precedings* 713. DOI: 10.1038/npre.2011.6001.1.

Mao, C. et al. (2015). “Curation, integration and visualization of bacterial virulence factors in PATRIC.” *Bioinformatics* 31.2, pp. 252–258. DOI: 10.1093/bioinformatics/btu631.

Marchler-Bauer, A. et al. (2015). “CDD: NCBI’s conserved domain database.” *Nucleic Acids Res.* 43 (Database issue), pp. D222–226. DOI: 10.1093/nar/gku1221.

Margulies, M. et al. (2005). “Genome sequencing in microfabricated high-density picolitre reactors.” *Nature* 437.7057, pp. 376–380. DOI: 10.1038/nature03959.

Markowitz, V. M. et al. (2009). “IMG ER: a system for microbial genome annotation expert review and curation.” *Bioinformatics* 25.17, pp. 2271–2278. DOI: 10.1093/bioinformatics/btp393.

Martin, M. (2011). “Cutadapt removes adapter sequences from high-throughput sequencing reads.” *EMBnetJ* 17.1, p. 10. DOI: 10.14806/ej.17.1.200.

Martin, W. and T. M. Embley (2004). “Evolutionary biology: Early evolution comes full circle.” *Nature* 431.7005, pp. 134–137. DOI: 10.1038/431134a.

Marx, V. (2016). “Microbiology: the road to strain-level identification.” *Nat. Methods* 13.5, pp. 401–404. DOI: 10.1038/nmeth.3837.

Matsuzaki, S. et al. (2014). “Perspective: The age of the phage.” *Nature* 509.7498, S9–S9. DOI: 10.1038/509S9a.

Mau, B. et al. (2006). “Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*.” *Genome Biol.* 7.5, R44. DOI: 10.1186/gb-2006-7-5-r44.

McAdam, P. R., E. J. Richardson, and J. R. Fitzgerald (2014). “High-throughput sequencing for the study of bacterial pathogen biology.” *Curr. Opin. Microbiol.* 19, pp. 106–113. DOI: 10.1016/j.mib.2014.06.002.

McInerney, J. O., A. McNally, and M. J. O’Connell (2017). “Why prokaryotes have pangenomes.” *Nat. Microbiol.* 2, p. 17040. DOI: 10.1038/nmicrobiol.2017.40.

McNally, A. et al. (2013). “The evolutionary path to extraintestinal pathogenic, drug-resistant *Escherichia coli* is marked by drastic reduction in detectable recombination within the core genome.” *Genome Biol. Evol.* 5.4, pp. 699–710. DOI: 10.1093/gbe/evt038.

McPherson, J. D. (2009). “Next-generation gap.” *Nat. Methods* 6 (11 Suppl), S2–5. DOI: 10.1038/nmeth.f.268.

Medini, D. et al. (2005). “The microbial pan-genome.” *Curr. Opin. Genet. Dev.* 15.6, pp. 589–594. DOI: 10.1016/j.gde.2005.09.006.

Medini, D. et al. (2008). “Microbiology in the post-genomic era.” *Nat. Rev. Microbiol.* 6.6, pp. 419–430. DOI: 10.1038/nrmicro1901.

Mellmann, A. et al. (2008). “Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic *Escherichia coli*.” *Emerging Infect. Dis.* 14.8, pp. 1287–1290. DOI: 10.3201/eid1408.071082.

Mellmann, A. et al. (2011). “Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology.” *PLoS ONE* 6.7, e22751. DOI: 10.1371/journal.pone.0022751.

Mellmann, A. et al. (2016). “Real-time genome sequencing of resistant bacteria provides precision infection control in an institutional setting.” *J. Clin. Microbiol.* 54.12, pp. 2874–2881. DOI: 10.1128/JCM.00790-16.

Mesibov, R. and J. Adler (1972). “Chemotaxis toward amino acids in *Escherichia coli*.” *J. Bacteriol.* 112.1, pp. 315–326.

Milkman, R. and M. M. Bridges (1990). “Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames.” *Genetics* 126.3, pp. 505–517.

Miquel, S. et al. (2010). “Complete genome sequence of Crohn’s disease-associated adherent-invasive *E. coli* strain LF82.” *PLoS ONE* 5.9. DOI: 10.1371/journal.pone.0012714.

Monecke, S. et al. (2011). “Presence of enterohemorrhagic *Escherichia coli* ST678/O104:H4 in France prior to 2011.” *Appl. Environ. Microbiol.* 77.24, pp. 8784–8786. doi: 10.1128/AEM.06524-11.

Monteiro, C. et al. (2009). “Characterization of cellulose production in *Escherichia coli* Nissle 1917 and its biological consequences.” *Environ. Microbiol.* 11.5, pp. 1105–1116. doi: 10.1111/j.1462-2920.2008.01840.x.

Morabito, S. et al. (1998). “Enteroaggregative, Shiga toxin-producing *Escherichia coli* O111:H2 associated with an outbreak of hemolytic-uremic syndrome.” *J. Clin. Microbiol.* 36.3, pp. 840–842.

Mordhorst, I. L. et al. (2009). “O-acetyltransferase gene *neuO* is segregated according to phylogenetic background and contributes to environmental desiccation resistance in *Escherichia coli* K1.” *Environ. Microbiol.* 11.12, pp. 3154–3165. doi: 10.1111/j.1462-2920.2009.02019.x.

Moreno-Hagelsieb, G. and K. Latimer (2008). “Choosing BLAST options for better detection of orthologs as reciprocal best hits.” *Bioinformatics* 24.3, pp. 319–324. doi: 10.1093/bioinformatics/btm585.

Moriel, D. G. et al. (2010). “Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*.” *Proc. Natl. Acad. Sci. U.S.A.* 107.20, pp. 9072–9077. doi: 10.1073/pnas.0915077107.

Mossoro, C. et al. (2002). “Chronic diarrhea, hemorrhagic colitis, and hemolytic-uremic syndrome associated with HEp-2 adherent *Escherichia coli* in adults infected with human immunodeficiency virus in Bangui, Central African Republic.” *J. Clin. Microbiol.* 40.8, pp. 3086–3088. doi: 10.1128/JCM.40.8.3086-3088.2002.

Moulin-Schouleur, M. et al. (2007). “Extraintestinal pathogenic *Escherichia coli* strains of avian and human origin: link between phylogenetic relationships and common virulence patterns.” *J. Clin. Microbiol.* 45.10, pp. 3366–3376. doi: 10.1128/JCM.00037-07.

Mullane, N. et al. (2008). “Molecular analysis of the *Enterobacter sakazakii* O-antigen gene locus.” *Appl. Environ. Microbiol.* 74.12, pp. 3783–3794. doi: 10.1128/AEM.02302-07.

Müller-Loennies, S., B. Lindner, and H. Brade (2002). “Structural analysis of deacylated lipopolysaccharide of *Escherichia coli* strains 2513 (R4 core-type) and F653 (R3 core-type).” *Eur. J. Biochem.* 269.23, pp. 5982–5991. doi: 10.1046/j.1432-1033.2002.03322.x.

Müller-Loennies, S., B. Lindner, and H. Brade (2003). “Structural analysis of oligosaccharides from lipopolysaccharide (LPS) of *Escherichia coli* K12 strain W3100 reveals a link between inner and outer core LPS biosynthesis.” *J. Biol. Chem.* 278.36, pp. 34090–34101. doi: 10.1074/jbc.M303985200.

Müller, D. et al. (2007). “Identification of unconventional intestinal pathogenic *Escherichia coli* isolates expressing intermediate virulence factor profiles by using a novel single-step multiplex PCR.” *Appl. Environ. Microbiol.* 73.10, pp. 3380–3390. doi: 10.1128/AEM.02855-06.

Nagarajan, N. et al. (2010). “Finishing genomes with limited resources: lessons from an ensemble of microbial genomes.” *BMC Genomics* 11, p. 242. doi: 10.1186/1471-2164-11-242.

Nash, J. H. et al. (2010). “Genome sequence of adherent-invasive *Escherichia coli* and comparative genomic analysis with other *E. coli* pathotypes.” *BMC Genomics* 11, p. 667. doi: 10.1186/1471-2164-11-667.

Nataro, J. P. et al. (1987). “Patterns of adherence of diarrheagenic *Escherichia coli* to HEp-2 cells.” *Pediatr. Infect. Dis. J.* 6.9, pp. 829–831.

Nataro, J. P. et al. (1992). “Aggregative adherence fimbriae I of enteroaggregative *Escherichia coli* mediate adherence to HEp-2 cells and hemagglutination of human erythrocytes.” *Infect. Immun.* 60.6, pp. 2297–2304.

Nataro, J. P. et al. (1995). “Heterogeneity of enteroaggregative *Escherichia coli* virulence demonstrated in volunteers.” *J. Infect. Dis.* 171.2, pp. 465–468. doi: 10.1093/infdis/171.2.465.

Navarro-García, F. et al. (1998). "*In vitro* effects of a high-molecular-weight heat-labile enterotoxin from enteroaggregative *Escherichia coli*." *Infect. Immun.* 66.7, pp. 3149–3154.

Nederbragt, L. (2016). "developments in NGS." *figshare*. DOI: 10.6084/m9.figshare.100940.v9.

Nemeth, J., C. A. Muckle, and R. Y. Lo (1991). "Serum resistance and the *traT* gene in bovine mastitis-causing *Escherichia coli*." *Vet. Microbiol.* 28.4, pp. 343–351. DOI: 10.1016/0378-1135(91)90069-R.

Nemeth, J., C. A. Muckle, and C. L. Gyles (1994). "*In vitro* comparison of bovine mastitis and fecal *Escherichia coli* isolates." *Vet. Microbiol.* 40 (3-4), pp. 231–238. DOI: 10.1016/0378-1135(94)90112-0.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2. URL: https://CRAN.R-project.org/package=RColorBrewer.

Nielsen, K. L. et al. (2016). "Adaptation of *Escherichia coli* traversing from the faecal environment to the urinary tract." *Int. J. Med. Microbiol.* 306.8, pp. 595–603. DOI: 10.1016/j.ijmm.2016.10.005.

Nougayrède, J.-P. et al. (2006). "*Escherichia coli* induces DNA double-strand breaks in eukaryotic cells." *Science* 313.5788, pp. 848–851. DOI: 10.1126/science.1127059.

Novick, R. P. and C. Roth (1968). "Plasmid-linked resistance to inorganic salts in *Staphylococcus aureus*." *J. Bacteriol.* 95.4, pp. 1335–1342.

Nowrouzian, F. L. et al. (2009). "Phylogenetic group B2 *Escherichia coli* strains from the bowel microbiota of Pakistani infants carry few virulence genes and lack the capacity for long-term persistence." *Clin. Microbiol. Infect.* 15.5, pp. 466–472. DOI: 10.1111/j.1469-0691.2009.02706.x.

Nowrouzian, F., I. Adlerberth, and A. E. Wold (2001). "P fimbriae, capsule and aerobactin characterize colonic resident *Escherichia coli*." *Epidemiol. Infect.* 126.1, pp. 11–18.

Nowrouzian, F. L. and E. Oswald (2012). "*Escherichia coli* strains with the capacity for long-term persistence in the bowel microbiota carry the potentially genotoxic *pks* island." *Microb. Pathog.* 53 (3-4), pp. 180–182. DOI: 10.1016/j.micpath.2012.05.011.

Nowrouzian, F. L., A. E. Wold, and I. Adlerberth (2005). "*Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants." *J. Infect. Dis.* 191.7, pp. 1078–1083. DOI: 10.1086/427996.

Nowrouzian, F. L., I. Adlerberth, and A. E. Wold (2006). "Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells." *Microbes Infect.* 8.3, pp. 834–840. DOI: 10.1016/j.micinf.2005.10.011.

Nowrouzian, F. et al. (2003). "*Escherichia coli* in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage." *Pediatr. Res.* 54.1, pp. 8–14. DOI: 10.1203/01.PDR.0000069843.20655.EE.

Nyholm, O. et al. (2015). "Comparative genomics and characterization of hybrid shigatoxigenic and enterotoxigenic *Escherichia coli* (STEC/ETEC) strains." *PLoS ONE* 10.8, e0135936. DOI: 10.1371/journal.pone.0135936.

O'Brien, C. L. et al. (2016). "Comparative genomics of Crohn's disease-associated adherent-invasive *Escherichia coli*." *Gut*. DOI: 10.1136/gutjnl-2015-311059.

Ochman, H. and R. K. Selander (1984). "Standard reference strains of *Escherichia coli* from natural populations." *J. Bacteriol.* 157.2, pp. 690–693.

Oehler, D. et al. (2012). "Genome-guided analysis of physiological and morphological traits of the fermentative acetate oxidizer *Thermacetogenium phaeum*." *BMC Genomics* 13, p. 723. DOI: 10.1186/1471-2164-13-723.

Ogura, Y. et al. (2009). "Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*." *Proc. Natl. Acad. Sci. U.S.A.* 106.42, pp. 17939–17944. DOI: 10.1073/pnas.0903585106.

Oikonomou, G. et al. (2012). "Microbial diversity of bovine mastitic milk as described by pyrosequencing of metagenomic 16s rDNA." *PLoS ONE* 7.10, e47671. DOI: 10.1371/journal.pone.0047671.

Okonechnikov, K., A. Conesa, and F. García-Alcalde (2016). "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data." *Bioinformatics* 32.2, pp. 292–294. DOI: 10.1093/bioinformatics/btv566.

Olesen, B. et al. (2012). "Enteroaggregative *Escherichia coli* O78:H10, the cause of an outbreak of urinary tract infection." *J. Clin. Microbiol.* 50.11, pp. 3703–3711. DOI: 10.1128/JCM.01909-12.

Olier, M. et al. (2012). "Genotoxicity of *Escherichia coli* Nissle 1917 strain cannot be dissociated from its probiotic activity." *Gut Microbes* 3.6, pp. 501–509. DOI: 10.4161/gmic.21737.

O'Neill, J. (2016). *The review on antimicrobial resistance. Tackling drug-resistant infections globally: final report and recommendations.* URL: https://amr-review.org/.

Orth, D. et al. (2007). "Variability in tellurite resistance and the *ter* gene cluster among Shiga toxin-producing *Escherichia coli* isolated from humans, animals and food." *Res. Microbiol.* 158.2, pp. 105–111. DOI: 10.1016/j.resmic.2006.10.007.

Oshima, K. et al. (2008). "Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult." *DNA Res.* 15.6, pp. 375–386. DOI: 10.1093/dnares/dsn026.

Ostblom, A. et al. (2011). "Pathogenicity island markers, virulence determinants *malX* and *usp*, and the capacity of *Escherichia coli* to persist in infants' commensal microbiotas." *Appl. Environ. Microbiol.* 77.7, pp. 2303–2308. DOI: 10.1128/AEM.02405-10.

Paape, M. J. et al. (2003). "The bovine neutrophil: structure and function in blood and milk." *Vet. Res.* 34.5, pp. 597–627. DOI: 10.1051/vetres:2003024.

Page, A. J. et al. (2015). "Roary: rapid large-scale prokaryote pan genome analysis." *Bioinformatics* 31.22, pp. 3691–3693. DOI: 10.1093/bioinformatics/btv421.

Pallen, M. J. (2014). "Diagnostic metagenomics: potential applications to bacterial, viral and parasitic infections." *Parasitology* 141.14, pp. 1856–1862. DOI: 10.1017/S0031182014000134.

Pallen, M. J. (2016). "Microbial bioinformatics 2020." *Microb. Biotechnol.* 9.5, pp. 681–686. DOI: 10.1111/1751-7915.12389.

Pallen, M. J. and N. J. Loman (2011). "Are diagnostic and public health bacteriology ready to become branches of genomic medicine?" *Genome Med.* 3, p. 53. DOI: 10.1186/gm269.

Passey, S., A. Bradley, and H. Mellor (2008). "*Escherichia coli* isolated from bovine mastitis invade mammary cells by a modified endocytic pathway." *Vet. Microbiol.* 130 (1-2), pp. 151–164. DOI: 10.1016/j.vetmic.2008.01.003.

Paterson, D. L. and R. A. Bonomo (2005). "Extended-spectrum beta-lactamases: a clinical update." *Clin. Microbiol. Rev.* 18.4, pp. 657–686. DOI: 10.1128/CMR.18.4.657-686.2005.

Peabody, M. A. et al. (2015). "Evaluation of shotgun metagenomics sequence classification methods using *in silico* and *in vitro* simulated communities." *BMC Bioinformatics* 16, p. 363. DOI: 10.1186/s12859-015-0788-5.

Perelle, S. et al. (2002). "Identification of the O-antigen biosynthesis genes of *Escherichia coli* O91 and development of a O91 PCR serotyping test." *J. Appl. Microbiol.* 93.5, pp. 758–764. DOI: 10.1046/j.1365-2672.2002.01743.x.

Perkel, J. (2016). "Democratic databases: science on GitHub." *Nature* 538.7623, pp. 127–128. DOI: 10.1038/538127a.

Perna, N. T. et al. (2001). "Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7." *Nature* 409.6819, pp. 529–533. DOI: 10.1038/35054089.

Petty, N. K. et al. (2014). "Global dissemination of a multidrug resistant *Escherichia coli* clone." *Proc. Natl. Acad. Sci. U.S.A.* 111.15, pp. 5694–5699. DOI: 10.1073/pnas.1322678111.

Petzl, W. et al. (2008). "*Escherichia coli*, but not *Staphylococcus aureus* triggers an early increased expression of factors contributing to the innate immune defense in the udder of the cow." *Vet. Res.* 39.2, p. 18. DOI: 10.1051/vetres:2007057.

Pevzner, P. and R. Shamir (2009). "Computing has changed biology–biology education must catch up." *Science* 325.5940, pp. 541–542. DOI: 10.1126/science.1173876.

Pföstl, A. et al. (2008). "Biosynthesis of dTDP-3-acetamido-3,6-dideoxy-alpha-D-glucose." *Biochem. J.* 410.1, pp. 187–194. DOI: 10.1042/BJ20071044.

Phillippy, A. M. (2017). "New advances in sequence assembly." *Genome Res.* 27.5, pp. xi–xiii. DOI: 10.1101/gr.223057.117.

Picard, B. et al. (1999). "The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection." *Infect. Immun.* 67.2, pp. 546–553.

Pirofski, L.-A. and A. Casadevall (2012). "Q and A: What is a pathogen? A question that begs the point." *BMC Biol.* 10, p. 6. DOI: 10.1186/1741-7007-10-6.

Poehlein, A. et al. (2015). "Genome sequence of *Clostridium sporogenes* DSM 795(T), an amino acid-degrading, nontoxic surrogate of neurotoxin-producing *Clostridium botulinum*." *Stand. Genomic Sci.* 10, p. 40. DOI: 10.1186/s40793-015-0016-y.

Pomeroy, B. et al. (2016). "Intramammary immunization with ultraviolet-killed *Escherichia coli* shows partial protection against late gestation intramammary challenge with a homologous strain." *J. Dairy Sci.* DOI: 10.3168/jds.2016-11149.

Pop, M. and S. L. Salzberg (2008). "Bioinformatics challenges of new sequencing technology." *Trends Genet.* 24.3, pp. 142–149. DOI: 10.1016/j.tig.2007.12.006.

Pop, M. and S. L. Salzberg (2015). "Use and mis-use of supplementary material in science publications." *BMC Bioinformatics* 16, p. 237. DOI: 10.1186/s12859-015-0668-z.

Popescu, A.-A., K. T. Huber, and E. Paradis (2012). "ape 3.0: new tools for distance-based phylogenetics and evolutionary analysis in R." *Bioinformatics* 28.11, pp. 1536–1537. DOI: 10.1093/bioinformatics/bts184.

Porcherie, A. et al. (2012). "Repertoire of *Escherichia coli* agonists sensed by innate immunity receptors of the bovine udder and mammary epithelial cells." *Vet. Res.* 43, p. 14. DOI: 10.1186/1297-9716-43-14.

Porcheron, G. et al. (2012). "Effect of fructooligosaccharide metabolism on chicken colonization by an extra-intestinal pathogenic *Escherichia coli* strain." *PLoS ONE* 7.4, e35475. DOI: 10.1371/journal.pone.0035475.

Putze, J. et al. (2009). "Genetic structure and distribution of the colibactin genomic island among members of the family *Enterobacteriaceae*." *Infect. Immun.* 77.11, pp. 4696–4703. DOI: 10.1128/IAI.00522-09.

Qin, J. et al. (2011). "Identification of the Shiga toxin-producing *Escherichia coli* O104:H4 strain responsible for a food poisoning outbreak in Germany by PCR." *J. Clin. Microbiol.* 49.9, pp. 3439–3440. DOI: 10.1128/JCM.01312-11.

Quick, J. et al. (2015). "Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*." *Genome Biol.* 16, p. 114. DOI: 10.1186/s13059-015-0677-2.

Quinn, G. P. and M. J. Keough (2002). *Experimental design and data analysis for biologists*. Cambridge University Press.

R Core Team (2017). *R: a language and environment for statistical computing*. Vienna, Austria. URL: https://www.R-project.org/.

Raetz, C. R. H. and C. Whitfield (2002). "Lipopolysaccharide endotoxins." *Annu. Rev. Biochem.* 71, pp. 635–700. DOI: 10.1146/annurev.biochem.71.110601.135414.

Rainard, P. and C. Riollet (2006). "Innate immunity of the bovine mammary gland." *Vet. Res.* 37.3, pp. 369–400. DOI: 10.1051/vetres:2006007.

Rainard, P., P. Cunha, and F. B. Gilbert (2016). "Innate and adaptive immunity synergize to trigger inflammation in the mammary gland." *PLoS ONE* 11.4, e0154172. DOI: 10.1371/journal.pone.0154172.

Ramette, A. (2007). "Multivariate analyses in microbial ecology." *FEMS Microbiol. Ecol.* 62.2, pp. 142–160. DOI: 10.1111/j.1574-6941.2007.00375.x.

Rasko, D. A. et al. (2008). "The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates." *J. Bacteriol.* 190.20, pp. 6881–6893. DOI: 10.1128/JB.00619-08.

Rasko, D. A. et al. (2011). "Origins of the *E. coli* strain causing an outbreak of hemolytic-uremic syndrome in Germany." *N. Engl. J. Med.* 365.8, pp. 709–717. DOI: 10.1056/NEJMoa1106920.

Reeves, P. P. and L. Wang (2002). "Genomic organization of LPS-specific loci." *Curr. Top. Microbiol. Immunol.* 264.1, pp. 109–135. DOI: 10.1007/978-3-642-56031-6_7.

Reid, S. D. et al. (2000). "Parallel evolution of virulence in pathogenic *Escherichia coli*." *Nature* 406.6791, pp. 64–67. DOI: 10.1038/35017546.

Reidl, S. et al. (2009). "Impact of O-glycosylation on the molecular and cellular adhesion properties of the *Escherichia coli* autotransporter protein Ag43." *Int. J. Med. Microbiol.* 299.6, pp. 389–401. DOI: 10.1016/j.ijmm.2009.01.001.

Relman, D. A. (2013). "Metagenomics, infectious disease diagnostics, and outbreak investigations: sequence first, ask questions later?" *JAMA* 309.14, pp. 1531–1532. DOI: 10.1001/jama.2013.3678.

Ren, C.-P. et al. (2004). "The ETT2 gene cluster, encoding a second type III secretion system from *Escherichia coli*, is present in the majority of strains but has undergone widespread mutational attrition." *J. Bacteriol.* 186.11, pp. 3547–3560. DOI: 10.1128/JB.186.11.3547-3560.2004.

Ren, C.-P. et al. (2005). "The Flag-2 locus, an ancestral gene cluster, is potentially associated with a novel flagellar system from *Escherichia coli*." *J. Bacteriol.* 187.4, pp. 1430–1440. DOI: 10.1128/JB.187.4.1430-1440.2005.

Restieri, C. et al. (2007). "Autotransporter-encoding sequences are phylogenetically distributed among *Escherichia coli* clinical isolates and reference strains." *Appl. Environ. Microbiol.* 73.5, pp. 1553–1562. DOI: 10.1128/AEM.01542-06.

Rice, P., I. Longden, and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." *Trends Genet.* 16.6, pp. 276–277. DOI: 10.1016/S0168-9525(00)02024-2.

Richards, V. P. et al. (2015). "Genome based phylogeny and comparative genomic analysis of intra-mammary pathogenic *Escherichia coli*." *PLoS ONE* 10.3, e0119799. DOI: 10.1371/journal.pone.0119799.

Richter, A. M. et al. (2014). "Cyclic-di-GMP signalling and biofilm-related properties of the Shiga toxin-producing 2011 German outbreak *Escherichia coli* O104:H4." *EMBO Mol. Med.* 6.12, pp. 1622–1637. DOI: 10.15252/emmm.201404309.

Rijavec, M. et al. (2008). "Virulence factors and biofilm production among *Escherichia coli* strains causing bacteraemia of urinary tract origin." *J. Med. Microbiol.* 57 (Pt 11), pp. 1329–1334. DOI: 10.1099/jmm.0.2008/002543-0.

Risse, J. et al. (2015). "A single chromosome assembly of *Bacteroides fragilis* strain BE1 from Illumina and MinION nanopore sequencing data." *Gigascience* 4.1, pp. 1–7. DOI: 10.1186/s13742-015-0101-6.

Robins-Browne, R. M. et al. (2016). "Are *Escherichia coli* pathotypes still relevant in the era of whole-genome sequencing?" *Front. Cell. Infect. Microbiol.* 6, p. 141. DOI: 10.3389/fcimb.2016.00141.

Rodriguez-Siek, K. E. et al. (2005). "Comparison of *Escherichia coli* isolates implicated in human urinary tract infection and avian colibacillosis." *Microbiology (Reading, Engl.)* 151 (Pt 6), pp. 2097–2110. DOI: 10.1099/mic.0.27499-0.

Roesch, P. L. et al. (2003). "Uropathogenic *Escherichia coli* use D-serine deaminase to modulate infection of the murine urinary tract." *Mol. Microbiol.* 49.1, pp. 55–67.

Roggenkamp, A. et al. (1996). "Deletion of amino acids 29 to 81 in adhesion protein YadA of *Yersinia enterocolitica* serotype O:8 results in selective abrogation of adherence to neutrophils." *Infect. Immun.* 64.7, pp. 2506–2514.

Rohde, H. et al. (2011). "Open-source genomic analysis of Shiga-toxin-producing *E. coli* O104:H4." *N. Engl. J. Med.* 365.8, pp. 718–724. DOI: 10.1056/NEJMoa1107643.

Roos, V. et al. (2006). "Asymptomatic bacteriuria *Escherichia coli* strain 83972 carries mutations in the *foc* locus and is unable to express F1C fimbriae." *Microbiology (Reading, Engl.)* 152 (Pt 6), pp. 1799–1806. DOI: 10.1099/mic.0.28711-0.

Rouli, L. et al. (2015). "The bacterial pangenome as a new tool for analysing pathogenic bacteria." *New Microbes New Infect.* 7, pp. 72–85. DOI: 10.1016/j.nmni.2015.06.005.

Rouquet, G. et al. (2009). "A metabolic operon in extraintestinal pathogenic *Escherichia coli* promotes fitness under stressful conditions and invasion of eukaryotic cells." *J. Bacteriol.* 191.13, pp. 4427–4440. DOI: 10.1128/JB.00103-09.

Rump, L. V. et al. (2011). "Draft genome sequences of six *Escherichia coli* isolates from the stepwise model of emergence of *Escherichia coli* O157:H7." *J. Bacteriol.* 193.8, pp. 2058–2059. DOI: 10.1128/JB.00118-11.

Rutherford, K. et al. (2000). "Artemis: sequence visualization and annotation." *Bioinformatics* 16.10, pp. 944–945. DOI: 10.1093/bioinformatics/16.10.944.

Sabaté, M. et al. (2006). "Pathogenicity island markers in commensal and uropathogenic *Escherichia coli* isolates." *Clin. Microbiol. Infect.* 12.9, pp. 880–886. DOI: 10.1111/j.1469-0691.2006.01461.x.

Sadovskaya, I. et al. (2000). "Structural characterization of the outer core and the O-chain linkage region of lipopolysaccharide from *Pseudomonas aeruginosa* serotype O5." *Eur. J. Biochem.* 267.6, pp. 1640–1650. DOI: 10.1046/j.1432-1327.2000.01156.x.

Sahl, J. W. et al. (2011). "A comparative genomic analysis of diverse clonal types of enterotoxigenic *Escherichia coli* reveals pathovar-specific conservation." *Infect. Immun.* 79.2, pp. 950–960. DOI: 10.1128/IAI.00932-10.

Sahl, J. W., M. N. Matalka, and D. A. Rasko (2012). "Phylomark, a tool to identify conserved phylogenetic markers from whole-genome alignments." *Appl. Environ. Microbiol.* 78.14, pp. 4884–4892. DOI: 10.1128/AEM.00929-12.

Sahl, J. W. et al. (2015). "Examination of the enterotoxigenic *Escherichia coli* population structure during human infection." *MBio* 6.3, e00501. DOI: 10.1128/mBio.00501-15.

Salipante, S. J. et al. (2015). "Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains." *Genome Res.* 25.1, pp. 119–128. DOI: 10.1101/gr.180190.114.

Salvador, E. et al. (2012). "Comparison of asymptomatic bacteriuria *Escherichia coli* isolates from healthy individuals versus those from hospital patients shows that long-term bladder colonization selects for attenuated virulence phenotypes." *Infect. Immun.* 80.2, pp. 668–678. DOI: 10.1128/IAI.06191-11.

Sambrook, J., E. F. Fritsch, and T. Maniatis (1989). *Molecular cloning: a laboratory manual*. 2nd ed. Cold spring harbor laboratory press.

Samuel, G. and P. Reeves (2003). "Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly." *Carbohydr. Res.* 338.23, pp. 2503–2519. DOI: 10.1016/j.carres.2003.07.009.

Sanger, F., S. Nicklen, and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." *Proc. Natl. Acad. Sci. U.S.A.* 74.12, pp. 5463–5467.

Sawardeker, J. S., J. H. Sloneker, and A. Jeanes (1965). "Quantitative determination of monosaccharides as their alditol acetates by gas liquid chromatography." *Anal. Chem.* 37.12, pp. 1602–1604. DOI: 10.1021/ac60231a048.

Sboner, A. et al. (2011). "The real cost of sequencing: higher than you think!" *Genome Biol.* 12.8, p. 125. DOI: 10.1186/gb-2011-12-8-125.

Schembri, M. A., D. Dalsgaard, and P. Klemm (2004). "Capsule shields the function of short bacterial adhesins." *J. Bacteriol.* 186.5, pp. 1249–1257. DOI: 10.1128/JB.186.5.1249-1257.2004.

Schierack, P. et al. (2008). "ExPEC-typical virulence-associated genes correlate with successful colonization by intestinal *E. coli* in a small piglet group." *Environ. Microbiol.* 10.7, pp. 1742–1751. DOI: 10.1111/j.1462-2920.2008.01595.x.

Schierack, P. et al. (2011). "*E. coli* Nissle 1917 affects *Salmonella* adhesion to porcine intestinal epithelial cells." *PLoS ONE* 6.2, e14712. DOI: 10.1371/journal.pone.0014712.

Schlee, M. et al. (2007). "Induction of human beta-defensin 2 by the probiotic *Escherichia coli* Nissle 1917 is mediated through flagellin." *Infect. Immun.* 75.5, pp. 2399–2407. DOI: 10.1128/IAI.01563-06.

Schloss, P. D. et al. (2009). "Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities." *Appl. Environ. Microbiol.* 75.23, pp. 7537–7541. DOI: 10.1128/AEM.01541-09.

Schmeisser, C. et al. (2009). "*Rhizobium* sp. strain NGR234 possesses a remarkable number of secretion systems." *Appl. Environ. Microbiol.* 75.12, pp. 4035–4045. DOI: 10.1128/AEM.00515-09.

Schneider, G. et al. (2004). "The pathogenicity island-associated K15 capsule determinant exhibits a novel genetic structure and correlates with virulence in uropathogenic *Escherichia coli* strain 536." *Infect. Immun.* 72.10, pp. 5993–6001. DOI: 10.1128/IAI.72.10.5993-6001.2004.

Schneider, G. et al. (2011). "Mobilisation and remobilisation of a large archetypal pathogenicity island of uropathogenic *Escherichia coli in vitro* support the role of conjugation for horizontal transfer of genomic islands." *BMC Microbiol.* 11, p. 210. DOI: 10.1186/1471-2180-11-210.

Scholz, M. et al. (2016). "Strain-level microbial epidemiology and population genomics from shotgun metagenomics." *Nat. Methods* 13.5, pp. 435–438. DOI: 10.1038/nmeth.3802.

Schouler, C. et al. (2009). "A genomic island of an extraintestinal pathogenic *Escherichia coli* strain enables the metabolism of fructooligosaccharides, which improves intestinal colonization." *J. Bacteriol.* 191.1, pp. 388–393. DOI: 10.1128/JB.01052-08.

Schreiber, H. L. et al. (2017). "Bacterial virulence phenotypes of *Escherichia coli* and host susceptibility determine risk for urinary tract infections." *Sci. Transl. Med.* 9.382. DOI: 10.1126/scitranslmed.aaf1283.

Schubert, S. et al. (2009). "Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species." *PLoS Pathog.* 5.1, e1000257. DOI: 10.1371/journal.ppat.1000257.

Schukken, Y. H. et al. (2011). "Host-response patterns of intramammary infections in dairy cows." *Vet. Immunol. Immunopathol.* 144 (3-4), pp. 270–289. DOI: 10.1016/j.vetimm.2011.08.022.

Schultz, M. (2008). "Clinical use of *E. coli* Nissle 1917 in inflammatory bowel disease." *Inflamm. Bowel Dis.* 14.7, pp. 1012–1018. DOI: 10.1002/ibd.20377.

Schürch, A. C. and W. van Schaik (2017). "Challenges and opportunities for whole-genome sequencing–based surveillance of antibiotic resistance." *Ann. N.Y. Acad. Sci.* 1388.1, pp. 108–120. DOI: 10.1111/nyas.13310.

Schwan, W. R. (2008). "Flagella allow uropathogenic *Escherichia coli* ascension into murine kidneys." *Int. J. Med. Microbiol.* 298 (5-6), pp. 441–447. DOI: 10.1016/j.ijmm.2007.05.009.

Sears, H. J. and I. Brownlee (1952). "Further observations on the persistence of individual strains of *Escherichia coli* in the intestinal tract of man." *J. Bacteriol.* 63.1, pp. 47–57.

Sears, H. J., I. Brownlee, and J. K. Uchiyama (1950). "Persistence of individual strains of *Escherichia coli* in the intestinal tract of man." *J. Bacteriol.* 59.2, pp. 293–301.

Seemann, T. (2013). "Ten recommendations for creating usable bioinformatics command line software." *Gigascience* 2.1, p. 15. DOI: 10.1186/2047-217X-2-15.

Seemann, T. (2014). "Prokka: rapid prokaryotic genome annotation." *Bioinformatics* 30.14, pp. 2068–2069. DOI: 10.1093/bioinformatics/btu153.

Shepard, S. M. et al. (2012). "Genome sequences and phylogenetic analysis of K88- and F18-positive porcine enterotoxigenic *Escherichia coli*." *J. Bacteriol.* 194.2, pp. 395–405. DOI: 10.1128/JB.06225-11.

Shpigel, N. Y., S. Elazar, and I. Rosenshine (2008). "Mammary pathogenic *Escherichia coli*." *Curr. Opin. Microbiol.* 11.1, pp. 60–65. DOI: 10.1016/j.mib.2008.01.004.

Sievers, F. et al. (2011). "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega." *Mol. Syst. Biol.* 7, p. 539. DOI: 10.1038/msb.2011.75.

Simpson, J. T. et al. (2017). "Detecting DNA cytosine methylation using nanopore sequencing." *Nat. Meth.* 14.4, pp. 407–410. DOI: 10.1038/nmeth.4184.

Sims, G. E. and S.-H. Kim (2011). "Whole-genome phylogeny of *Escherichia coli*/*Shigella* group by feature frequency profiles (FFPs)." *Proc. Natl. Acad. Sci. U.S.A.* 108.20, pp. 8329–8334. DOI: 10.1073/pnas.1105168108.

Skerra, A. (1994). "Use of the tetracycline promoter for the tightly regulated production of a murine antibody fragment in *Escherichia coli*." *Gene* 151 (1-2), pp. 131–135.

Smajs, D. et al. (2010). "Bacteriocin synthesis in uropathogenic and commensal *Escherichia coli*: colicin E1 is a potential virulence factor." *BMC Microbiol.* 10, p. 288. DOI: 10.1186/1471-2180-10-288.

Smet, A. et al. (2010). "Complete nucleotide sequence of CTX-M-15-plasmids from clinical *Escherichia coli* isolates: insertional events of transposons and insertion sequences." *PLoS ONE* 5.6, e11202. DOI: 10.1371/journal.pone.0011202.

Smith, D. H. (1967). "R factors mediate resistance to mercury, nickel, and cobalt." *Science* 156.3778, pp. 1114–1116. DOI: 10.1126/science.156.3778.1114.

Söderblom, T. et al. (2002). "Toxin-induced calcium oscillations: a novel strategy to affect gene regulation in target cells." *Int. J. Med. Microbiol.* 291 (6-7), pp. 511–515.

Sordillo, L. M. and K. L. Streicher (2002). "Mammary gland immunity and mastitis susceptibility." *J Mammary Gland Biol Neoplasia* 7.2, pp. 135–146. DOI: 10.1023/A:1020347818725.

Staden, R., K. F. Beal, and J. K. Bonfield (2000). "The Staden package, 1998." *Methods Mol. Biol.* 132, pp. 115–130. DOI: 10.1385/1-59259-192-2:115.

Stahlhut, S. G. et al. (2009). "Comparative structure-function analysis of mannose-specific FimH adhesins from *Klebsiella pneumoniae* and *Escherichia coli*." *J. Bacteriol.* 191.21, pp. 6592–6601. DOI: 10.1128/JB.00786-09.

Stajich, J. E. et al. (2002). "The Bioperl toolkit: Perl modules for the life sciences." *Genome Res.* 12.10, pp. 1611–1618. DOI: 10.1101/gr.361602.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." *Bioinformatics* 22.21, pp. 2688–2690. DOI: 10.1093/bioinformatics/btl446.

Stamatakis, A. (2014). "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics* 30.9, pp. 1312–1313. DOI: 10.1093/bioinformatics/btu033.

Stamatakis, A. and M. Ott (2008). "Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures." *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 363.1512, pp. 3977–3984. DOI: 10.1098/rstb.2008.0163.

Steinberg, K. M. and B. R. Levin (2007). "Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage." *Proc. Biol. Sci.* 274.1621, pp. 1921–1929. DOI: 10.1098/rspb.2007.0245.

Stenutz, R., A. Weintraub, and G. Widmalm (2006). "The structures of *Escherichia coli* O-polysaccharide antigens." *FEMS Microbiol. Rev.* 30.3, pp. 382–403. DOI: 10.1111/j.1574-6976.2006.00016.x.

Stetinova, V. et al. (2010). "Caco-2 cell monolayer integrity and effect of probiotic *Escherichia coli* Nissle 1917 components." *Neuro Endocrinol. Lett.* 31 Suppl 2, pp. 51–56.

Stevenson, G. et al. (1994). "Structure of the O antigen of *Escherichia coli* K-12 and the sequence of its *rfb* gene cluster." *J. Bacteriol.* 176.13, pp. 4144–4156. DOI: 10.1128/jb.176.13.4144-4156.1994.

Stoiber, M. H. et al. (2016). "*De novo* identification of DNA modifications enabled by genome-guided nanopore signal processing." *bioRxiv*, p. 094672. DOI: 10.1101/094672.

Storm, D. W. et al. (2011). "*In vitro* analysis of the bactericidal activity of *Escherichia coli* Nissle 1917 against pediatric uropathogens." *J. Urol.* 186 (4 Suppl), pp. 1678–1683. DOI: 10.1016/j.juro.2011.04.021.

Sullivan, M. J., N. K. Petty, and S. A. Beatson (2011). "Easyfig: a genome comparison visualizer." *Bioinformatics* 27.7, pp. 1009–1010. DOI: 10.1093/bioinformatics/btr039.

Summers, A. O. and S. Silver (1972). "Mercury resistance in a plasmid-bearing strain of *Escherichia coli*." *J. Bacteriol.* 112.3, pp. 1228–1236.

Sundén, F. et al. (2010). "*Escherichia coli* 83972 bacteriuria protects against recurrent lower urinary tract infections in patients with incomplete bladder emptying." *J. Urol.* 184.1, pp. 179–185. DOI: 10.1016/j.juro.2010.03.024.

Suojala, L., L. Kaartinen, and S. Pyörälä (2013). “Treatment for bovine *Escherichia coli* mastitis - an evidence-based approach.” *J. Vet. Pharmacol. Ther.* 36.6, pp. 521–531. DOI: 10.1111/jvp.12057.

Suojala, L. et al. (2011). “Phylogeny, virulence factors and antimicrobial susceptibility of *Escherichia coli* isolated in clinical bovine mastitis.” *Vet. Microbiol.* 147 (3-4), pp. 383–388. DOI: 10.1016/j.vetmic.2010.07.011.

Sváb, D. et al. (2013a). “Draft genome sequence of an *Escherichia coli* O157:H43 strain isolated from cattle.” *Genome Announc.* 1.3, e00263–13. DOI: 10.1128/genomeA.00263-13.

Sváb, D. et al. (2013b). “The long polar fimbriae operon and its flanking regions in bovine *Escherichia coli* O157:H43 and STEC O136:H12 strains.” *Pathog. Dis.* 68.1, pp. 1–7. DOI: 10.1111/2049-632X.12038.

Tamura, K. and M. Nei (1993). “Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees.” *Mol. Biol. Evol.* 10.3, pp. 512–526. DOI: 10.1093/oxfordjournals.molbev.a040023.

Tamura, K. et al. (2011). “MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.” *Mol. Biol. Evol.* 28.10, pp. 2731–2739. DOI: 10.1093/molbev/msr121.

Tarr, P. I., C. A. Gordon, and W. L. Chandler (2005). “Shiga-toxin-producing *Escherichia coli* and haemolytic uraemic syndrome.” *Lancet* 365.9464, pp. 1073–1086. DOI: 10.1016/S0140-6736(05)71144-2.

Tatusov, R. L. et al. (2003). “The COG database: an updated version includes eukaryotes.” *BMC Bioinformatics* 4, p. 41. DOI: 10.1186/1471-2105-4-41.

Taylor, D. E. et al. (2002). “Genomic variability of O islands encoding tellurite resistance in enterohemorrhagic *Escherichia coli* O157:H7 isolates.” *J. Bacteriol.* 184.17, pp. 4690–4698. DOI: 10.1128/JB.184.17.4690-4698.2002.

Tech, M. and R. Merkl (2003). “YACOP: enhanced gene prediction obtained by a combination of existing methods.” *In Silico Biol.* 3.4, pp. 441–451.

Tenaillon, O. et al. (2010). “The population genetics of commensal *Escherichia coli*.” *Nat. Rev. Microbiol.* 8.3, pp. 207–217. DOI: 10.1038/nrmicro2298.

Tettelin, H. et al. (2005). “Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”.” *Proc. Natl. Acad. Sci. U.S.A.* 102.39, pp. 13950–13955. DOI: 10.1073/pnas.0506758102.

Tettelin, H. et al. (2008). “Comparative genomics: the bacterial pan-genome.” *Curr. Opin. Microbiol.* 11.5, pp. 472–477. DOI: 10.1016/j.mib.2008.09.006.

Tietze, E. et al. (2015). “Comparative genomic analysis of two novel sporadic Shiga toxin-producing *Escherichia coli* O104:H4 strains isolated 2011 in Germany.” *PLoS ONE* 10.4, e0122074. DOI: 10.1371/journal.pone.0122074.

Toh, H. et al. (2010). “Complete genome sequence of the wild-type commensal *Escherichia coli* strain SE15, belonging to phylogenetic group B2.” *J. Bacteriol.* 192.4, pp. 1165–1166. DOI: 10.1128/JB.01543-09.

Torres, A. G. et al. (2009). “Genes related to long polar fimbriae of pathogenic *Escherichia coli* strains as reliable markers to identify virulent isolates.” *J. Clin. Microbiol.* 47.8, pp. 2442–2451. DOI: 10.1128/JCM.00566-09.

Totsika, M. et al. (2012). “Molecular characterization of the EhaG and UpaG trimeric autotransporter proteins from pathogenic *Escherichia coli*.” *Appl. Environ. Microbiol.* 78.7, pp. 2179–2189. DOI: 10.1128/AEM.06680-11.

Touchon, M. et al. (2009). “Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths.” *PLoS Genet.* 5.1, e1000344. DOI: 10.1371/journal.pgen.1000344.

Tourret, J. and E. Denamur (2016). “Population phylogenomics of extraintestinal pathogenic *Escherichia coli*.” *Microbiol. Spectr.* 4.1, UTI–0010–2012. DOI: 10.1128/microbiolspec.UTI-0010-2012.

Toval, F. et al. (2014a). “Characterization of *Escherichia coli* isolates from hospital inpatients or outpatients with urinary tract infection.” *J. Clin. Microbiol.* 52.2, pp. 407–418. DOI: 10.1128/JCM.02069-13.

Toval, F. et al. (2014b). "Characterization of urinary tract infection-associated Shiga toxin-producing *Escherichia coli*." *Infect. Immun.* 82.11, pp. 4631–4642. DOI: 10.1128/IAI.01701-14.

Treangen, T. J. and S. L. Salzberg (2011). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." *Nat. Rev. Genet.* 13.1, pp. 36–46. DOI: 10.1038/nrg3117.

Uhlén, P. et al. (2000). "Alpha-haemolysin of uropathogenic *E. coli* induces Ca2+ oscillations in renal epithelial cells." *Nature* 405.6787, pp. 694–697. DOI: 10.1038/35015091.

Ukena, S. N. et al. (2005). "The host response to the probiotic *Escherichia coli* strain Nissle 1917: specific up-regulation of the proinflammatory chemokine MCP-1." *BMC Med. Genet.* 6, p. 43. DOI: 10.1186/1471-2350-6-43.

Ukena, S. N. et al. (2007). "Probiotic *Escherichia coli* Nissle 1917 inhibits leaky gut by enhancing mucosal integrity." *PLoS ONE* 2.12, e1308. DOI: 10.1371/journal.pone.0001308.

Ulett, G. C., R. I. Webb, and M. A. Schembri (2006). "Antigen-43-mediated autoaggregation impairs motility in *Escherichia coli*." *Microbiology (Reading, Engl.)* 152 (Pt 7), pp. 2101–2110. DOI: 10.1099/mic.0.28607-0.

Ulett, G. C. et al. (2007). "Functional analysis of antigen 43 in uropathogenic *Escherichia coli* reveals a role in long-term persistence in the urinary tract." *Infect. Immun.* 75.7, pp. 3233–3244. DOI: 10.1128/IAI.01952-06.

Ullrich, S. R. et al. (2015). "Permanent draft genome sequence of *Acidiphilium* sp. JA12-A1." *Stand. Genomic Sci.* 10, p. 56. DOI: 10.1186/s40793-015-0040-y.

UniProt Consortium (2014). "Activities at the Universal Protein Resource (UniProt)." *Nucleic Acids Res.* 42 (Database issue), pp. D191–198. DOI: 10.1093/nar/gkt1140.

Urbina, F. et al. (2005). "Structural elucidation of the O-antigenic polysaccharide from the enteroaggregative *Escherichia coli* strain 180/C3 and its immunochemical relationship with *E. coli* O5 and O65." *Carbohydr. Res.* 340.4, pp. 645–650. DOI: 10.1016/j.carres.2005.01.001.

Valle, J. et al. (2008). "UpaG, a new member of the trimeric autotransporter family of adhesins in uropathogenic *Escherichia coli*." *J. Bacteriol.* 190.12, pp. 4147–4161. DOI: 10.1128/JB.00122-08.

Vallenet, D. et al. (2017). "MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes." *Nucleic Acids Res.* 45 (D1), pp. D517–D528. DOI: 10.1093/nar/gkw1101.

Van der Woude, M. W. and I. R. Henderson (2008). "Regulation and function of Ag43 (*flu*)." *Annu. Rev. Microbiol.* 62, pp. 153–169. DOI: 10.1146/annurev.micro.62.081307.162938.

Van Helden, P. (2013). "Data-driven hypotheses." *EMBO Rep.* 14.2, p. 104. DOI: 10.1038/embor.2012.207.

Van Opijnen, T. and A. Camilli (2013). "Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms." *Nat. Rev. Microbiol.* 11.7, pp. 435–442. DOI: 10.1038/nrmicro3033.

Vangroenweghe, F. et al. (2004). "Increase of *Escherichia coli* inoculum doses induces faster innate immune response in primiparous cows." *J. Dairy Sci.* 87.12, pp. 4132–4144. DOI: 10.3168/jds.S0022-0302(04)73556-0.

Vejborg, R. M. et al. (2010). "A virulent parent with probiotic progeny: comparative genomics of *Escherichia coli* strains CFT073, Nissle 1917 and ABU 83972." *Mol. Genet. Genomics* 283.5, pp. 469–484. DOI: 10.1007/s00438-010-0532-9.

Vieira, G. et al. (2011). "Core and panmetabolism in *Escherichia coli*." *J. Bacteriol.* 193.6, pp. 1461–1472. DOI: 10.1128/JB.01192-10.

Vinogradov, E. V. et al. (1999). "The structures of the carbohydrate backbones of the lipopolysaccharides from *Escherichia coli* rough mutants F470 (R1 core type) and F576 (R2 core type)." *Eur. J. Biochem.* 261.3, pp. 629–639. DOI: 10.1046/j.1432-1327.1999.00280.x.

Von Buenau, R. et al. (2005). "*Escherichia coli* strain Nissle 1917: significant reduction of neonatal calf diarrhea." *J. Dairy Sci.* 88.1, pp. 317–323. DOI: 10.3168/jds.S0022-0302(05)72690-4.

Von Mentzer, A. et al. (2014). "Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution." *Nat. Genet.* 46.12, pp. 1321–1326. DOI: 10.1038/ng.3145.

Waack, S. et al. (2006). "Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models." *BMC Bioinformatics* 7, p. 142. DOI: 10.1186/1471-2105-7-142.

Walk, S. T. et al. (2009). "Cryptic lineages of the genus *Escherichia*." *Appl. Environ. Microbiol.* 75.20, pp. 6534–6544. DOI: 10.1128/AEM.01262-09.

Wang, L. et al. (2005). "Molecular markers for detection of pathogenic *Escherichia coli* strains belonging to serogroups O 138 and O 139." *Vet. Microbiol.* 111 (3-4), pp. 181–190. DOI: 10.1016/j.vetmic.2005.10.006.

Wang, Q. et al. (2009). "Genetic and structural analyses of *Escherichia coli* O107 and O117 O-antigens." *FEMS Immunol. Med. Microbiol.* 55.1, pp. 47–54. DOI: 10.1111/j.1574-695X.2008.00494.x.

Wang, X. et al. (2006). "Impact of biofilm matrix components on interaction of commensal *Escherichia coli* with the gastrointestinal cell line HT-29." *Cell. Mol. Life Sci.* 63 (19-20), pp. 2352–2363. DOI: 10.1007/s00018-006-6222-4.

Ward, N. and G. Moreno-Hagelsieb (2014). "Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss?" *PLoS ONE* 9.7, e101850. DOI: 10.1371/journal.pone.0101850.

Warnes, G. R. et al. (2016). *gplots: various R programming tools for plotting data*. R package version 3.0.1. URL: https://CRAN.R-project.org/package=gplots.

Wattam, A. R. et al. (2017). "Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center." *Nucleic Acids Res.* 45 (D1), pp. D535–D542. DOI: 10.1093/nar/gkw1017.

Welch, R. A. et al. (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*." *Proc. Natl. Acad. Sci. U.S.A.* 99.26, pp. 17020–17024. DOI: 10.1073/pnas.252529799.

Wellnitz, O. et al. (2006). "Immune relevant gene expression of mammary epithelial cells and their influence on leukocyte chemotaxis in response to different mastitis pathogens." *Vet. Med.* 51.4. WOS:000237607100001, pp. 125–132.

Wells, T. J. et al. (2008). "EhaA is a novel autotransporter protein of enterohemorrhagic *Escherichia coli* O157:H7 that contributes to adhesion and biofilm formation." *Environ. Microbiol.* 10.3, pp. 589–604. DOI: 10.1111/j.1462-2920.2007.01479.x.

Wells, T. J. et al. (2009). "The *Escherichia coli* O157:H7 EhaB autotransporter protein binds to laminin and collagen I and induces a serum IgA response in O157:H7 challenged cattle." *Environ. Microbiol.* 11.7, pp. 1803–1814. DOI: 10.1111/j.1462-2920.2009.01905.x.

Wells, T. J., M. Totsika, and M. A. Schembri (2010). "Autotransporters of *Escherichia coli*: a sequence-based characterization." *Microbiology (Reading, Engl.)* 156 (Pt 8), pp. 2459–2469. DOI: 10.1099/mic.0.039024-0.

Wenz, J. R. et al. (2006). "*Escherichia coli* isolates' serotypes, genotypes, and virulence genes and clinical coliform mastitis severity." *J. Dairy Sci.* 89.9, pp. 3408–3412. DOI: 10.3168/jds.S0022-0302(06)72377-3.

Westermann, A. J., S. A. Gorski, and J. Vogel (2012). "Dual RNA-seq of pathogen and host." *Nat. Rev. Microbiol.* 10.9, pp. 618–630. DOI: 10.1038/nrmicro2852.

Westermann, A. J. et al. (2016). "Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions." *Nature* 529.7587, pp. 496–501. DOI: 10.1038/nature16547.

Westphal, O and K. Jann (1965). "Bacterial lipopolysaccharides: extraction with phenol-water and further application of the procedure." *Met. Carbohydr. Chem.* 5, pp. 83–89.

Whelan, K. F., E. Colleran, and D. E. Taylor (1995). "Phage inhibition, colicin resistance, and tellurite resistance are encoded by a single cluster of genes on the IncHI2 plasmid R478." *J. Bacteriol.* 177.17, pp. 5016–5027. DOI: 10.1128/jb.177.17.5016-5027.1995.

Whiteside, M. D. et al. (2016). "SuperPhy: predictive genomics for the bacterial pathogen *Escherichia coli*." *BMC Microbiol.* 16.1, p. 65. DOI: 10.1186/s12866-016-0680-0.

Whitfield, C. and I. S. Roberts (1999). "Structure, assembly and regulation of expression of capsules in *Escherichia coli*." *Mol. Microbiol.* 31.5, pp. 1307–1319. DOI: 10.1046/j.1365-2958.1999.01276.x.

Whittam, T. S. et al. (1993). "Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea." *Infect. Immun.* 61.5, pp. 1619–1629.

WHO (2014). *Antimicrobial resistance: global report on surveillance 2014*. URL: http://www.who.int/drugresistance/documents/surveillancereport/en/.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer New York. DOI: 10.1007/978-0-387-98141-3. URL: http://link.springer.com/10.1007/978-0-387-98141-3.

Wiedenbeck, J. and F. M. Cohan (2011). "Origins of bacterial diversity through horizontal genetic transfer and adaptation to new ecological niches." *FEMS Microbiol. Rev.* 35.5, pp. 957–976. DOI: 10.1111/j.1574-6976.2011.00292.x.

Wildeman, P. et al. (2016). "*Propionibacterium avidum* as an etiological agent of prosthetic hip joint infection." *PLoS ONE* 11.6, e0158164. DOI: 10.1371/journal.pone.0158164.

Wilkinson, M. D. et al. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Sci. Data* 3, p. 160018. DOI: 10.1038/sdata.2016.18.

Wilson, D. J. (2012). "Insights from genomics into bacterial pathogen populations." *PLoS Pathog.* 8.9, e1002874. DOI: 10.1371/journal.ppat.1002874.

Wirth, T. et al. (2006). "Sex and virulence in *Escherichia coli*: an evolutionary perspective." *Mol. Microbiol.* 60.5, pp. 1136–1151. DOI: 10.1111/j.1365-2958.2006.05172.x.

Wold, A. E. et al. (1992). "Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics." *J. Infect. Dis.* 165.1, pp. 46–52.

Woodford, N. et al. (2009). "Complete nucleotide sequences of plasmids pEK204, pEK499, and pEK516, encoding CTX-M enzymes in three major *Escherichia coli* lineages from the United Kingdom, all belonging to the international O25:H4-ST131 clone." *Antimicrob. Agents Chemother.* 53.10, pp. 4472–4482. DOI: 10.1128/AAC.00688-09.

Wurpel, D. J. et al. (2013). "Chaperone-usher fimbriae of *Escherichia coli*." *PLoS ONE* 8.1, e52835. DOI: 10.1371/journal.pone.0052835.

Xiang, S. H., M. Hobbs, and P. R. Reeves (1994). "Molecular analysis of the *rfb* gene cluster of a group D2 *Salmonella enterica* strain: evidence for its origin from an insertion sequence-mediated recombination event between group E and D1 strains." *J. Bacteriol.* 176.14, pp. 4357–4365. DOI: 10.1128/jb.176.14.4357-4365.1994.

Yang, J. et al. (2008a). "VFDB 2008 release: an enhanced web-based resource for comparative pathogenomics." *Nucleic Acids Res.* 36 (Database issue), pp. D539–542. DOI: 10.1093/nar/gkm951.

Yang, W. et al. (2008b). "Bovine TLR2 and TLR4 properly transduce signals from *Staphylococcus aureus* and *E. coli*, but *S. aureus* fails to both activate NF-kappaB in mammary epithelial cells and to quickly induce TNFalpha and interleukin-8 (CXCL8) expression in the udder." *Mol. Immunol.* 45.5, pp. 1385–1397. DOI: 10.1016/j.molimm.2007.09.004.

Yao, Y. et al. (2009). "The type III secretion system is involved in the invasion and intracellular survival of *Escherichia coli* K1 in human brain microvascular endothelial cells." *FEMS Microbiol. Lett.* 300.1, pp. 18–24. DOI: 10.1111/j.1574-6968.2009.01763.x.

Yeo, H.-J. et al. (2004). "Structural basis for host recognition by the *Haemophilus influenzae* Hia autotransporter." *EMBO J.* 23.6, pp. 1245–1256. DOI: 10.1038/sj.emboj.7600142.

Yi, H. et al. (2011). "Genome sequence of *Escherichia coli* AA86, isolated from cow feces." *J. Bacteriol.* 193.14, p. 3681. DOI: 10.1128/JB.05193-11.

Younis, S., Q. Javed, and M. Blumenberg (2016). "Meta-analysis of transcriptional responses to mastitis-causing *Escherichia coli*." *PLoS ONE* 11.3, e0148562. DOI: 10.1371/journal.pone.0148562.

Zadik, P. M., P. A. Chapman, and C. A. Siddons (1993). "Use of tellurite for the selection of verocytotoxigenic *Escherichia coli* O157." *J. Med. Microbiol.* 39.2, pp. 155–158. DOI: 10.1099/00222615-39-2-155.

Zadoks, R. N. et al. (2011). "Molecular epidemiology of mastitis pathogens of dairy cattle and comparative relevance to humans." *J. Mammary Gland. Biol. Neoplasia* 16.4, pp. 357–372. DOI: 10.1007/s10911-011-9236-y.

Zankari, E. et al. (2012). "Identification of acquired antimicrobial resistance genes." *J. Antimicrob. Chemother.* 67.11, pp. 2640–2644. DOI: 10.1093/jac/dks261.

Zarrouk, H. et al. (1997). "Use of mass spectrometry to compare three O-chain-linked and free lipopolysaccharide cores: differences found in *Bordetella parapertussis*." *J. Endotoxin Res.* 4.6, pp. 453–458. DOI: 10.1177/096805199700400609.

Zdziarski, J. et al. (2008). "Molecular basis of commensalism in the urinary tract: low virulence or virulence attenuation?" *Infect. Immun.* 76.2, pp. 695–703. DOI: 10.1128/IAI.01215-07.

Zdziarski, J. et al. (2010). "Host imprints on bacterial genomes–rapid, divergent evolution in individual patients." *PLoS Pathog.* 6.8, e1001078. DOI: 10.1371/journal.ppat.1001078.

Zhang, W. et al. (2012). "Real-time multiplex PCR for detecting Shiga toxin 2-producing *Escherichia coli* O104:H4 in human stools." *J. Clin. Microbiol.* 50.5, pp. 1752–1754. DOI: 10.1128/JCM.06817-11.

Zhou, Y. et al. (2011). "PHAST: a fast phage search tool." *Nucleic Acids Res.* 39 (Web Server issue), W347–352. DOI: 10.1093/nar/gkr485.

Zhou, Z. et al. (2010). "Derivation of *Escherichia coli* O157:H7 from its O55:H7 precursor." *PLoS ONE* 5.1, e8700. DOI: 10.1371/journal.pone.0008700.

Ziebell, K. et al. (2011). "Gene cluster conferring streptomycin, sulfonamide, and tetracycline resistance in *Escherichia coli* O157:H7 phage types 23, 45, and 67." *Appl. Environ. Microbiol.* 77.5, pp. 1900–1903. DOI: 10.1128/AEM.01934-10.

Zude, I., A. Leimbach, and U. Dobrindt (2014). "Prevalence of autotransporters in *Escherichia coli*: what is the impact of phylogeny and pathotype?" *Int. J. Med. Microbiol.* 304 (3-4), pp. 243–256. DOI: 10.1016/j.ijmm.2013.10.006.

AFFIDAVIT

I hereby confirm that my thesis entitled "*Genomics of pathogenic and commensal* Escherichia coli" is the result of my own work. I did not receive any help or support from commercial consultants. All sources and/or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Würzburg, July 2017

Andreas Leimbach

EIDESSTATTLICHE ERKLÄRUNG

Hiermit erkläre ich an Eides statt, die Dissertation "*Genomik pathogener und kommensaler* Escherichia coli" eigenständig, d.h. insbesondere selbständig und ohne Hilfe eines kommerziellen Promotionsberaters, angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Dissertation weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Würzburg, July 2017

Andreas Leimbach

COLOPHON

This thesis was typeset in LaTeX 2$_{\varepsilon}$ using pdfLaTeX (v3.14159265-2.6-1.40.16), `TeXstudio` (v2.10.8), and the typographical `classicthesis` (v4.2) style developed by André Miede. Hermann Zapf's fantastic *Palatino* (`newpxtext`) and *Euler* (`eulervm`) type faces were employed. The bibliography was typeset using `biblatex` (v3.7).