Determination of virulence and fitness genes associated with the \textit{pheU}, \textit{pheV} and \textit{selC} integration sites of LEE-negative food-borne Shiga toxin-producing \textit{Escherichia coli} strains

Nadja Saile$^{1\dagger}$, Elisabeth Schuh$^{1,2\dagger}$, Torsten Semmler$^3$, Inga Eichhorn$^4$, Lothar H. Wieler$^3$, Andreas Bauwens$^5$ and Herbert Schmidt$^1$*

Abstract

Background: In the current study, nine foodborne “Locus of Enterocyte Effacement” (LEE)-negative Shiga toxin-producing \textit{Escherichia coli} (STEC) strains were selected for whole genome sequencing and analysis for yet unknown genetic elements within the already known LEE integration sites \textit{selC}, \textit{pheU} and \textit{pheV}. Foreign DNA ranging in size from 3.4 to 57 kbp was detected and further analyzed. Five STEC strains contained an insertion of foreign DNA adjacent to the \textit{selC} tRNA gene and five and seven strains contained foreign DNA adjacent to the \textit{pheU} and \textit{pheV} tRNA genes, respectively. We characterized the foreign DNA insertion associated with \textit{selC} (STEC O91:H21 strain 17584/1), \textit{pheU} (STEC O8:H4 strain RF1a and O55:Hnt strain K30) and \textit{pheV} (STEC O91:H21 strain 17584/1 and O113:H21 strain TS18/08) as examples.

Results: In total, 293 open reading frames partially encoding putative virulence factors such as TonB-dependent receptors, DNA helicases, a hemolysin activator protein precursor, antigen 43, anti-restriction protein KlcA, ShiA, and phosphoethanolamine transferases were detected. A virulence type IV toxin-antitoxin system was detected in three strains. Additionally, the \textit{ato} system was found in one strain. In strain 17584/1 we were able to define a new genomic island which we designated \textit{GI}_{selC}^{17584/1}. The island contained integrases and mobile elements in addition to genes for increased fitness and those playing a putative role in pathogenicity.

Conclusion: The data presented highlight the important role of the three tRNAs \textit{selC}, \textit{pheU}, and \textit{pheV} for the genomic flexibility of \textit{E. coli}.

Keywords: STEC, Food, \textit{pheU}, \textit{pheV}, \textit{selC} integration site, Genomic island

Background

The genome sequence of the \textit{Escherichia coli} type strain (U5/41T) was recently published and revealed a size of about 5 Mbp, containing 4762 protein-coding genes [1]. In comparison to the non-pathogenic \textit{E. coli} strains with genome sizes of 4.5–5 Mbp, pathogenic \textit{E. coli} frequently carry additional DNA and reach genome sizes up to 5.9 Mbp [2, 3]. This additional DNA is often located on plasmids, prophages or genomic islands, all of which might be acquired by horizontal gene transfer [4]. Genomic islands (GEI) are distinct DNA regions that are usually larger than 30 kbp, and their GC-content differs from the host genome. They are often flanked by insertion sequences or direct repeats. When additional pathogenicity-associated genes are present, they are also referred to as pathogenicity islands (PAIs). PAIs are found exclusively in pathogenic members of a species, contain multiple (cryptic) mobile genetic elements, have

*Correspondence: Herbert.Schmidt@uni-hohenheim.de
†Nadja Saile and Elisabeth Schuh contributed equally to this work
$^1$ Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstr. 28, 70599 Stuttgart, Germany
Full list of author information is available at the end of the article

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
fitness or virulence factors and are often associated with tRNA loci which function as integration sites for additional DNA. Such PAI are summarized as “foreign DNA” in the current study [5, 6]. Comparative genomics studies of a group of Shiga toxin-producing E. coli (STEC), the so-called enterohemorrhagic E. coli (EHEC), were performed by Ogura and colleagues and revealed that EHEC possess more tRNA genes than other E. coli and Shigella strains. Additionally, they showed that tRNA genes are often the target of DNA insertions, often as insertion loci for two or three genetic elements at once [3, 7]. The best described PAI in EHEC which is inserted adjacent to a tRNA gene is the “Locus of Enterocyte Effacement” (LEE) with a basic size of about 35 kbp. The genes of this pathogenicity island are often close to the tRNA genes selC, pheU or pheV [8]. The LEE is an important virulence factor, and its gene products support the intimate binding of the bacterium to the host cell and the release of several effector proteins into the host cell cytoplasm [8–10]. Besides the LEE, researchers have also described other PAIs as being integrated into the selC locus. These include PAI-1 in the uropathogenic E. coli (UPEC) strain 536 and SPI-3 in Salmonella enterica as well as the 33 kbp locus of proteolysis activity in the STEC strain 4797/97 and the toxigenic invasion locus A in the enterotoxigenic E. coli (ETEC) strain H10407 [11–14]. Others have also found an attachment site of the E. coli retrophage ΦR73 integrated into selC [15]. For uropathogenic E. coli strains, the percentage of genomic islands is nearly 13% of the genome, demonstrating the importance of horizontally acquired DNA [16]. Strain CFT073, for example, was shown to harbour 13 genomic islands larger than 30 kbp. One of them is integrated in selC with a size of 68 kbp and is named intC-c4581 [16].

Both integration sites pheU and pheV have an identical sequence but differ in their gene surroundings. For pheU and pheV, different variants of the LEE locus with sizes between 36 and 111 kbp have been described [8].

One of these is the hybrid PAI ICL3, which contains the LEE core, parts of two different genomic islands detected in EHEC strain EDL933 (O1-48 and O1-122) and DNA homologous to Yersinia pestis. The PAI ICL3 is integrated in pheV of Citrobacter rodentium and in STEC [17, 18]. Another PAI originally found in pheU but more often detected in pheV was described as including an adhesin encoded by an afa-8 gene cluster which was found in human and bovine E. coli isolates [19]. Uropathogenic E. coli (UPEC) and Shigella strains were also shown to harbour genomic islands at these integration sites. In UPEC J96, the PAI VJ96 with 110 kbp is located adjacent to pheU and the PAI IVJ96 with more than 170 kbp was detected at the pheV integration site [20]. Shigella boydii owns an iron transport-associated PAI of 21 kbp named SHI-3 in pheU and Shigella flexneri 2a a PAI of 46.6 kbp in pheV [21, 22].

STEC are important foodborne pathogens with more than 400 described serotypes and a high diversity of isolates from contaminated food [23]. Therefore, the aim of the study was to investigate whether LEE-negative foodborne STEC strains harbour foreign DNA at the LEE integration sites and whether this DNA may contribute to the fitness and pathogenicity of these strains. Nine strains with at least one occupied integration site selC, pheU or pheV were selected and subjected to whole genome sequencing and characterization of their foreign DNA.

Methods

Bacterial strains and culture conditions

The foodborne STEC strains TS18/08, LM27558, RF1a, TS25/08, LM27564, LM14603/08, K30, TS21/08 and 17584/1 have been isolated from risk foods (Table 1) and were selected because they were LEE-negative and a former study had shown that at least one of the tRNA sites selC, pheU or pheV was occupied by additional DNA [24, 25]. The strains were cultured overnight in LB broth (1% (w/v) tryptone, 0.5% (w/v) yeast extract, 1% (w/v) NaCl, pH 7) at 37 °C with agitation at 180 rpm. DNA was isolated using a QiaGen Blood and Tissue Kit following the manufacturer’s instructions (Qiagen, Hilden, Germany). DNA concentration and purity were measured using a NanoDrop 2000 device (Thermo Fisher Scientific, Schwerte, Germany).

Whole genome sequencing and sequence analysis

The concentration of the purified DNA was evaluated using the Qubit dsDNA HS Assay (Life Technologies, Darmstadt, Germany). MiSeq libraries containing 1 ng of DNA were prepared with Nextera XT chemistry (Illumina, San Diego, CA, USA) and were sequenced in a paired-end run (2 × 300 bp) on an Illumina MiSeq sequencer as recommended by the manufacturer with a minimum coverage of 90×. Raw data was de novo assembled using CLC Genomics Workbench (http://www.qiagenbioinformatics.com) resulting in assemblies with N50 values between 79 and 224 kbp consisting of between 181 and 419 contigs. The draft genomes were further analysed using Geneious software ver. 7.1, 9.1.8 and 10.0.7 (http://www.geneious.com). Annotations were carried out using the RASTk annotation tool within the PATRIC web resources [26–28]. For comparisons, the Blastn and Blastx algorithms were used (http://blast.ncbi.nlm.nih.gov). The genome sequences of the investigated strains have been deposited in the NCBI database under the accession numbers given in Table 1. For integration site detection and description, selC, pheU
and pheV site-specific primers sequences, which were described before, were used for in silico analyses [24]. For strain LM14603/08 the pheU and pheV integration sites were confirmed for DNA insertion by PCR, as recently described [24].

Results
In this study, nine LEE-negative foodborne STEC strains of different serotypes (Table 1) were investigated by whole genome sequencing to gain further insight into the identity of additional DNA at LEE integration sites. For each strain, at least one of the integration sites selC, pheU or pheV was occupied as previously analysed by PCR. Using the respective primers, strains without integration of additional DNA in those integration sites showed the expected sequences and amplicon sizes (selC locus 2173 bp, pheU locus 664 bp, pheV locus 1306 bp).

General characteristics of LEE-negative food-borne STEC
Whole genome sequencing was performed with nine STEC strains and the number of contigs achieved range from 181 to 419 among the analysed strains. More detailed information is given in Table 1.

In agreement with the PCR results, five of the nine strains contained additional DNA in selC, five strains within pheU and seven within pheV. Strain LM14603/08 was found to harbour additional DNA in pheU and pheV in contrast to previous PCR analyses. We analyzed the pheU and pheV loci in strain LM14603/08 by PCR, and did not obtain an amplicon for either site (see Additional file 1), indicating an integration of DNA close to pheU and pheV. An overview of the sizes of integrated DNA within the three sites is given in Table 2 and is shown schematically in Fig. 1. For the strains LM27558_stx2 and TS25/08, no contig with pheV was detected at the downstream site. Additionally, no contig spanning the complete pheU or pheV locus including additional DNA was obtained for any strain whereas one contig (contig 22) spanning the entire selC locus with additional DNA was obtained for only one strain (17584/1).

In regard to the achieved contig length, we received from the whole genome sequencing procedure we have chosen four strains, 17584/1, RF1a, K30, and TS18/08, and analysed the selC-, pheU- and/or pheV-located nucleotide sequences in more detail and identified the corresponding open reading frames. For strain 17584/1 we received one contig (contig 22) that includes DNA spanning the whole selC integration site and two contigs (contig 20 and 35) that include parts of the DNA.

| Strain | selC | pheU | pheV |
|--------|------|------|------|
| TS18/08 | None | None | >57,010 bp |
| LM27558_stx2 | >5002 bp | >8453 bp | >3395 bp |
| RF1a | >13,763 bp | >34,112 bp | None |
| TS25/08 | None | None | >19,489 bp |
| LM27564 | None | >3797 bp | >3686 bp |
| LM14603/08 | >25,398 | >6437 bp | >6424 bp |
| K30 | None | >26,836 bp | None |
| TS21/08 | >32,214 bp | None | >13,402 bp |
| 17584/1 | 50,646 bp | None | >54,896 bp |

Table 1 General characteristics of LEE-negative foodborne STEC used for sequence analysis and the respective accession numbers

| Strain | Serotype | Source | Predicted occupied integration sites | No. of contigs | N50 value (kbp) | Calculated genome size (Mbp) | Gen Bank Accession number |
|--------|----------|--------|-------------------------------------|----------------|----------------|-------------------------------|-------------------------|
| TS18/08 | O113:H21 | Minced meat | pheV | 226 | 150 | 5.1 | MPTX00000000 |
| LM27558_stx2 | Orough:H43 | Deer meat | selC, pheU, pheV | 410 | 111 | 5.7 | MPTY00000000 |
| RF1a | O8:H6 | beef | selC, pheU | 233 | 173 | 5.2 | MPTZ00000000 |
| TS25/08 | Ont:Hnt | Minced meat | pheV | 419 | 79 | 5.2 | MPUA00000000 |
| LM27564 | O113:Hnm | Deer meat | pheU, pheV | 413 | 101 | 5.4 | MPUC00000000 |
| LM14603/08 | O21:H21 | Deer meat | selC, pheU, pheV | 271 | 169 | 5.4 | MPUD00000000 |
| K30 | O55:Hnt | Raw milk | pheU | 181 | 224 | 5.0 | MPUE00000000 |
| TS21/08 | O113:H21 | Minced meat | selC, pheV | 297 | 177 | 5.2 | MPUF00000000 |
| 17584/1 | O91:H21 | Mettwurst | selC, pheV | 341 | 140 | 5.0 | MPUG00000000 |

nt not typeable, nm non motile
a Data from Slanec et al. [25]
b Data from Hauser et al. [24]
integrated within *pheV*. For the strains RF1a and K30, we identified two contig each (contig 2 and 30, and contig 16 and 22, respectively) that include parts of the DNA inserted within *pheU*. In addition, two contigs (contig 18 and 39) were also found for the strain TS18/08 including DNA inserted within *pheV*.

**Borders of the nucleotide sequences integrated in tRNA genes**

The analysed nucleotide sequences of the strains 17584/1, RF1a, K30 and TS18/08 are integrated downstream of the tRNA genes *pheU*, *pheV* or *selC* and were compared to the sequences of the integration sites of *E. coli* K-12 sub-str. MG1655 (Acc. No. NC_000913) and EDL933 (Acc. No. NZ_CP008957) (Fig. 2). The tRNA genes *pheU*, *pheV* and *selC* are located near the S′ termini of the nucleotide sequences adjacent to *pheU*, *pheV* and *selC* tRNA genes. The characters in the boxes are parts of the annotated tRNA genes. Colored characters indicate identities between the nucleotide sequences of the strains.
and selC of the analysed strains are similar to the genes of E. coli K-12. Differences occur downstream of the tRNA genes especially for the selC and with a greater extent for the pheV associated insertions.

The insertions stop on the 3’ terminus with a direct repeat (24 bp for selC, 22 bp for pheU and pheV) that is part of the corresponding annotated tRNA gene of E. coli K-12. For the direct repeats of pheU and pheV, a reading frame shift in the analysed sequences was recorded when compared to the E. coli K-12 sequences.

Structural characterization of inserted DNA in the integration sites pheU, pheV or selC

Escherichia coli O113:H21 strain TS18/08 expresses the toxins SubAB, Cdt-V and Stx [25, 29]. The screening of the LEE-typical insertion sites merely indicates an occupation of the pheV-associated insertion site only. Even if we did not receive a contig spanning the whole insertion sequence, we identified two contigs that include the 3’ terminus of pheV (38.9 kbp) encoding 45 CDS and the 5’ terminus of the pheV corresponding with a direct repeat (16.9 kbp) with 33 CDS (CDS 46–78) (Additional file 1: Figure S2). The CDS have a length between 114 and 9201 bp. With RASTk analysis, 48 CDS were designated as hypothetical proteins. By blastx analyses 34 of the 48 CDS could be confirmed as hypothetical proteins but 14 were characterized as common proteins mostly containing mobile elements (Additional file 1: Table S1).

The second STEC strain that was analysed in detail is RF1a of the serotype O8:H6, which has insertions near selC and pheU. Since the contig lengths spanning parts covering the selC associated nucleotide sequence were low, we only analysed the pheU-affiliated parts of contigs 2 (7.3 kbp) and 30 (26.2 kbp) (Additional file 1: Figure S3). The pheU 3’ terminus of contig 2 contains four CDS and the downstream pheU direct repeat 5’ terminus of contig 30 contains 40 CDS (CDS 5–44) ranging from 120 to 3516 bp (Additional file 1: Table S2). After blastx analyses, 10 CDS are of unknown function, while the remaining CDS could be identified in terms of their encoded protein.

STEC O55:Hnt strain K30 contains foreign DNA only within the pheU gene with 20 CDS within the pheU-associated part of contig 16 (23.1 kbp) and six CDS within contig 22 (3.1 kbp) (Additional file 1: Figure S4). The length of the CDS varies between 114 and 5511 bp, and seven of the CDS were of unknown function (Additional file 1: Table S3).

For O91:H21 strain 17584/1 we could analyse both occupied integration sites pheV and selC. For pheV we found two contigs containing parts of the inserted DNA, contig 20 (32.1 kbp) and contig 35 (21.5 kbp) (Additional file 1: Figure S5), and for selC we identified contig 22, including the entire nucleotide sequence located in selC (48.5 kbp) of strain 17584/1 (Fig. 3). In contigs 20, 35 and 22 we detected 49, 35 and 60 CDS, respectively. The CDS lengths of the pheV-associated DNA ranges between 114 and 3120 bp (Additional file 1: Table S4) and for the selC associated DNA between 114 and 9780 bp (Additional file 1: Table S5) with 22 CDS of coding for hypothetical proteins.

Predicted proteins encoded by DNA insertions in pheU, pheV or selC

The predicted proteins belong to the ato system, the type IV toxin-antitoxin system or to the B12 uptake system. Moreover, hemolysins/hemagglutinins, transporters, proteins containing domains of unknown function, methyltransferases, regulators, other hypothetical proteins or proteins belonging to mobile genetic elements were detected (Table 3).
Table 3  Predicted proteins encoded by inserted DNA adjacent to pheV, pheU or selC

| Proteins                                                                 | TS18/08 pheV | RF1a pheU | K30 pheU | 17584/1 pheV | selC |
|--------------------------------------------------------------------------|--------------|-----------|----------|--------------|------|
| Mobile elements                                                          | x            | x         | x        | x            | x    |
| Transporters                                                             |              |           |          |              |      |
| SapA³                                                                   |              |           | x        |              | x    |
| ShiA-homolog³                                                            |              |           |          |              | x    |
| Per-activated serine protease autotransporter enterotoxin EspC          |              |           | x        |              |      |
| Outer membrane porin OmpF                                               | x            |           | x        |              |      |
| Antigen 43³                                                              | x            |           | x        |              |      |
| Ato system                                                               |              |           | x        |              |      |
| AtoS/AtoS₅                                                               |              |           |          |              |      |
| AtoC                                                                     |              |           |          |              |      |
| AtoD                                                                     |              |           |          |              |      |
| AtoA                                                                     |              |           |          |              |      |
| AtoE                                                                     |              |           |          |              |      |
| AtoB³                                                                    |              |           |          |              |      |
| Type IV toxin-antitoxin system                                           |              |           |          |              |      |
| YeeP/NgrB⁶                                                               | x            | x         |          | x            | x    |
| YeeR³                                                                    |              |           |          |              |      |
| YeeS³                                                                    | x            | x         |          | x            | x    |
| YeeT⁹                                                                    | x            | x         | x        | x            | x    |
| YeeU⁹                                                                    | x            | x         | x        |              |      |
| YeeV⁹                                                                    | x            | x         | x        |              |      |
| Domains of unknown function                                              |              |           |          |              |      |
| DUF3987 containing protein                                              |              |           | x        |              |      |
| DUF957 containing protein                                               |              |           | x        | x            | x    |
| DUF4338 containing protein                                              |              |           |          |              |      |
| DUF4222 containing protein                                              |              |           |          |              |      |
| DUF2251 containing protein                                              |              |           |          |              |      |
| DUF1705 containing protein                                              |              |           |          |              | x    |
| DUF2569 containing protein                                              |              |           |          |              | x    |
| Methyltransferases                                                       |              |           |          |              |      |
| Z1226⁴                                                                  | x            | x         | x        |              |      |
| DNA-cytosine methyltransferase                                          |              |           |          |              |      |
| B12 uptake system                                                        |              |           |          |              |      |
| TonB-dependent receptor⁸                                                | x            |           |          |              |      |
| Outer membrane vitamin B12 receptor BtuB                                 | x            |           |          |              |      |
| Hemolysin/hemagglutinin                                                  |              |           |          |              |      |
| Putative member of ShlA/HecA/FhaA exoprotein family                     |              |           |          |              |      |
| Putative large exoprotein involved in heme utilization or adhesion of    | x            |           |          |              |      |
| ShlA/HecA/FhaA family                                                   |              |           |          |              |      |
| Putative adhesin/hemagglutinin/hemolysin                                |              |           |          |              |      |
| RTX toxin activating lysine-acyltransferase                             |              |           | x        |              |      |
| Hemolysin expression modulating protein                                 |              |           |          |              |      |
| Hemolysin activator protein precursor⁴                                  |              |           | x        |              |      |
| Regulators                                                               |              |           |          |              |      |
| DNA-binding proteins³                                                    | x            | x         |          | x            |      |
| dNTP triphosphohydrolase                                                 |              |           |          |              |      |
| Transcriptional regulators³                                             |              |           | x        | x            |      |
| HecB-like protein                                                       |              |           | x        |              |      |
Definition of the new genomic island GI<sub>selC<sub><sup>17584/1</sup></sub></sub>

As mentioned above, we were able to identify the complete foreign DNA insertion adjacent to <i>selC</i> of strain 17584/1 by whole-genome sequencing. Calculation of the GC-content of the <i>selC</i>-associated nucleotide sequence resulted in 48.4% and thus differs by 2.5% in comparison to the complete genome showing a GC-content of 50.9%. Because we detected the presence of integrases and mobile elements in addition to genes for increased fitness and putative roles in pathogenicity, we could define a new genomic island that was named GI<sub>selC<sub><sup>17584/1</sup></sub></sub> (Fig. 3).

A mapping of the contigs of the other sequenced strains using the GI<sub>selC<sub><sup>17584/1</sup></sub></sub> as reference (Geneious mapper) identified sequences spanning the GEI in <i>selC</i> of strain LM14603/08 within contigs 10, 122, 68, 151, 76 and 32 (see Fig. 4) and parts of the GEI in strain TS21/08 (contigs 3, 52) from pos. 1 to 27,118 (not shown).

Comparison of GI<sub>selC<sub><sup>17584/1</sup></sub></sub> to already described GEIs

Sequence comparison of GI<sub>selC<sub><sup>17584/1</sup></sub></sub> with other Enterobacteria revealed in parts similarity to other already described GEIs (see Fig. 5). Parts of GI<sub>selC<sub><sup>17584/1</sup></sub></sub> can be found in <i>E. coli</i> CFT073 (NC_004431.1), K-12 and EDL933 or in GEI PAI I<sub>536</sub>, PAI II<sub>536</sub>, PAI III<sub>536</sub> and PAI V<sub>536</sub> of UPEC 536 (Acc. No. NC_008253), SHI-2 of <i>Shigella flexneri</i> strain M90T (Acc. No. AF141323), EPI-I of ExPEC strain BEN2908 (Acc. No. AY857617), ARI<sub>EC20020119</sub> of <i>E. coli</i> O157:H7 strain EC20020119 (Acc

---

**Table 3 (continued)**

| Proteins                                      | TS18/08 | RF1a | K30 | 17584/1 |
|-----------------------------------------------|---------|------|-----|---------|
|                                               | pheV    | pheU | pheU| pheV    |
| ProQ/FINO family protein                      | x       |      |     |         |
| Others                                        |         | x    | x   |         |
| Helicases<sup>a</sup>                         |         | x    |     |         |
| Yjfd                                         |         |      |     |         |
| Yaf<sup>a</sup>                               |         |     |     |         |
| Inovirus Gp2 family protein<sup>a</sup>       |         | x    | x   |         |
| Outer membrane protein X precursor            |         | x    |     |         |
| KcA<sup>a</sup>                               |         | x    | x   | x       |
| Z5092<sup>a</sup>                             |         | x    |     | x       |
| EAL domain-containing protein                 |         |     |     |         |
| N-acetylgalactosamine 6-sulfate sulfatase (GALNS) | x     |     | x   |         |
| Lipoprotein bor                               |         |     |     |         |
| YfjI                                         |         | x    |     |         |
| Ash family protein                            |         |     |     |         |
| Restriction endonuclease                      |         |     |     |         |
| Putative phosphoethanolamine transferase<sup>a</sup> | x     | x    | x   | x       |
| l-lactate permease                            |         |     |     |         |
| Aec62<sup>a</sup>                             |         | x    |     |         |
| Putative competence protein                   |         |     |     |         |
| Hypothetical proteins                         |         | x    | x   | x       |

<sup>a</sup> Grouped or partly grouped as virulence factors by RASTk analysis

---

**Fig. 4** Alignment of <i>selC</i> genomic region of strain 17584/1 (yellow highlighted) with contigs of strain 14603/8
No. HQ018801) and PAI$_{4797/97}$ of STEC strain 4797/97 (Acc no. AJ278144).

The neighbourhood of selC was also detected as integration site for SHI-2 PAI of *Shigella flexneri* strain M90T. The GEI described here shares the first 2877 bp (pos. 338–3215, 99% identity) with SHI-2 including *shiA* (CDS 3). In addition, position 1–3658 bp shares 96.8% sequence identity to a 49.6 kbp-large genomic island EPI-I with an avian background and is described as comprising genes for carbohydrate metabolism, mobility and virulence [30]. Additionally, the direct repeats described in that study were also detected in the GI selC$_{17584/1}$ (2× TTTGACTCTTGTTAG at position 299 bp and TTTGTTGTTAAGGTTAG at pos. 433 bp and TTTGTTGTTAAGGTTAG at pos. 48,737 bp). Further, eight pairs of direct repeats between

---

**Fig. 5** Schematic overview of Blast results for selC GEI in comparison with different strains and pathogenicity islands. The query sequence (blue bars) of the new GEI shows information of nucleotide positions and, below that, the allocated sequence of the respective strains or PAIs (grey line) with significant sequence matches (red bars).
Comparison of the GEI with PAIs of UPEC strain 536 (Acc. No. NC_008253). CDS 10–12 (pos. 7632–9904 bp) is highly similar (100% identity) to the restriction-modification system originally described for *E. coli* strain HK31 (HK31IM Acc. No. X82231). CDS 22 has similarities to a transglutaminase-like enzyme and shows 98% homology with parts of a described antimicrobial resistance island in *E. coli* O157:H7 strain EC20020119 (Acc. No. HQ018801). This island was also identified as a genomic location for the sequence from pos. 14,956 to 15,783 with parts of CDS 21 and 23 of GIselC₁₇₅₈₄/1 by Blast comparison.

Other CDS of GIselC₁₇₅₈₄/1 (24–28, 35–37, 47–53) were also detected encoded on pathogenicity islands either in PAI I or PAI II of UPEC strain 536 (Acc. No. NC_008253). Comparison of the GEI with PAIs of UPEC strain 536 shows about 27,000 bp sequence overlap in several fragments and parts of PAI I, II and III within the GEI which is depicted in Fig. 5.

CDS 48, annotated as NgrB, shows 99% homology to a GTPase of the YeeP family and is also described as being encoded in PAI II of UPEC strain 536. CDS 60, annotated as Z1226 with 99% identity to a restriction methylase, is also found in the mentioned strain encoded on PAI V and homologs are found in UPEC strain CFT073. Also CDS 49–60 and the following sequence of the 3′ end with a size of 6581 bp share 99% homology with this UPEC strain.

**Discussion**

The results of this study have shown that LEE-negative STEC isolated from different foods contain foreign DNA in the three known LEE integration sites that may contribute to their fitness, potentially resulting in adaptation capacity in the host and also supporting their pathogenicity.

We have analysed the genome sequences of nine LEE-negative STEC strains for horizontally-acquired DNA adjacent to the typical LEE integration sites *pheU*, *pheV* and *selC* and found foreign DNA in at least one integration site in each strain. All three analysed integration sites of the strains LM27558, *x*₂₈ and LM14603/08 are occupied. Four strains were analysed in more depth, revealing genetic information for putative virulence and fitness factors at these integration sites. Many CDS encoded hypothetical proteins and all integration sites included mobile elements such as transposases and phage integrases (see Additional file 1: Tables S1–S5). Furthermore, all analysed strains carry genes at their integration sites that presumably encode for proteins with special domains of unknown functions.

The putative restriction methylase Z1226 locus was commonly identified near the 3′ terminus of the corresponding insertion in all strains. Moreover, all analysed integration sites contained the putative phosphoethanolamine transferases VtgX or YhbX.

The following transporter genes were identified in three (TS18/08, K30 and 17584/1) of the four strains. The product of the transporter gene *shiA* shows similarities to a quinone reductase/NADPH oxidoreductase protein [31]. OmpF forms pores in the outer membrane allowing small molecules to diffuse and could also be found in *E. coli* K-12. SapA is a putative ABC transporter. Antigen 43 is an outer membrane protein, autotransporter and a putative adhesion protein that was found in HUS sera [32–34]. The Per-activated serine protease autotransporter enterotoxin EspC, present in *pheV* of strain 17584/1, contains the superfamily Peptidase_S6 motif and is also designated as a hemoglobin-binding protease (hbp) (EGW81517), serum protease (WP_001367507), autotransporter (KLH79608), autotransporter outer membrane beta-barrel domain containing protein (ANES9238), serine protease pic autotransporter (EH135518) or serine protease SepA autotransporter precursor (BAX13941). Peptidase_S6 is present in many serine proteases as Hbp, EspP, Pet, EatA, EspC or Pic from *E. coli* or EspA from *Shigella flexneri*. In these proteins the Peptidase_S6 domain is combined with an autotransporter domain. The autotransporter domain is missing in the Per-activated serine protease autotransporter enterotoxin EspC analysed here. The gene is located at the 3′ terminus of contig 20 and the autotransporter domain was not captured on the same contig. To prove this we performed a PCR and amplified the putative whole gene, including the autotransporter domain and the signal peptide sequence. With the receipt of an amplificate length of 3993 bp, we could confirm the presence of the entire Per-activated serine protease autotransporter enterotoxin EspC gene in strain 17584/1 (see Additional file 1: Figure S6).

The *pheV*-adjacent foreign DNA of strain 17584/1 carries an *ato* system that was recently described in the Locus of Adhesion and Autoaggregation (LAA) pathogenicity island [35]. AtoS (sensor kinase) and AtoC (response regulator) belong to a two-component regulatory system that stimulates the expression of the *atoDAEB*-operon in the presence of acetate or spermidine [36–38]. The *atoDAEB*-operon is
fundamental for cellular processes such as short-chain fatty acid catabolism, poly-(R)-3-hydroxybutyrate biosynthesis and chemotaxis [39].

Surprisingly, three (TS18/08, RF1a, 17584/1) of the four strains encode a toxin-antitoxin gene pair (yeelV/yeelI). The toxin usually binds to an essential enzyme in the cell and inhibits the enzymatic activity. The antitoxin binds the toxin and restores viability. Growth inhibition was detected and could be restored by antitoxins but the physiological role is still under investigation. For chromosomally encoded toxin-antitoxin systems, two models for cellular function and role have been proposed: The first leads to programmed cell death in response to starvation by transcriptional attenuation using the toxin-antitoxin system and therefore providing nutrients for the remaining population [40]. The second function is to modulate the rate of metabolic processes in response to environmental stress [41]. The genes for YeeP/NgrB (50S ribosome-binding GTPase), YeeR (inner membrane protein), YeeS (metalloprotease) and YeeF (unknown function) are located in the same direction upstream of yeelI and yeelV. The toxin-antitoxin system was first described within the cryptic prophage CP4–44.

A vitamin B12 uptake system was detected in both pheV-integrated DNA segments in the strains TS18/08 and 17584/1.

Two strains carry genes encoding for hemagglutinin or hemolysin, both which are part of GsselC$_{17584/1}$. Blast comparison of CDS 28, the annotated hemolysin activator protein precursor (two partner secretion) family, shows sequence homologies (1732 of 1767 bp) to cdiB of the contact-dependent growth inhibition (CDI) system. This system is used by bacteria to express two-partner secretion proteins encoded by cdiA and cdiB to bind to BamA in the outer membranes of target cells and inhibit their growth [42, 43]. The CDI system was found in different genomic and pathogenicity islands and analysed in detail in uropathogenic E. coli [44]. CDS 25 with homology to an adhesin within the CDI system and CDS 27 annotated as CdiA, secreted exoprotein with conserved domains of hemagglutination activity, are both components of the CDI system. Annotated CDS 20 and CDS 23 are also both similar to hemagglutinin when compared to the NCBI database with Blastx. Furthermore, in pheV of TS18/08 the hha homologous gene for a hemolysin-expressing modulating protein is present, which down-regulates expression of hemolysin in a complex with H-NS in E. coli O6:H1. The RTX toxin activating lysin-acyltransferase, also termed Hemolysin C (HlyC), converts HlyA to an active toxin.

All analysed strains contain regulators that may control the expression of the surrounding genes or of more distantly encoded genes. The ProQ/FinO family is involved in the control of F plasmid transfer [45].

There are additional proteins that are not characteristically grouped but are mostly present in the analysed strains as, for example, the anti-restriction protein KlcA described by Serfiotis-Mitsa and coworkers [46], which is often encoded on plasmids, conjugative transposons and phages and is supposed to increase the chances of entering a new bacterial host due to a Type I DNA restriction and modification (RM) system, which would usually destroy the invading DNA [46]. YjI, Ykff (DUF905 domain containing protein) and YafZ are described in E. coli K-12 with unknown protein functions. As Ykff and YafZ, the inovirus Gp2 family proteins were partially grouped to the virulence factors by RASTk analysis but their function is still unknown. Z5092 encodes an uncharacterised RNA-directed DNA polymerase. The EAL domain is found in signaling proteins [47]. The lambda phage lipoprotein Bor is an outer membrane protein of E. coli and confers serum resistance [48].

The GsselC$_{17584/1}$ was partially found in another characterised, already published, strain. Bertin and coworkers described an altered selC locus with CP4-int at the 3′ terminus and similarities to the locus of proteolysis activity (Acc No. AJ278144 strain 4797/97 STEC) and SHI-2 (Acc No. AF141323 strain M90T S. flexneri) and demonstrated that the selC locus is frequently used as an integration site for PAIs with CP4 integrase genes [49]. Homologies to CP-4 phage-derived proteins were also found in the GsselC$_{17584/1}$. The described genomic island is of mosaic composition and harbours several already described parts from other genomic/pathogenicity islands. Several parts of UPEC strain 536 were found throughout the GEI and the last approx. 6000 bp at the 3′ terminus show homologies to UPEC strain CFT073 (see Fig. 5). The complete sequence of the GEI is only found in STEC strains 117 and 453 (Acc. No. NZ_MPGR0100001.1 and NZ_MPGR0100001.1) isolated from deer feaces in Finland with 99% similarity each, which may hint at a recent clonal distribution of the GEI. Large parts were also detected in two E. coli strains FHI98 and FHI29 from human feaces in Norway (Acc. No. LM997367.1 and Acc. No. LM995856.1). The mosaic structure of the GsselC$_{17584/1}$ with sequences partly described in several different other genomic islands illustrates again the high genome plasticity within the E. coli species independent of the different pathotypes. We can only speculate on the role of this additional DNA adjacent to selC in foodborne strain 17584/1 and others. Since many of the already described protein functions deal with increased fitness, response to environmental stress or adhesion and fine-tuning of the metabolism, we suppose this genomic island is a collection of DNA and functions as a survival
kit for the bacterium in unfavorable environmental conditions.

In conclusion, we detected foreign DNA in the well-known LEE integration sites phell, pheV and selC in different LEE-negative STEC strains. Some annotated CDS located in these integration sites can be designated as virulence factors. However, whether foreign DNA in the different integration sites supports virulence or pathogenicity to the respective strains or whether the occupation of the integration sites hinders the reception of other high impact pathogenicity islands like the LEE and is, therefore, a cause for decreased pathogenicity will be an interesting question for future studies.

Additional file

Additional file 1. Additional Figures S1–S6 and Tables S1–S4.

Abbreviations

LEE: Locus of Enterocyte Effacement, STEC: Shiga toxin-producing E. coli; GEI: genomic island; PAI: pathogenicity island; EHEC: enterohemorrhagic E. coli; UPEC: uropathogenic E. coli; nt: not typeable; nm: non motile; DUF: domain of unknown function; CDS: coding DNA sequence.

Authors’ contributions

NS performed experiments, reanalyzed the genome data, revised the manuscript and interpreted data. ES planned and performed experiments, performed the primary analysis of the genome data, interpreted the results and wrote the manuscript. TS and AB contributed with NGS genome sequencing and analysis, IE supported the work with genome analysis, and revision and interpretation of the results, and revised the manuscript. HS wrote the grant proposal, made the conception for the project, interpreted data and revised the manuscript and interpreted data. All authors read and approved the final manuscript.

Author details

1 Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstr 28, 70599 Stuttgart, Germany. 2 Department Biological Safety, National Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstr 28, 70599 Stuttgart, Germany. 2 Department Biological Safety, National Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstr 28, 70599 Stuttgart, Germany. 3 Department Biological Safety, National Institute of Food Science and Biotechnology, University of Hohenheim, Garbenstr 28, 70599 Stuttgart, Germany. 4 Institute of Microbiology and Epizootics, Reference Laboratory for Escherichia coli, Federal Institute for Risk Assessment (BfR), Braunschweiger Weg 1, 12277 Berlin, Germany. 5 Robert Koch Institute, Nordufer 20, 13353 Berlin, Germany. 6 Institute of Microbiology and Epizootics, Freie Universität Berlin, Robert-von-Ostertag-Str. 7-13, 14163 Berlin, Germany. 7 Institute for Hygiene, University of Münster, Robert-Koch-Str. 41, 48149 Münster, Germany.

Acknowledgements

We thank Annette Bruckbauer and Claudia Lis for skillful technical assistance. This work was financially supported by the German Federal Ministry of Education and Research (BMBF) Grants 01KI1012A, 01KI1012B, and 01KI1012C.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

All data generated or analysed during this study are included in this published article (and its additional information files).

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

This work was financially supported by the German Federal Ministry of Education and Research (BMBF) Grant 01KI1012C.

Publisher’s Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 10 July 2018 Accepted: 1 October 2018

Published online: 08 October 2018

References

1. Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, Rohde C, Rohde M, Farthmann B, Goodwin LA, Chertkov O, Reddy T, Pati A, Ivanova NN, Markowitz V, Kyripets NC, Woyte K, Goker M, Klenk HP. Complete genome sequence of DSM 30083(T), the type strain (US(41(T)) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. Stand Genomic Sci. 2014;9:2.

2. Berghorsson U, Ochman H. Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol Biol Evol. 1998;15:6–16.

3. Ogura Y, Okata T, Iwaguchi T, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc Natl Acad Sci USA. 2009;106:17939–44.

4. Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA. 1998;95:949–9.

5. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol. 1997;23:1089–97.

6. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 2000;54:641–79.

7. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeswyl S, Bidet P, Bingen E, Boncorici 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, Matic I, Oztas S, Petit MA, Pichon C, Rouy Z, Ruff CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médiouge C, Rocha EC, Denamur E. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;1:e1000344.

8. Jones J, Rumer L, Wieler LH. Impact of the locus of enteroagglutin- pathogenicity island on the evolution of pathogenic Escherichia coli. Int J Med Microbiol. 2004;294:103–13.

9. Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB, Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA. 2004;94:103–13.

10. Fleckenstein JM, Lindler LE, Elsinghorst EA, Dale JB. Identification of a virulent isolate pathogenicity island on the evolution of pathogenic Escherichia coli. Mol Biol Evol. 2004;21:267–80.

11. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 2000;54:641–79.

12. Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, Rohde C, Rohde M, Farthmann B, Goodwin LA, Chertkov O, Reddy T, Pati A, Ivanova NN, Markowitz V, Kyripets NC, Woyte K, Goker M, Klenk HP. Complete genome sequence of DSM 30083(T), the type strain (US(41(T)) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. Stand Genomic Sci. 2014;9:2.

13. Berghorsson U, Ochman H. Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol Biol Evol. 1998;15:6–16.

14. Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA. 1998;95:949–9.

15. Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol. 1997;23:1089–97.

16. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 2000;54:641–79.

17. Touchon M, Hoede C, Tenaillon O, Barbe V, Baeswyl S, Bidet P, Bingen E, Boncorici 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, Matic I, Oztas S, Petit MA, Pichon C, Rouy Z, Ruff CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médiouge C, Rocha EC, Denamur E. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009;1:e1000344.

18. Jones J, Rumer L, Wieler LH. Impact of the locus of enteroagglutin- pathogenicity island on the evolution of pathogenic Escherichia coli. Int J Med Microbiol. 2004;294:103–13.

19. Perna NT, Mayhew GF, Posfai G, Elliott S, Donnenberg MS, Kaper JB, Lawrence JG, Ochman H. Molecular archaeology of the Escherichia coli genome. Proc Natl Acad Sci USA. 2004;94:103–13.

20. Fleckenstein JM, Lindler LE, Elsinghorst EA, Dale JB. Identification of a virulent isolate pathogenicity island on the evolution of pathogenic Escherichia coli. Mol Biol Evol. 2004;21:267–80.

21. Hacker J, Kaper JB. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol. 2000;54:641–79.
15. Sun J, Inouye M, Inouye S. Association of a retroelement with a P4-like cryptic prophage (retrophage phi R73) integrated into the selenocystein tRNA gene of Escherichia coli. J Bacteriol. 1991;173:417–81.

16. Lloyd AL, Rasko DA, Mobley HL. Defining genomic islands and uropathogen-specific genes in uropathogenic Escherichia coli. J Bacteriol. 2007;189:3532–46.

17. Girardeau JP, Bertin Y, Martin C. Genomic analysis of the PAI ICL3 locus in pathogenic LEE-negative Shiga toxin-producing Escherichia coli and Citrobacter rodentium. Microbiology. 2009;155:1016–27.

18. Shen S, Mascarenhas M, Rahn K, Kaper JB, Karmali MA. Evidence for a hybrid genomic island in verocytotoxin-producing Escherichia coli CL3 (serotype O113:H21) containing segments of EDL933 (serotype O157:H7). O islands 122 and 48. Infect Immun. 2004;72:1496–503.

19. Lalouli L, Le Bouguenec C. afa-E Gene cluster is carried by a pathogenicity island inserted into the tRNA(Phe) of human and bovine pathogenic Escherichia coli isolates. Infect Immun. 2001;69:937–48.

20. Swenson DL, Bukkanov NO, Berg DE, Welch RA. Two pathogenicity islands in uropathogenic Escherichia coli. J Bacteriol. 1997;179:3971–83.

21. Al-Hasani K, Rajakumar K, Bulach D, Robins-Browne R, Adler B, Sakkellari H. Genetic organization of the shig gene pathogenicity island in Shigella flexneri 2a. Microb Pathog. 2001;30:1–8.

22. Purdy GE, Payne SM. The SHI-1 iron transport island of Shigella boydii 0–1392 carries the genes for aerobactin synthesis and transport. J Bacteriol. 2001;183:4176–82.

23. Mathusa EC, Chen Y, Enache E, Hontz L. Non-O157 Shiga toxin-producing Escherichia coli in foods. J Food Prot. 2010;73:1721–36.

24. Hauser E, Mellmann A, Semmler T, Stoeber H, Wieler LH, Karch H, Kuebler RL. Improvements to PATRIC, the all-bacterial bioinformatics database. Sci Rep. 2015;5:8365.

25. Brettin T, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Wattam AR, Davis JJ, Disz T, Edwards RA, Finlay B, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Parrello B, Pusch GD, Shukla M, Thomason JA 3rd, Stevens R, Velasco J, Del Canto F, Vidal R, Williams RL. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res. 2017;45:DS35–42.

26. Chouikha I, Germon P, Bree A, Gontard A. The selC-associated genomic island of the extraintestinal avian pathogenic Escherichia coli strain BEN2908 is involved in carbohydrate uptake and virulence. J Bacteriol. 2006;188:977–87.

27. Moss JE, Cardozo TJ, Zychlinsky A, Grossman EA. The selC-associated SHI-2 pathogenicity island of Shigella flexneri. Mol Microbiol. 1999;33:74–83.