Stitching genomics data to protein structures: Virulence factors in non-O157 Shiga toxin-producing *Escherichia coli*

Sony Malhotra[1], Ashley Ward[2,3,], Lauren Giles[1], Tom Gerrard[4], Martyn Winn[1*], and Nicola Holden[2*]

[1]Scientific Computing Department, Science and Technology Facilities Council, Research Complex at Harwell, Didcot, OX110QX, United Kingdom.
[2]School of Veterinary Medicine and Biosciences, Scotland's Rural College, Aberdeen AB21 9YA, United Kingdom
[3]School of Veterinary Science, University of Liverpool, Department of Livestock and One Health, Neston CH64 7TE, United Kingdom
[4]University of Bristol, School of Cellular and Molecular medicine, Biomedical Sciences building, University Walk, Bristol, BS8 1TD and University of the West of England, School of Applied Sciences, Frenchay Campus, Coldharbour Lane, Bristol, BS16 1QY, United Kingdom

*corresponding authors: martyn.winn@stfc.ac.uk, nicola.holden@sruc.ac.uk

Abstract

The genetic diversity of the Shiga toxin-producing *Escherichia coli* (STEC) sub-species group poses major challenges for assigning pathogenic potential, and for accurate diagnostics. Although existing risk-based frameworks work for the predominant serotypes e.g. STEC O157 or O26, they are less useful for more diverse genotypes, with variation in the genetic complement of virulence factors that is responsible for disease outcome. Moreover, genetic diversity occurs at the virulence factor allele level. We investigated the functional potential of genetic variation at a whole genome level and at an allele variant level, based on the premise that genomic variability underlies phenotypic traits including disease outcomes. We analysed 286 non-O157 STEC genomes for their virulence factor complement and determined co-occurrence patterns. Genetic variation in the well-characterized virulence factor intimin (*eae*) and its receptor (*tir*) was analyzed at a sequence level, and by modelling the three-dimensional protein–protein complex using AlphaFold3 and homology modelling. Different subtypes of Eae were shown to preserve their interaction with the receptor Tir by retaining the interactions at the protein-protein interface. The study shows that virulence factor characterisation in association with 3-D modelling of protein complexes can aid genomic analysis for risk assessment, and inform on functional mechanisms for pathogenicity.

1. Introduction

Shiga toxin-producing *Escherichia coli* (STEC) are a group of foodborne pathogenic bacteria that are predominantly transmitted to humans through contaminated food or from direct contact, particularly from farmed ruminant animals. Diseases caused by STEC range from diarrhoea (D), blood diarrhoea (BD) to more severe haemolytic uremic syndrome (HUS). Animals harbouring STEC do not typically demonstrate signs of infection, creating barriers to risk identification. Transmission of the pathogens through environmental routes, like water, creates a risk of infection from ready-to-eat produce. Treatment of infected patients is limited as antibiotic usage is (for the most part) contra-indicated due to increased risk of release of the Shiga toxin. As such, risk management with surveillance of multiple critical control points is the main means to prevent STEC transmission into the food chain.

Detection of the STEC group of pathogens is challenging due to their serological and genetic diversity. Serotypes other than O157 are increasingly detected and have become more prevalent[1,2], with >338 detected in clinical infections[3]. In many European nations, serogroup O157 continues to be associated with the highest number of HUS cases, hospitalisation and bloody diarrhoea, followed by O26[4]. Identification of pathogenic non-O157 STEC is complicated as not all have the ability to cause disease[5]. As such, there is a pressing need for methods that can first identify them easily, and second, accurately indicate the likely level of risk.

Risk classification of STEC is currently based on the carriage of a particular set of virulence genes, as defined by a Joint FAO / WHO Expert Meeting on Microbiological Risk Assessment (JEMRA) Group[6]. The set of criteria and resulting decision tree (known as the JEMRA STEC risk matrix) are universally used as a means to classify pathogenicity, based on the Shiga toxin (*stx*) variant and carriage of adherence factors including initimin (*eae*) and / or Aggregative Adherence Fimbriae (*aaf*) adhesin (*aggR*). Hazard classification took into account other putative adherence factors, different serotypes and hybrid pathotypes of *E. coli*[6]. The JEMRA risk classification matrix can be supplemented by ranking risks associated with outcomes of clinical disease severity[4] and using them as predictors for disease[7], also based on *stx* variant and carriage of *eae*. The presence of *stx2* is strongly associated with severe disease progression, with highly enhanced potency compared to *stx1*[8]. Subclasses of *stx* can bind different glycosphingolipid receptors, although Gb3 is the principal entry point[9]. The *eae* gene is located on the 35-kb LEE[10] (locus of enterocyte effacement) pathogenicity island along with other effectors and genes for the type 3 secretion system (T3SS). The presence of *eae* together with

*stx*2 is commonly used as a molecular signal to predict pathogenic outcome in clinical settings[11,12]. Intimin (Eae) binds to the translocated intimin receptor (Tir), with the corresponding *tir* gene also located on the LEE. Tir is injected by the T3SS into host cells[13] and embeds itself into the host membrane so that it can bind Eae[14]. For enterohaemorrhagic strains (EHEC *i.e.* most STEC), this is a precursor to the formation of attaching and effacing (A/E) lesions and actin-filament rearrangements resulting in pedestal formation[15]. Similarly in enteropathogenic *E. coli* (EPEC, a different diarrheagenic *E. coli* that is also characterised by A/E lesions but does not encode the Shiga toxin gene), *eae-tir* also facilitates intimate host binding[16].

It is notable that a relatively high proportion of hospitalisation and bloody diarrhoea disease occur in the absence of *eae* carriage[4]. As the minimum virulence gene repertoire is still unclear, computational methods based on genomic content are now being tested to help with this determination. Machine learning approaches have predicted that serious disease (BD and HUS) is associated with additional factors in the O157 serotype beyond established virulence factors[17].

The numbers of STEC serotypes other than O157 is increasing year-on-year across the UK, and in some regions has become more dominant. In England, there were 2,762 total cases of non-O157 STEC reported in 2024, an increase of 22% from 2023 and 175% compared to 2019[1]. In contrast, there were 564 confirmed cases of STEC serotype O157 in England (2024[1]), a decrease of 26% compared to 2022[18]. The same trend is seen in Scotland with increasing non-O157 STEC, of 182 laboratory reports in 2023[2] compared to 170 in 2022[18]. The level of STEC serotype O157 is proportionally lower in Scotland, with 133 laboratory reports in 2023, an decrease from a spike of 270 reported in 2022 (PHS GIZ 2020-2021 report[19]). In European member states there has been a continuous increase in serogroups O26, O146, O145, and O103, although O157 remains the most frequently reported (19%)[20]. The main non-O157 STEC serotypes that are associated with clinical disease include O26:H11 and O145:H28 (as listed in Enterobase[21], 1952-2024, Table S1). Genomic diversity that occurs between serotypes encompasses surface factors affecting attachment phenotypes, virulence factors affecting pathogenicity and host colonisation, and horizontally acquired elements including prophage, insertion elements and plasmids[22].

Genomic variation is a reflection of selection pressure on the gene products and corresponding protein function. This occurs where a genomic complement or gene allele type has been under positive selection in a given environment or host, and so represents the successful clone. In

turn, it raises the possibility that such genomic variation can be used to indicate particular traits as a result of host or habitat niche adaptation, such as pathogenicity within a human host or colonisation of secondary hosts. This broad principle has been applied to another set of pathogens, *Salmonella* species, where machine learning has been applied to protein functional domains to assign biomarkers related to disease severity potential[23].

To investigate the pathogenic potential of less well characterised non-O157 STEC, a set of clinical isolates (n=286) was collated and the key virulence factors identified through genome analysis. A reference panel of known virulence factors[24,25] was used to map these isolates for the presence and distribution of the corresponding genes. Co-occurrence patterns of these genes were then assessed to explore possible functional relationships that may dictate their pathogenic potential. As expected, the *eae* and *tir* virulence genes, associated with A/E lesions, were consistently associated, indicating the presence of the LEE pathogenicity island in a subset of isolates[26]. Within this subset, we could identify multiple eae subtypes and further allelic variation within the subtypes.

We propose that the genomic variability underlying such traits can be used in diagnostics applications. Therefore, investigating allelic variations can help understand their impact on potential protein function and pathogenicity. Three-dimensional protein structure modelling (using AlphaFold3[27]) was used to model the Intimin-Tir complexes and map the allelic variations onto the structures for assessing the impacts on protein structure and function.

2. Materials and Methods

2.1 Data Collection

Human (clinical) isolates were selected from Enterobase[21] (queried Dec 2024), using the following filters:

a. Geographical location: United Kingdom; UK; Great Britain; GB; England; Scotland; Wales; Northern Ireland /NI (but not 'Ireland')
b. Source type: Human
c. '*E. coli'* only and Species: exclude all isolates including "*Shigella*"
d. Serotype filter: anything with O and H type included, all other descriptors removed. Included ones with O type only, removed all blanks or listed as 'Escherichia coli'. All the O157 were removed as the focus of this study is non-O157.
e. Path/Non path: Pathogenic & blanks included
f. Pathotypes related to UTI or ExPEC removed. Included EIEC.
g. Lab contact: Anything that contain 'Public Health England' included in email addresses
h. Collection year: from 2010 onwards

These entries in Enterobase do not have clinical outcome data because questionnaires are not routinely administered to non-O157 cases in the same way as they are STEC O157 cases (pers. communication Claire Jenkins).

Within the non-O157 STEC dataset, O26 STEC isolates were eliminated as it is the most prevalent non-O157 STEC serotype in the UK[1,2] and could disproportionately skew the dataset and subsequent analyses. Furthermore, O26 is already one of the most extensively characterized non-O157 serotypes. The aim was to obtain a broader snapshot of recently emerged and less-characterized strains. After applying the filters described above and removing O26, to further capture H-antigen diversity and maximise the diversity of serotypes, an isolate per H-antigen type was selected for each non-O157 serotype.

Further, these isolates were filtered for the presence of *stx* gene to identify the (non-O157) STEC genomes. For this, the paired end reads from Sequence Read Archive[28] (SRA) were

downloaded using the fastq-dump utility, and filtered using the BBTools[29]. The following criteria were used for cleaning the downloaded reads:

i. discard reads with average quality below 20.
ii. Trim adapters based on pair overlap detection (does not require known adapter) using BBMerge
iii. Trim both reads to the same length
iv. Also used the generalized adapters and trimmed provided by BBTools package which includes Illumina Truseq and Nextera adapter sequences

The obtained cleaned reads were then mapped to the *E. coli* stx gene sequences downloaded from the VirulenceFinder 2.0 (as described in section 2.2), using bowtie2[30] and samtools[31]. The downloaded set of stx sequences from VirulenceFinder 2.0 cover a variety of stx subtypes and include subtypes stx1a, stx1c, stx1d, stx1e, stx2a, stx2b, stx2c, stx2d, stx2e, stx2f, stx2g, stx2h, stx2i, stx2j, stx2k, stx2l, stx2m, stx2n and stx2o.

The resulting stx+ genomes were assembled using the Shovill pipeline (https://github.com/tseemann/shovill) and the Spades assembler[32] which scans over multiple kmer lengths. Prokka[33] was used to annotate the assembled genomes and identify genes.

2.2 Virulence Factors Mapping

The list of *E. coli* virulence factors (virulence_ecoli.fsa) was downloaded from the Virulence Finder 2.0[24,25] (database version 2022, last updated in April 2024, https://bitbucket.org/genomicepidemiology/virulencefinder_db/src/master/). For some of the virulence factors such as intimin (*eae*), the downloaded annotations include the subtype mapping. The whole list comprises 4942 sequences of virulence factors including the subtypes and their multiple sequences, which maps to 680 sequences for different subtypes of virulence factors by removing alleles.

Virulence Finder provides a very broad list that includes any factor associated with pathogenicity, but many are simply colonisation factors and not true virulence genes. The sequence identity between related entries can also be high. Thus, *stx1* and *stx2* subtypes are just a few SNPs different and there can be multiple subtypes per genome which may cause the assembly to split the gene.

An in-house database was built using BLAST[34] for the reference virulence factor sequences which was then used to search for virulence factors in the Prokka-annotated isolate genomes. tblastn[35] was used to query the database with an evalue cut off 10^{-4}. The hits were then filtered using the following criteria:

1. percent identity >=90%
2. query coverage >= 75%, where query coverage is the aligned query/length of query

Using these two filters if a sequence maps to more than one virulence factor gene, the hit with the greatest percent identity was selected. Further filtering was done using the subject (virulence factor gene) coverage cut off of >=75%, where subject coverage is the aligned subject/subject length to filter out the genome sequences which are just partial sequences.

For isolates carrying the intimin gene, the main genetic subtype (alpha, beta, gamma, etc) was identified based on the Virulence Finder annotation of the closest sequence match, noting that the database includes several entries for each subtype. Although Virulence Finder contains sequences for the intimin receptor, encoded by the *tir* gene, there is no equivalent subtype annotation for these entries.

The term ‘variant’ is used throughout the text and is defined in context, with respect to STEC serotype, gene presence/absence at the genome level, specific gene subtypes, or within gene point mutations.

2.3 Co-occurrence of virulence factors

Entries in a co-occurrence matrix were defined as the number of isolates in which a pair of virulence factors occur together, normalised by the total occurrence of the virulence factor on the row, so that the matrix is asymmetric with respect to the two virulence factors. An enrichment factor was also defined of a virulence factor *A* with respect to a virulence factor *B* as the ratio of the occurrence of *A* in *B*+ isolates to its occurrence in *B*- isolates, adjusted for the relative number of each group of isolates:

Enrichment of virA in presence of virB: $E_A^{(B)}$ = (virAB/virB) / ((virA – virAB) / (286 – virB))

A value greater than one implies that a virulence factor *A* is more likely to be found in an isolate containing *B* than one where *B* is absent. The Jaccard similarity was also considered, J_{AB}, defined as the number of isolates containing both virulence factors divided by the number of isolates that contain at least one.

2.4 Sequence alignment and phylogenetic trees

Sequence alignments were performed using MUSCLE[36]. RaxML[37] was used to generate phylogenetic trees for different sets of sequences. For extracting the LEE island, the LEE sequence from the *E. coli* O157:H7 strain Sakai (GenBank accession no. BA000007) was used. BLAST[34] against the Sakai sequence was performed using all the contigs from the *eae*+ genomes. A contig was considered to contain a LEE island if it shared at least 80% sequence identity with the Sakai LEE sequence and the alignment length was at least 20 kb. The LEE sequences were extracted from the *eae*+ isolates (n=85) and the pangenome was built using Panaroo[38].

The pangenome for the *stx*+ isolates (n=286) was built using Panaroo[38] and the annotation previously output from Prokka. RaxML (v8.2.12)[37] was used to build an alignment of the core genome. Visualisation of the annotated trees was done in Iroki[39].

Sequence alignments were used as an input to SplitsTree[40] to compute a split decomposition network (selecting network -> split decomposition).

2.5 Three dimensional structural modelling of virulence factors

AlphaFold3[27] was used to model the structures of protein complexes. The protein sequences of Eae (D1, D2 and D3 domains) and Tir (Tir intimin binding domain) in Fasta format were used as input to the AlphaFold3 server (https://alphafoldserver.com/). In total, ten subtypes of Eae were modelled using sequences taken from isolates for which the subtype had been identified, with the cognate Tir sequence taken from the same isolate. The following accession IDs were used for structural modelling of Eae subtypes: alpha:SRR3581427, beta:SRR4180889, gamma:SRR3581353, rho:SRR3241980, theta:SRR4179996, epsilon:SRR3574270, xi:SRR16203499, delta:SRR29176816, iota:SRR4195732 and zeta:SRR4184837. For each subtype, a separate job was run, using the Eae protein sequence as entity 1 (copy = 1) and the Tir protein sequence from the same genome as entity 2 (copy = 1). The server returns top five

scoring complexes and the one with highest ranking_score was selected. Ranking_score combines pTM (predicted TM-score for full structure), ipTM (predicted interface TM-score), penalises clashes and encourages the absence of spurious helices in disordered regions[27].

Additional homology based structural models were built for some of the subtypes (rho and theta) using Modeller (v9.10)[41]. The template used for the homology modelling is PDB ID: 1F02 and the model with lowest DOPE (Discrete Optimised Potential Energy) score was selected as the best model.

3. Results

3.1 Collation of STEC isolate genomes

3.1.1 Initial download

A retrieval of all E. coli isolates from EnteroBase yielded 86,609 accessions. The filtering criteria described in Section 2.1 (Methods) were applied to retain clinical, non-O157 isolates, resulting in 7,855 isolates. The most common serotypes (Figure 1A) in this subset were O26 (n = 1,027), O145 (n = 554), and O146 (n = 515). The O26 serotype was subsequently excluded because it is already well characterised and overrepresented in public databases. To ensure a broader snapshot of recently emerged and less-characterized strains, one isolate per H-antigen was retained for each remaining non-O157 serotype. This resulted in a filtered subset of 1177 isolates.

Entries in Enterobase do not have clear metadata on pathogenicity or clinical outcome. We make the assumption that all of the human isolate sequences selected from Enterobase using this protocol were associated with clinical disease symptoms, since the isolates in this collection were all deposited by reference or diagnostic laboratories. They are hereafter referred to as 'clinical isolates'.

3.1.2 Presence of Shiga toxin gene

Isolate genomes could not be filtered in Enterobase by the presence of *stx* gene(s) because of incomplete metadata, nor were they generally designated explicitly as STEC. Therefore, the genome sequence reads were downloaded and mapped to the *E. coli stx* gene alleles from the Virulence Finder v2.0 database[24,25]. Out of the 1177 genomes in the original Enterobase set, 326 genomes had at least one read mapped to one of the *stx* gene subtypes. This does not necessarily imply that all these 326 genomes are true STEC due to the low threshold of number of reads mapped to the *stx* genes. Therefore, confirmation of *stx* presence was carried out at the annotation level following assembly of the 326 genomes.

3.1.3 Filtering for STEC genomes

The 326 putative STEC genomes were assembled and annotated, as described in section 2.1. A more stringent check for the presence of *stx* genes was based on *tblastn* searches, using the protein sequences obtained after the annotation step (section 2.2). As the reference *stx* genes

(from Virulence Finder v2.0) includes both subunit A and subunit B genes, the BLAST hits were further validated, and only genomes that included both *stx* toxin A and B subunits were selected. This resulted in 286 *stx*+ isolates*,* henceforth referred to as non-O157 STEC isolates. The Enterobase accession numbers for these 286 isolates are provided in Table S2. The distribution of *stx* types among serotypes present in more than five isolates is shown in Figure 1B. Sole carriage of *stx2* only occurred in serotype O8 and O80, while conversely sole carriage of *stx1* occurred in O78 and O79 serotypes. All other serotypes (9) encoded both *stx1* and *stx2* subtype reads.

3.2 Virulence factor genes in STEC genomes

3.2.1 Virulence factors

The complete set of virulence factor alleles was mapped onto non-O157 STEC genomes (n=286) using tblastn[35] following the annotation of the assembled genomes (see Section 2.2). For ease of presentation, virulence factors are referred to by their 3-letter gene name, of which there were 85 present in the set of non-O157 STEC isolates (Table S3) . Where appropriate, individual genes were identified, of which there were 130 in the genome set when allelic variants were merged (Table S4). Isolate genomes varied in the number of virulence factors identified, and Figure 1C shows the distribution of virulence factor load. The genome with maximum number of virulence factors is accession SRR3579389 with 52 factors mapped to it, in contrast to minimum number of virulence factors (n=12) in the genome of accession SRR29241024.

For the two main *stx* groups, 128 genomes encoded only the *stx1* subtype, 110 had only *stx2* subtype, and the remaining 48 genomes contained representatives of both *stx1* and *stx2*. Further division into *stx* subtypes was carried out using the reference sequences in Virulence Finder v2.0 (Figure 1D). We observed two subtypes of *stx1* (1a and 1c, roughly in equal numbers) and seven subtypes of *stx2* (2a,2b,2c,2d,2f,2g,2k) across the 286 isolates (Figure 1D). Of these seven stx2 subtypes, the most common were *stx2b* and *stx2c*, with low numbers of five other subtypes.

The most frequent virulence factors were assessed, classed as present in at least 50% of the genomes (Figure 2A). *nlpI* (lipoprotein NlpI precursor) and *terC* (tellurium ion resistance protein) are present in all 286 genomes, followed by *yeh* (4 genes, YHD fimbrial cluster) and *csgA* (curli

major subunit) which are present in 283 and 281 isolates respectively. Nine other virulence factors are widespread in our set of isolates, though still absent from a significant fraction.

Co-occurrence of virulence factors was determined for the set of 286 non-O157 STEC genomes (Figure 2B). The co-occurrence values were normalised for a pair of virulence factors by the total occurrence of the virulence factor on the row, so that the matrix is asymmetric with respect to the two virulence factors.

First, co-occurrence of virulence factors with intimin was a point of focus, as it is one of the key virulence factors for the STEC group[41] in terms of incidence and disease outcome (see row 'eae' in Figure 2B) [42]. To help identify significant correlations with intimin, an enrichment factor $E_A^{(eae)}$ was defined as the ratio of the occurrence of a virulence factor in *eae*+ isolates to its occurrence in *eae*- isolates, adjusted for the relative number of each group of isolates. The Jaccard similarity, $J_{A,eae}$, between the set of isolates containing a specified virulence factor and the set carrying *eae* was also considered. Values for all virulence factors are listed in Tables S3 and S4. By these definitions (see Methods), *tir* has $E_{tir}^{(eae)}$ of infinity as it never occurs without its binding partner, and $J_{tir,eae} = 1$ since it has an identical distribution across the isolates. On the other hand, *etp* has $E_{etp}^{(eae)}$ of infinity as it only occurs in *eae*+ isolates, but $J_{etp,eae} = 0.27$ since it only occurs in 23 of the *eae*+ isolates and thus has a different distribution.

The gene group *nleA, nleB, nleC* comprises non-LEE effectors which are functionally dependent on the LEE-encoded T3SS. At least one *nle* gene is found in 83 out of 85 *eae*+ isolates, while all three are found in a single *eae*- isolate. The presence of *eae* and *nle* genes is reported to be a signature for highly virulent STEC strains and are reported to co-occur in non-O157 serotypes[43]. *nle* is reported to correlate with outbreak and HUS potential[44].

For the *cif* gene, 37 out of 38 occurrences are in *eae*+ isolates, while for both *espJ* and *efa1* it is 44 out of 45. Several of these genes have been observed on a lambdoid prophage ΦE22[45] and so may be carried together. A single isolate (AIPCOOIH) carries the only occurrences of *nleA, nleB, nleC, cif, espJ* and *efa1* that are not associated with intimin.

Other genes enriched in *eae*+ isolates include *espP*, *tox*, *katP* and the bacteriocin *cia*. Table 1 lists other virulence factors that were observed to co-occur with eae, though some are also widely seen in *eae*- isolates. Some of these are toxins (such as hemolysin, enterohemolysin),

adherence and colonisation factors (*efa, fdeC*), or help in evading the bacterial complement system (*iss*).

Table 1: Virulence factor genes that show significant co-occurrence with *eae* in non-O157 STEC isolates. The co-occurrence is defined as the number of isolates carrying both the virulence factor of interest and *eae*, as a fraction of the total number of *eae*+ isolates, corresponding to the values on the *eae* row of Figure 2B. The following column gives the absolute number of isolates carrying the virulence factor, divided between *eae*+ and *eae*- isolates.

Gene co-occurring with *eae*	Co-occurrence	Number in *eae*+ / *eae*- isolates	Short description
cif	0.44	37 / 1	Type III secreted effector
efa1	0.52	44 / 1	EC factor for adherence; lymphostatin *Efa1-LifA*
ehxA	0.76	65 / 105	Enterohaemolysin:known to be associated with HUS and diarrheal disease
espA, espB, espC, espF, espI, espJ, espP, espY	1.00	85 / 48	E.coli secreted protein; *espA*, *espB*, *espF* carried on LEE island; *espP* serine protease
fdeC	0.73	62 / 175	intimin-like adhesin *fdeC*: biofilm formation; structural similarity to intimin
hlyA, hlyE, hlyF	0.67	57 / 98	Hemolysin; known to be associated with HUS and diarrheal disease
iha	0.46	39 / 121	*irgA* homologue adhesin (*Iha*)
iss	0.69	59 / 141	Increased serum survival; complement resistance
lpfA	0.45	38 / 148	Long polar fimbriae: adhesive factor;also known to be expressed in LEE-negative clinical isolates
nleA, nleB, nleC	0.98	83 / 1	Non-LEE effectors
ompT	0.94	80 / 137	Outer membrane protease (protein protease 7):
traJ, traT	0.68	58 / 166	*traJ* (Positive regulator of conjugal transfer operon)) and *traT* (Outer membrane protein

			complement resistance:);
yghJ	0.73	62 / 132	Secreted and surface associated lipoprotein; aids in colonisation

In contrast, many pairs of virulence factors are mutually exclusive, meaning that there are no isolates that contain both. This may simply be because of low overall occurrence in the dataset, or may reflect some aspect of genomic structure *e.g.* competition for the same phage integration site. In particular, the genes *subA*, *kpsM*, *pic* and *senB* occur 79, 92, 24 and 67 times respectively, but never in *eae+* isolates. The *subA* gene encodes a component of the SubAB toxin, which has been observed in LEE-negative STEC strains including clinical human isolates[46]. The gene *pic* encodes for a member of the serine protease auto-transporter (SPATE) family and has been shown to have mucolytic activity[47].

The *ireA* gene occurs in 99 eae- isolates and a single *eae+* isolate (SRR3579389). Similarly, *aal* genes occur in 28 *eae-* isolates and *fae* genes occur in 29 *eae-* isolates, while a single *eae+* isolate (SRR3573434) contains a copy of *aalF, aalH, faeE, faeF, faeC, faeI, faeD*.

Next, co-occurrence of groups of genes independent of intimin was determined. A co-occurrence matrix for a set of 32 genes was generated (Figure S1). Genes that occur either too rarely or which are very common were excluded, as were those that showed no significant co-occurrence. Several sets of genes were identified that showed a degree of co-occurrence between each other (Figure S1). The largest group consisted of *'ireA', 'kpsM', 'mchC', 'mchF', 'senB', 'shiA', 'sitA', 'subA', 'tia'*. These are all enriched in *eae-* isolates, and include three of the genes (*subA*, *kpsM* and *senB*) that are never found in *eae+* isolates. Similarly, adhesion-associated locus genes (*aal*) and fimbrial genes (*fae*) are found to co-occur, and are found more frequently in *eae-* isolates.

Another significant group consists of genes *cif*, *efa1*, *iucC*, *iutA*, *katP*. Three of these were observed above to co-occur with intimin, and therefore will co-occur with each other. Finally, there are pairs of genes which occur together, e.g. the Colicins B and M (*cba*, *cma*), as has been seen previously[47], which are almost always found adjacently on the same large conjugative plasmid. The genes *cba* and *cma* have intimin enrichment factors of 0.26 and 0.43 suggesting a slight enrichment in *eae-* isolates.

3.2.2 Intimin (*eae*) subtype mapping

Within the set of 85 (30%) isolates that are both *eae* and *stx* positive, several *eae* genetic subtypes can be distinguished (based on their best match upon BLAST searches against the reference *eae* sequences, section 2.2). The *E. coli* virulence factors in the virulence finder database (Section 2.2) include 20 subtypes of *eae*: alpha, beta, delta, epsilon, eta, gamma, iota, kappa, lambda, mu, nu, omicron, pi, rho, sigma, tau, theta, upsilon, xi and zeta. Of these, 10 *eae* subtypes were identified in the set of non-O157 *stx*+ isolates; namely alpha, beta, delta, epsilon, gamma, iota, rho, theta, xi and zeta.

The *eae*+ genomes were mapped to the *stx* gene subtypes in order to investigate if any *eae* subtype occurrence correlates with *stx* gene type (Figure 3A). 6/85 eae+ isolates possess both *stx1* and *stx2*, while 44 and 35 possess only *stx2* or *stx1*, respectively. Four *stx* gene subtypes, namely *stx1a, stx2c, stx2d* and *stx2f*, co-exist with the *eae* gene. *stx2* has been shown to be associated with severe disease phenotype[49,50] and it is interesting to note that in the non-O157 STEC dataset, *eae*-gamma and *eae*-iota co-occur exclusively with the *stx2c* subtype. stx2c is associated with six further *eae* subtypes, although in those cases not exclusively. *stx2d* which is reported to be very virulent in mice and associated with HUS[51] was seen co-occurring mainly with subtype *eae*-xi, but also with *eae*-epsilon. The less virulent form *stx1a* was present with 6 *eae* subtypes.

The *stx* subtype mapping for *eae*- isolates (n=201) shows that 21% (n=42) of the isolates possess both *stx1* and *stx2*, while 33% (n=93) and 46% (n=66) of the isolates possess only *stx1* or *stx2* respectively. Further, mapping these to *stx1* and *stx2* subtypes, it is interesting to note that some subtypes (*stx1c, stx2a, stx2b, stx2e, stx2g, stx2k and stx2l*) were exclusively seen in the *eae*- set (Figure 3B).

3.2.3 Evolution of the *eae* (intimin) gene

Across the set of 85 *stx*+/*eae*+ isolates, there are 31 unique translated intimin sequences, categorised into 10 subtypes. The diversity across the observed intimin sequences is displayed as a heatmap (Figure S2) generated using percent identity values between the protein sequences. The minimum percent identity is between the single iota sequence and two zeta subtype sequences (79.1%) whereas epsilon and xi sequences are most similar with ~97% identity (Figure S2). To understand the evolution of the *eae* gene and the variation within the different subtypes, the intimin subtypes were mapped onto phylogenetic trees. As intimin is encoded by a gene on the LEE island, originally acquired through horizontal gene transfer, we

sought to separate different evolutionary drivers by constructing phylogenies at the level of the core genome of *stx*+ isolates (n=286), the whole LEE island of *stx*+/*eae*+ isolates (n=85) and the individual translated intimin sequences (n=31).

The selected set of non-O157 STEC isolates (n=286) were observed to have a core genome (99% <= strains <= 100%) of only 3119 genes out of the total ~36k genes, with most genes in the shell and cloud regions. This agrees with the known observation of pathogenic *E. coli* genomes to be larger and more diverse than their commensal counterparts[52]. Virulence factors such as *eae*/*tir* are assigned to the cloud region, in accordance with the low level of occurrence (highlighted above). A phylogeny was obtained using the core genome of the isolates, and annotated with the *stx* and *eae* subtypes (Figure 4A). The distribution of *eae* across different clades is consistent with multiple acquisitions of the LEE island via horizontal gene transfer. The phylogeny shows reasonable clustering of the *eae* subtypes, although several subtypes are split across different branches which may be due to inaccuracies in the phylogeny or from recombination events.

The phylogeny obtained using the full-length LEE sequences from 85 *stx*+/*eae*+ isolates (Figure 4B) results in better clustering of the *eae* subtypes compared to the core genome phylogeny, confirming the fundamental role of genomic location. Further, it is interesting to note that the *eae* subtypes cluster even better in a phylogeny derived using only the LEE-located gene for the Type 3 secretion system Secretin (Figure 4C). This gene is essential to the function of T3SS, and its evolution is a good measure of that of the LEE island. Identifying the core genome of the LEE island using pangenome analysis (as previously applied to the whole bacterial genome), and constructing a phylogeny based on the set of core genes, produces similar results (Figure S3A). These phylogenies show that evolution of intimin subtypes largely follows evolution of the LEE island, except for some splitting of the beta and theta clades. Finally, a phylogeny was constructed from the 31 unique translated intimin sequences, reflecting only nonsynonymous variations (Figure 4D). This shows perfect separation of intimin subtypes, except for the inclusion of the single delta intimin sequence within the beta clade.

Next, the evolutionary mechanism of the intimin gene was investigated. While point mutations produce branches in a tree-like pattern, recombination events lead to networks with closed loops. Networks were generated using split decomposition analyses of the full length (Figure S3B), N terminus (1-700 amino acids, Figure S3C) and C-terminus (701-939 amino acids

including the D2 and D3 domains, Figure S3D) of the intimin sequences, using SplitsTree[40]. These network patterns indicate horizontal gene transfer through recombination is more prominent in the N-terminus of the intimin sequence. Split decomposition analysis using an alternative division into N-terminus (1-550) and C-terminus (551-939 including domains D0, D1, D2 and D3) regions showed a similar trend, with the network using the N-terminus region more reticulate than the C-terminus region (Figure S4).

3.3 Structural analysis of exemplar virulence factors: Intimin-Tir

3.3.1 Variations in Intimin

Allelic variation in *eae* was shown to be associated with STEC genotype, indicating a degree of clonality[53]. While there are no experimental structures of the full-length Intimin and Tir proteins, there are crystal or NMR structures of individual domains. Intimin is usually ~939 amino acids long and comprises a flexible periplasmic N-terminal region (up to ~188 amino acids), followed by the transmembrane beta-barrel domain (189-549) and the extracellular C-terminal region (550-939) (Figure 5A). The super-domain D2/D3 in the C-terminal region is responsible for binding the cognate receptor Tir and forming host attachment. We use residue numbering based on the UniProt reference sequences P19809 (Eae) and Q9KWH9 (Tir), which are represented by the structure PDB ID: 1F02[54]. This reference protein sequence for Intimin has 99.79% identity with the translated sequence of an alpha subtype *eae* annotated in Virulence Finder, although other reference alpha subtypes from Virulence Finder have lower identities down to 86% (illustrating variation within subtypes). The P19809 sequence also clusters with alpha subtype sequences within our set of isolates.

The 85 intimin protein sequences from our *stx+/eae+* isolates (comprising 31 unique sequences) were aligned to the UniProt sequence P19809 (subtype alpha), resulting in the identification of 706 point variations. These were mapped onto the domains using the following domain boundaries: 1-39, 65-115, 189-430, 450-550, 559-653, 658-751, 755-842, 843-939, corresponding to the SP (Signal Peptide), LysM, barrel, D00, D0, D1, D2 and D3 domains respectively (Figure 5A). Overall 661/706 variations lie within these domains, with the rest present in linker regions, and the majority of these (78%) lie in the exposed domains D1, D2 and D3 (Figure 5B).

Residues that contribute to the interface with Tir are different in our isolates to those observed in the P19809 reference sequence (PDB ID: 1F02), with a limited number of variant types permitted (Figure 5C, Table 2). Residue 926 is mutated to a valine in all 85 *eae*+ isolates, including those identified as alpha subtype, suggesting a global shift rather than variability. On the other hand, the neighbouring residue 927 shows 8 residue types (including the reference Lys) suggesting more variability at this site. Some of these variants involve a change in the amino acid type and side chain length such as (I897T, I897S, T895I), whereas others involve more subtle physicochemical changes such as I926V, K919R, Y934F, R850K. The variations at the interface are investigated in more detail below (section 3.3.4).

Table 2: Amino acid variations observed in the interface residues in Eae sequences from the dataset of *stx*+/*eae*+ isolates. Residues are ordered according to their observed variability, namely the number of distinct variations with respect to the reference sequence P19809.

Reference residue (1F02)	Types of variation	Number of isolates with ref residue mutated	Mutated residue(s)
K908	7	85	S, R, T, Q, N, I, P
K927	7	82	G, D, H, T, E, S, N
N924	6	82	P,T, I,A, Q, -
E930	6	85	K, A, T, N, D, S
Q894	5	74	K, L, R, E, N
Q905	5	72	D,A,V, K, S
S909	4	42	Q, G, E, K
N925	4	49	G, S, D, R
L923	4	38	V, R, H, I
I897	3	72	T, S, N
K919	3	70	T, R, A
T895	3	55	S, A, I

R850	3	50	K, Q, P
S929	3	85	K, N, T
I896	2	31	L, M
I926	1	85	V
Y934	1	44	F
V911	1	26	W
A933	1	23	V
D916	1	12	N

3.3.2 Variations in Tir (translocated intimin receptor) protein sequence

The cognate receptor of intimin, Tir, is known to be rich in disordered regions, with flexible intracellular domains (*i.e.* N-tir and C-tir) that allow it to engage with a variety of host proteins[55]. Tir binds Intimin via its extracellular domain (residues 271-336, using the sequence numbering of PDB ID: 1F02) with much of the interaction mediated by a beta-hairpin (residues 296-315)[56].

The extracellular domain of Tir was observed to be more conserved than the Tir binding region of Eae. There were 49 variant types of which 15 types are in the beta hairpin of Tir located at 9 distinct sites (Figure 5D). All of these variations in beta hairpin were present in less than 50% of the genomes except for the A303E which was observed in nearly all (84) of the eae+ sequences (Figure 5D).

3.3.3. Structural mapping of Eae variations

706 variant types identified in our set of eae+ isolates (section 3.3.1), map to 320 residues. AlphaFold3 was used to model the full length structure of Eae (UniProt ID:P19809, PDBID: 1F02) and these 320 residues (in black sticks) were mapped onto the modelled structure (Figure S5A). The N-terminal region is modeled with low confidence as indicated by poor pLDDT scores (<50 or between 50-70). Additionally, higher PAE (predicted aligned error) scores (Figure S5B) suggests that AlphaFold3 is uncertain about the relative positioning of the

domains (except for D1,D2 and D3 domains) in the modeled structure. The mapping of observed mutations suggests most of the variations are in the extracellular regions of Eae. This poor quality, low confidence structure is only used to map the sequence variations onto the structure and this low quality structure is not used for drawing any functional or structural implication of these variations.

The reference crystal structure (PDB ID:1F02) corresponds to domains D1, D2 and D3 (residues 658-939) of Eae along with the Eae binding region of Tir (residues 271-336), and we mapped the point variations observed in these regions onto this structure. Figure 5E shows the crystal structure, with Eae coloured according to the frequency of mutation with respect to the sequence of the crystal structure (P19809). The most frequently mutated residues (observed in > 80 isolates) are shown in gold ball-and-stick representation. As discussed above, these mutations consist of one or more mutation types. Of these high frequency mutations, 55% were observed in the D3 domain (Tir binding domain). The variations in the interface residues of Tir (residues 296-315) are shown as blue sticks (Figure 5E).

3.3.4 Structural modelling of intimin subtypes

In order to investigate the functional consequences of sequence variation, the structures of Intimin-Tir complexes were modelled using AlphaFold3[27]. As noted previously (section 3.2.2) there are ten *eae* genetic subtypes represented in our set of *stx*+/*eae*+ isolates. Therefore, ten structural models were generated corresponding to these *eae* subtypes in complex with their cognate Tir partner. The quality of the modelled complexes is indicated by the ipTM (interface Predicted TM) score, which measures the accuracy of the predicted positions of the subunits in the modelled complex. ipTM scores greater than 0.8 are predictions with high confidence and predictions with ipTM <0.6 are failed predictions[27]. All the modelled complexes except rho and theta have high confidence predictions for the relative subunit positions (Table 3). Therefore for rho and theta subtypes, homology modelling (Modeller v 9.10) was used to generate the Eae-Tir structural complex using PDB structure 1F02 as a template.

Furthermore, the quality of the Eae-Tir interface of these modelled complexes was assessed using an in-house deep learning based method (DPI-score, Bhujel *et al.*, manuscript under preparation). All the modelled complexes were predicted to have a high quality score for the protein-protein interface (Table 3).

Table 3: Confidence scores for the Eae-Tir modelled complexes. All the complexes except for rho and theta were modelled using AlphaFold3 (AF3). For the subtypes rho and theta, homology modelling was used to model the complex using PDB structure 1F02 as a template. *Interface scores were obtained for AF3 models for all subtypes except rho and theta, for which it is calculated on the homology models generated using Modeller.

Intimin Subtype	ipTM score (AF3 models)	Interface score
gamma	0.88	0.98
alpha	0.83	0.98
beta	0.82	0.96
delta	0.79	0.98
epsilon	0.81	0.98
iota	0.81	0.99
rho	0.34	0.99*
theta	0.17	0.99*
xi	0.83	0.98
zeta	0.83	0.99

Upon validation of the quality of modelled complexes, the effects of genomic variations that were observed at the interface between Intimin and Tir were investigated. The variants present at the interface between Intimin and its cognate Tir were assessed in order to elucidate the gain of interactions required to preserve the receptor binding and maintain function. Some of these co-occurring variations (including two, E312 and N300, of three Eae binding hotspots as listed in Ross and Miller, 2007[56]) and their impacts are highlighted here.

E312V(Tir):

E312(Tir) is a known crucial interface residue (PDB ID: 1F02, Figure 6A) and is identified as one of the intimin binding hotspot residues[56]. In the reference crystal structure, it is observed to interact with K927(Eae) and K298(Tir). In the zeta, alpha and gamma intimin subtypes in our set of non-O157 *eae*+ isolates, their cognate Tir protein sequences were observed to have the mutation E312V (Figures 6B, 6C and 6D respectively). E312V results in the loss of these interactions due to the non-polar side chain of valine.

However, to counter the effects of this loss of interactions, K927D, K927E and K927N were observed in the zeta, alpha and gamma *eae* subtypes respectively (Figure 6B-D). The neighboring residue, K298(Tir), is a conserved residue and plays a role in preserving the Eae-Tir interactions in the *eae* subtypes. Additionally, S929(Eae) was observed to be mutated to S929N(Eae) in these three subtypes, which help to preserve the interactions with the conserved residue K298. However, out of the three variants, gamma has the least number of compensatory interactions that involve K298 (Figure 6D), although there may be compensatory changes at other interaction sites.

N300(Tir)

N300(Tir) is another Eae intimin binding hotspot[56] and may form a hydrogen-bond with E930 in the crystal structure (PDB ID: 1F02, Figure 7A). N300(Tir) was observed to be mutated to D, I or S in the isolates having *eae* subtypes zeta, rho and theta respectively. Upon structural inspection, N300D appears to form a salt bridge with E930K in the zeta subtypes (Figure 7B). This pair of co-occurring mutations shows a reversal of residue type, with the glutamic acid of 930(Eae) being replaced by the aspartic acid of 300(Tir), and illustrates the compensation required to maintain function, that is Eae-Tir binding.

The structural models for rho and theta produced using AlphaFold 3 predicted a different interaction site between Eae and Tir than the one observed in the crystal structure (1F02, Figures S6A and S6B). To try to obtain the expected quaternary structure, we modelled the complex with AlphaFold 3 including only the D3 region of the Eae subtype sequence together with the Tir extracellular domain. However, superposition of the predicted complex structure with the crystal structure showed many changes in the secondary structure of Eae which are probably artefacts due to the loss of D2 (Figure S6C and S6D). Therefore, it was necessary to resort to the traditional approach of homology modelling using Modeller[41] (v10.4) and the crystal structure as template.

These homology models were used to examine the variations in the rho and theta subtypes. There did not appear to be any major impacts of N300S in the theta subtype on the interface between Eae and Tir (Figure 7C), despite the loss of a specific interaction with Eae. E930A is the co-occurring mutation observed in its Eae sequence. For the rho model, no gains of interaction were observed for N300I(Tir). However, the co-occurring mutation E930S(Eae) may form a H-bond with N932 (Eae) to counter the effect of variations and preserve the structure and function (Figure 7D).

4. Discussion

In this study, a dataset comprising 286 non-O157 STEC genomes sourced from EnteroBase was curated and analysed to determine whether the virulence factor complement coupled with allelic variation could be used to infer their role in pathogenicity. We used the premise that genetic variation that results in structural changes impacts protein function and potentially their contribution to virulence. A selection of isolates was made to represent a diverse range of serotypes, with highly prevalent and well-characterised serotypes such as O26 excluded to minimise bias. The presence of the *stx* toxin gene was used to define these isolates as STEC. The genomic diversity of the set of non-O157 STEC isolates is illustrated by the fact that the genomes share only 3119 core genes out of a pangenome of ~36k genes.

The study took a two-step approach, to first investigate the individual and co-occurrence of genetic factors that are thought to contribute to disease, defined as Virulence Factors[24,25]. Secondly, since the majority of virulence factors exist in different allelic variants, the potential functional impact of variation was investigated, focusing on the well characterised STEC virulence factor Intimin, Eae and its cognate partner, Tir. Variations were considered both at the subtype level and at the level of individual point mutations. This serves as an example for other protein interactions of virulence factor variants, and potential impacts on overall pathogenicity of non-O157 STEC isolates.

The non-O157 STEC are characterised by their genetic diversity. Using serotype alone as a distinguishing feature identified 111 variants in our dataset. Like other members of the *E. coli* species, it is evident that the STEC sub-species comprise mosaic genomes, with multiple instances of recombination, genomic rearrangements and horizontally acquired elements, making classification challenging. The decision tree from the JEMRA risk matrix is based on the *stx* gene subtype and the presence of the intimin gene, *eae*, and the transcriptional activator of aggregative adherence fimbria expression, *aggR*. However, the intimin gene is not ubiquitous in clinical samples and was only present in 85 / 286 (30 %) *stx*+ non-O157 STEC genomes investigated. Furthermore, *aggR* is not widespread in EHEC and does not occur at all in our *stx*+ set.

Some serotypes of STEC that lack initimin are notable as they still cause severe disease[57]. STEC O91 is reported to be one of the most common *eae*- non-O157 serotypes associated with human disease[58]. O91 and O113 have been reported as prevalent serotypes (top 9) in England

in 2019[59], and are represented in our dataset. STEC O55 outbreaks have also been reported in England[41], represented by nine *eae*+ isolates and two *eae*- isolates in the dataset .

Besides *stx*, and *eae*, additional virulence factors are associated with pathogenicity of non-O157 STECs, and can contribute to severe disease phenotypes such as bloody diarrhoea and HUS. Genes that help with adherence, colonisation, production and secretion of toxins are known to be associated with causing severe disease. Within our dataset, there was a large degree of variation in the complement of virulence factors, with 130 different types of virulence factors across the set. Some factors are present in most isolates (e.g. *nlpI, terC, yeh, csgA*), while others are only represented in single isolates, such as *csm* and *fot*. The virulence factors that occurred frequently are known to be abundant in *E. coli* genomes[60]. They can be considered as factors that facilitate host colonisation or adhesion to a surface, and in doing so contribute to pathogenicity and lead to disease.

The co-occurrence of some pairs of virulence factors was evident, while other pairs were apparently mutually exclusive in the set of isolates. Several different sets of genes co-occurred, with the largest group encompassing a range of functions associated with virulence, including: *ireA*, a siderophore associated with pathology in UTI[61] ; capsular polysaccharide production genes, *kps*[62]; microcin genes, *mchC* and *F,* associated with virulence in the STEC O91 serogroup[63]; the *Shigella* enterotoxin gene *sen,* also associated with bacteraemic *E. coli* isolates[64]; *shiA,* responsible for immunomodulatory functions in *Shigella* species[65], with homologs apparently with similar functions in uropathogenic *E. coli*[66]; the subtilis cytotoxin gene *subA*, associated with STEC[67] including isolates from cattle[68]; and the associated *tia* gene, shown to be involved in cellular invasion of some STEC strains[69]. Notably, the latter two virulence factors are encoded on the subtilase-encoding pathogenicity island (SE-PAI) and are mutually exclusive with the locus for enterocyte effacement[70], presumably because of the role in invasion as opposed to A/E lesion formation.

Co-occurrence of virulence factors with the initimin gene, *eae*, was quantified to obtain a clearer indication of potentially interacting networks based on this key virulence factor. 52 % and 41 % of the *eae*+ isolates were *stx2* and *stx1*, respectively, with only 7 % occurring with both *stx*1+ and *stx*2+. As expected, some virulence factors are co-located with intimin on the LEE island e.g. *tir*, *espA, espB, espF*, although not all. This is in-line with the concept of a network of T3SS effectors, which exhibit co-dependency and context-dependent functionality[71]. Virulence factors

that exhibited strong co-occurrence with *eae* include the non-LEE encoded T3SS effectors *nleA/B/C* and *cif,* and the plasmid-encoded genes *esp, tox* and *katP*.

Several of the genes identified in co-occurrence networks of virulence factors were first identified on the pO157 plasmid, which was genetically characterised in serotype O157:H7, strain Sakai[72]. It encodes genes for a type 2 secretion system and the alpha-hemolysin gene cluster, *hyl*[73],as well as its associated protease, *katP*. *katP* has been detected in multiple non-O157 isolates on other plasmids, e.g. in O111[74] and O26[75], with sub-type a in particular associated with clinical disease[76]. It is normally linked with other plasmid encoded virulence factors: the heamolysin gene cluster (*hyl*), a lipid A modification system (*ecf* operon) and the serine protease (*espP*), although they do not always co-occur[74]. The *katP* gene is normally associated with LEE-encoding STEC strains, and not in non-LEE strains, presumably down to genetic incompatibility[77]. The KatP protease is responsible for cleavage and inactivation of Hyl heamolysin. STEC isolates frequently encode multiple hemolysins, of which four have been reported: alpha-hemolysin (HylA), enterohemolysin (ExhA) also encoded on a plasmid or phage (e-HylA) and silent hemolysin / cytolysin A (SheA)[78]. Both HylA and ExhA were detected in the STEC genome dataset. Hemolysins are encoded in other pathotypes of *E. coli*, although there has been shown to be a strong association between both ExhA and HylA and presence of *stx* genes and to some extent, with presence of *eae*[78].

Using reference sequences taken from the Virulence Finder database, ten *eae* genetic subtypes were identified across the dataset of 85 isolates. Multiple subtypes of the *eae* gene have been reported previously, and correlated with host specificity and tissue tropism[79]. The *eae* gene was initially sequenced from EPEC lineage 1 strain E2348/69, forming the reference sequence for α-intimin, while EPEC lineage 2 strains encode subtype β-intimin. Further subtypes are defined for EHEC, such as γ-intimin from O157:H7 and O55:H7 strains[80]. Divergent intimins appear to be encoded by mosaic alleles composed of DNA segments with different evolutionary histories[81].

Variation in intimin subtypes has long been recognised[81], with at least 18 different genetic variants identified in a single non-O157 STEC genome set[79]. The *eae* variants identified here showed significant identity amongst themselves, with the epsilon and xi subtypes sharing ~97% identity. Recombination events have been implicated in the evolution of *eae*, and split decomposition analysis on this dataset suggested a reticulate network indicative of recombination. The genomic variations (non-synonymous) in intimin subtypes were inspected using the sequence of *eae*-α as reference. ~700 variations were observed in all intimin protein

sequences and 78% of these lie in the extracellular region of intimin (domains D1, D2 and D3), which interacts with the receptor Tir. However, the extracellular region of Tir was observed to be more conserved than the extracellular region of intimin.

Protein structural modelling enabled a better understanding of the functional implications of the genomic variations. AlphaFold3[27] was used to generate the three-dimensional structural models of the intimin subtypes with their cognate Tir receptor, validated with an in-house machine learning based method. With AlphaFold3, monomeric protein structures are modelled with high accuracy although protein-protein complexes modelling is still a developing analytical tool[82]. As such, three-dimensional structure modelled complexes need to be validated using confidence metrics such as pLDDT (per atom confidence score), ipTM (confidence in relative placement of subunits in a modelled complex). Although the structures are models and the atom positions are approximate, it nonetheless still provides a good estimate of interactions at the interface.

Mapping of Eae variations onto structural models shows that most occur in the extra-cellular domains, and of these most are in the interaction region with the cognate partner Tir. Mutations in the extra-cellular region may be associated with immune escape, but nevertheless binding with Tir must be maintained in order to retain effective pedestal formation. For some specific examples, we have shown that variations preserve the functional interaction between the two protein partners through a compensatory mechanism, wherein loss of some interactions at the interface is compensated by the formation of new interfacial contacts. Understanding the importance of specific interactions involved in receptor binding will help improve the interpretation of STEC genomic variants and may aid the development of interventions.

Through this work, we would like to emphasise the importance of integrating genomic analysis with protein structural analysis which provides insights into mechanism of function, in this case pathogenicity of eae+ STEC isolates. The examples of co-occurring variations at protein-protein interfaces underscore the importance of integrating genomic data with structural biology to deepen our understanding of protein function.

Funding

This work was supported by STFC Food Safety Network+ (award R24686 to NH and MW), Food Safety Research Network (award 43266FSRN-2023S12 to NH), and AIBIO-UK ((22-AIBN) [BB/Y006933/1]) Pilot funding, round1 (to SM).

Author Contributions

Sony Malhotra- Conceptualization, Formal analysis, Funding acquisition, Methodology, Writing-original draft, Writing- review and editing
Ashley Ward- Data curation
Lauren Giles- Data curation, Writing- review and editing
Tom Gerrard- Formal Analysis
Martyn Winn- Conceptualization, Formal analysis, Funding acquisition, Methodology, Writing-original draft, Writing- review and editing
Nicola Holden- Conceptualization, Data curation, Funding acquisition, Methodology, Project Administration, Writing-original draft, Writing- review and editing

Conflict of Interest

None

References

1. Shiga toxin-producing Escherichia coli (STEC) data: 2024 - GOV.UK. https://www.gov.uk/government/publications/escherichia-coli-e-coli-o157-annual-totals/shiga-toxin-producing-escherichia-coli-stec-data-2024.
2. Shiga toxin-producing Escherichia coli (STEC) data: 2023. *GOV.UK* https://www.gov.uk/government/publications/escherichia-coli-e-coli-o157-annual-totals/shiga-toxin-producing-escherichia-coli-stec-data-2023.
3. King, G. *et al.* Epidemiology of Shiga toxin-producing Escherichia coli other than serotype O157:H7 in England, 2016–2023. *Journal of Medical Microbiology* **74**, 001947 (2025).
4. Panel, E. B. *et al.* Pathogenicity assessment of Shiga toxin-producing Escherichia coli (STEC) and the public health risk posed by contamination of food with STEC. *EFSA Journal* **18**, e05967 (2020).
5. Shiga toxin-producing Escherichia coli (STEC) and food: attribution, characterization, and monitoring: report. https://www.who.int/publications/i/item/9789241514279.
6. FAO/WHO STEC EXPERT GROUP. Hazard Identification and Characterization: Criteria for Categorizing Shiga Toxin-Producing Escherichia coli on a Risk Basis†. *J Food Prot* **82**, 7–21 (2019).
7. Rauw, K. D., Buyl, R., Jacquinet, S. & Piérard, D. Risk determinants for the development of typical haemolytic uremic syndrome in Belgium and proposition of a new virulence typing algorithm for Shiga toxin-producing Escherichia coli. *Epidemiology & Infection* **147**, e6 (2019).
8. Fuller, C. A., Pellino, C. A., Flagler, M. J., Strasser, J. E. & Weiss, A. A. Shiga Toxin Subtypes Display Dramatic Differences in Potency. *Infection and Immunity* **79**, 1329–1337 (2011).
9. Gallegos, K. M. *et al.* Shiga Toxin Binding to Glycolipids and Glycans. *PLoS One* **7**, e30368 (2012).

10. Goffaux, F., China, B., Janssen, L., Pirson, V. & Mainil, J. The locus for enterocyte effacement (LEE) of enteropathogenic Escherichia coli (EPEC) from dogs and cats. *Adv Exp Med Biol* **473**, 129–136 (1999).

11. IRSHAD, H., COOKSON, A. L., PRATTLEY, D. J., DUFOUR, M. & FRENCH, N. P. Distribution of Escherichia coli strains harbouring Shiga toxin-producing E. coli (STEC)-associated virulence factors (stx1, stx2, eae, ehxA) from very young calves in the North Island of New Zealand. *Epidemiol Infect* **142**, 2548–2558 (2014).

12. Rodwell, E. V. *et al.* Epidemiology and genomic analysis of Shiga toxin-producing Escherichia coli clonal complex 165 in the UK. *J Med Microbiol* **70**, 001471 (2021).

13. Slater, S. L., Sågfors, A. M., Pollard, D. J., Ruano-Gallego, D. & Frankel, G. The Type III Secretion System of Pathogenic Escherichia coli. *Curr Top Microbiol Immunol* **416**, 51–72 (2018).

14. Gauthier, A., de Grado, M. & Finlay, B. B. Mechanical fractionation reveals structural requirements for enteropathogenic Escherichia coli Tir insertion into host membranes. *Infect Immun* **68**, 4344–4348 (2000).

15. Battle, S. E., Brady, M. J., Vanaja, S. K., Leong, J. M. & Hecht, G. A. Actin Pedestal Formation by Enterohemorrhagic Escherichia coli Enhances Bacterial Host Cell Attachment and Concomitant Type III Translocation. *Infection and Immunity* **82**, 3713–3722 (2014).

16. Croxen, M. A. & Finlay, B. B. Molecular mechanisms of Escherichia coli pathogenicity. *Nat Rev Microbiol* **8**, 26–38 (2010).

17. Paganini, J. A. *et al.* Predicting clinical outcome of Escherichia coli O157:H7 infections using explainable Machine Learning. 2025.06.05.25329036 Preprint at https://doi.org/10.1101/2025.06.05.25329036 (2025).

18. Shiga toxin-producing Escherichia coli (STEC) data: 2022. *GOV.UK* https://www.gov.uk/government/publications/escherichia-coli-e-coli-o157-annual-totals/shiga-toxin-producing-escherichia-coli-stec-data-2022.

19. Gastrointestinal and Zoonoses - Biennial report 2020 to 2021 - Gastrointestinal and Zoonoses - Publications - Public Health Scotland. https://publichealthscotland.scot/publications/gastrointestinal-and-zoonoses/gastrointestinal-and-zoonoses-biennial-report-2020-to-2021/.

20. Data and surveillance - Shiga toxin-producing Escherichia coli (STEC) - Infectious diseases - Health protection - Population health - Public Health Scotland. https://publichealthscotland.scot/population-health/health-protection/infectious-diseases/shiga-toxin-producing-escherichia-coli-stec/data-and-surveillance/.

21. Dyer, N. P. *et al.* EnteroBase in 2025: exploring the genomic epidemiology of bacterial pathogens. *Nucleic Acids Research* **53**, D757–D762 (2025).

22. Ishii, S., Meyer, K. P. & Sadowsky, M. J. Relationship between Phylogenetic Groups, Genotypic Clusters, and Virulence Gene Profiles of Escherichia coli Strains from Diverse Human and Animal Sources. *Applied and Environmental Microbiology* **73**, 5703–5710 (2007).

23. Bhattacharyya, A., Panday, S. & Ussery, D. Rapid assessment of clinical severity for salmonellosis cases via protein family domain analysis and machine learning. *Academia Molecular Biology and Genomics* **2**, (2025).

24. Malberg Tetzschner, A. M., Johnson, J. R., Johnston, B. D., Lund, O. & Scheutz, F. In Silico Genotyping of Escherichia coli Isolates for Extraintestinal Virulence Genes by Use of Whole-Genome Sequencing Data. *Journal of Clinical Microbiology* **58**, 10.1128/jcm.01269-20 (2020).

25. Joensen, K. G. *et al.* Real-Time Whole-Genome Sequencing for Routine Typing, Surveillance, and Outbreak Detection of Verotoxigenic Escherichia coli. *Journal of Clinical Microbiology* **52**, 1501–1510 (2014).

26. Roe, A. J., Hoey, D. E. E. & Gally, D. L. Regulation, secretion and activity of type III-secreted proteins of enterohaemorrhagic Escherichia coli O157. *Biochemical Society Transactions* **31**, 98–103 (2003).

27. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

28. Leinonen, R., Sugawara, H. & Shumway, M. The Sequence Read Archive. *Nucleic Acids Res* **39**, D19–D21 (2011).

29. BBMap. *SourceForge* https://sourceforge.net/projects/bbmap/ (2025).

30. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

31. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

32. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A. & Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr Protoc Bioinformatics* **70**, e102 (2020).

33. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).

34. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

35. Gertz, E. M., Yu, Y.-K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biology* **4**, 41 (2006).

36. Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

37. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

38. Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology* **21**, 180 (2020).

39. Moore, R. M., Harrison, A. O., McAllister, S. M., Polson, S. W. & Wommack, K. E. Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ* **8**, e8584 (2020).

40. Huson, D. H. & Bryant, D. The SplitsTree App: interactive analysis and visualization using phylogenetic trees and networks. *Nat Methods* **21**, 1773–1774 (2024).

41. Sali, A. Comparative protein modeling by satisfaction of spatial restraints. *Mol Med Today* **1**, 270–277 (1995).

42. Panel, E. B. *et al.* Pathogenicity assessment of Shiga toxin-producing Escherichia coli (STEC) and the public health risk posed by contamination of food with STEC. *EFSA Journal* **18**, e05967 (2020).

43. Bugarel, M., Beutin, L. & Fach, P. Low-Density Macroarray Targeting Non-Locus of Enterocyte Effacement Effectors (nle Genes) and Major Virulence Factors of Shiga Toxin-Producing Escherichia coli (STEC): a New Approach for Molecular Risk Assessment of STEC Isolates. *Appl Environ Microbiol* **76**, 203–211 (2010).

44. Coombes, B. K. *et al.* Molecular Analysis as an Aid To Assess the Public Health Risk of Non-O157 Shiga Toxin-Producing Escherichia coli Strains. *Appl Environ Microbiol* **74**, 2153–2160 (2008).

45. Loukiadis, E. *et al.* Distribution, functional expression, and genetic organization of Cif, a phage-encoded type III-secreted effector from enteropathogenic and enterohemorrhagic Escherichia coli. *J Bacteriol* **190**, 275–285 (2008).

46. Tsutsuki, H. *et al.* Subtilase cytotoxin from Shiga-toxigenic *Escherichia coli* impairs the inflammasome and exacerbates enteropathogenic bacterial infection. *iScience* **25**, 104050 (2022).

47. Navarro-Garcia, F. *et al.* Pic, an Autotransporter Protein Secreted by Different Pathogens in the Enterobacteriaceae Family, Is a Potent Mucus Secretagogue. *Infect Immun* **78**, 4101–4109 (2010).

48. Christenson, J. K. & Gordon, D. M. Evolution of colicin BM plasmids: the loss of the colicin B activity gene. *Microbiology* **155**, 1645–1655 (2009).

49. Boerlin, P. *et al.* Associations between virulence factors of Shiga toxin-producing

Escherichia coli and disease in humans. *J Clin Microbiol* **37**, 497–503 (1999).

50. Werber, D. *et al.* Strong association between shiga toxin-producing Escherichia coli O157 and virulence genes stx2 and eae as possible explanation for predominance of serogroup O157 in patients with haemolytic uraemic syndrome. *Eur J Clin Microbiol Infect Dis* **22**, 726–730 (2003).

51. Melton-Celsa, A. R., Kokai-Kun, J. F. & O'Brien, A. D. Activation of Shiga toxin type 2d (Stx2d) by elastase involves cleavage of the C-terminal two amino acids of the A2 peptide in the context of the appropriate B pentamer. *Molecular Microbiology* **43**, 207–215 (2002).

52. Croxen, M. A. & Finlay, B. B. Molecular mechanisms of Escherichia coli pathogenicity. *Nat Rev Microbiol* **8**, 26–38 (2010).

53. Creuzburg, K. *et al.* Evolutionary analysis and distribution of type III effector genes in pathogenic Escherichia coli from human, animal and food sources. *Environ Microbiol* **13**, 439–452 (2011).

54. Luo, Y. *et al.* Crystal structure of enteropathogenic Escherichia coli intimin-receptor complex. *Nature* **405**, 1073–1077 (2000).

55. Vieira, M. F. M. *et al.* The pathogen-encoded signalling receptor Tir exploits host-like intrinsic disorder for infection. *Commun Biol* **7**, 1–21 (2024).

56. Ross, N. T. & Miller, B. L. Characterization of the binding surface of the translocated intimin receptor, an essential protein for EPEC and EHEC cell adhesion. *Protein Sci* **16**, 2677–2683 (2007).

57. Feng, P., Delannoy, S., Lacher, D. W., Bosilevac, J. M. & Fach, P. Characterization and Virulence Potential of Serogroup O113 Shiga Toxin–Producing *Escherichia coli* Strains Isolated from Beef and Cattle in the United States. *Journal of Food Protection* **80**, 383–391 (2017).

58. Bielaszewska, M. *et al.* Shiga Toxin, Cytolethal Distending Toxin, and Hemolysin Repertoires in Clinical Escherichia coli O91 Isolates. *Journal of Clinical Microbiology* **47**,

2061 (2009).

59. Shiga toxin-producing Escherichia coli (STEC) data: 2019. *GOV.UK* https://www.gov.uk/government/publications/escherichia-coli-e-coli-o157-annual-totals/shiga-toxin-producing-escherichia-coli-stec-data-2019.

60. Touchon, M. *et al.* Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLOS Genetics* **5**, e1000344 (2009).

61. Hagan, E. C. & Mobley, H. L. T. Uropathogenic Escherichia coli Outer Membrane Antigens Expressed during Urinary Tract Infection. *Infection and Immunity* **75**, 3941–3949 (2007).

62. Russo, T. A., Wenderoth, S., Carlino, U. B., Merrick, J. M. & Lesse, A. J. Identification, Genomic Organization, and Analysis of the Group III Capsular Polysaccharide Genes kpsD, kpsM,kpsT, and kpsE from an Extraintestinal Isolate of Escherichia coli (CP9, O4/K54/H5). *Journal of Bacteriology* **180**, 338–349 (1998).

63. Nüesch-Inderbinen, M. *et al.* Distribution of virulence factors, antimicrobial resistance genes and phylogenetic relatedness among Shiga toxin-producing Escherichia coli serogroup O91 from human infections. *Int J Med Microbiol* **311**, 151541 (2021).

64. Telli, M. *et al.* Prevalence of enterotoxins among Escherichia coli isolates causing bacteraemia. *FEMS Microbiol Lett* **306**, 117–121 (2010).

65. Ingersoll, M. A. *et al.* The ShiA protein encoded by the Shigella flexneri SHI-2 pathogenicity island attenuates inflammation. *Cellular Microbiology* **5**, 797–807 (2003).

66. Lloyd, A. L., Smith, S. N., Eaton, K. A. & Mobley, H. L. T. Uropathogenic Escherichia coli Suppresses the Host Inflammatory Response via Pathogenicity Island Genes sisA and sisB. *Infection and Immunity* **77**, 5322–5333 (2009).

67. Paton, A. W., Srimanote, P., Talbot, U. M., Wang, H. & Paton, J. C. A New Family of Potent AB5 Cytotoxins Produced by Shiga Toxigenic Escherichia coli. *J Exp Med* **200**, 35–46 (2004).

68. Cáceres, M. E., Etcheverría, A. I., Fernández, D., Rodríguez, E. M. & Padola, N. L. Variation in the Distribution of Putative Virulence and Colonization Factors in Shiga Toxin-Producing Escherichia coli Isolated from Different Categories of Cattle. *Front. Cell. Infect. Microbiol.* **7**, (2017).

69. Bondì, R. *et al.* The Gene tia, Harbored by the Subtilase-Encoding Pathogenicity Island, Is Involved in the Ability of Locus of Enterocyte Effacement-Negative Shiga Toxin-Producing Escherichia coli Strains To Invade Monolayers of Epithelial Cells. *Infection and Immunity* **85**, 10.1128/iai.00613-17 (2017).

70. Michelacci, V. *et al.* A new pathogenicity island carrying an allelic variant of the Subtilase cytotoxin is common among Shiga toxin producing Escherichia coli of human and ovine origin. *Clinical Microbiology and Infection* **19**, E149–E156 (2013).

71. Sanchez-Garrido, J., Ruano-Gallego, D., Choudhary, J. S. & Frankel, G. The type III secretion system effector network hypothesis. *Trends Microbiol* **30**, 524–533 (2022).

72. Makino, K. *et al.* Complete Nucleotide Sequences of 93-kb and 3.3-kb Plasmids of an Enterohemorrhagic Escherichia coli O157:H7 Derived from Sakai Outbreak. *DNA Res* **5**, 1–9 (1998).

73. Brockmeyer, J. *et al.* Enterohaemorrhagic Escherichia coli haemolysin is cleaved and inactivated by serine protease EspPα. *Environmental Microbiology* **13**, 1327–1341 (2011).

74. Ogura, Y. *et al.* Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. *Proceedings of the National Academy of Sciences* **106**, 17939–17944 (2009).

75. Fratamico, P. M. *et al.* The complete DNA sequence and analysis of the virulence plasmid and of five additional plasmids carried by Shiga toxin-producing *Escherichia coli* O26:H11 strain H30. *International Journal of Medical Microbiology* **301**, 192–203 (2011).

76. Khan, A. B. *et al.* Serine protease *espP* subtype *α*, but not *β* or *γ*, of Shiga toxin-producing *Escherichia coli* is associated with highly pathogenic serogroups. *International*

Journal of Medical Microbiology **299**, 247–254 (2009).

77. Newton, H. J. *et al.* Shiga Toxin–producing Escherichia coli Strains Negative for Locus of Enterocyte Effacement. *Emerg Infect Dis* **15**, 372–380 (2009).

78. Lorenz, S. C. *et al.* Prevalence of Hemolysin Genes and Comparison of ehxA Subtype Patterns in Shiga Toxin-Producing Escherichia coli (STEC) and Non-STEC Strains from Clinical, Food, and Animal Sources. *Applied and Environmental Microbiology* **79**, 6301–6311 (2013).

79. Yang, X. *et al.* Genetic diversity of the intimin gene (eae) in non-O157 Shiga toxin-producing Escherichia coli strains in China. *Sci Rep* **10**, 3275 (2020).

80. Fitzhenry, R. J. *et al.* Tissue Tropism of Enteropathogenic Escherichia coli Strains Belonging to the O55 Serogroup. *Infection and Immunity* **70**, 4362–4368 (2002).

81. McGraw, E. A., Li, J., Selander, R. K. & Whittam, T. S. Molecular evolution and mosaic structure of alpha, beta, and gamma intimins of pathogenic Escherichia coli. *Mol Biol Evol* **16**, 12–22 (1999).

82. Lu, D., Yu, S., Huang, Y. & Gong, X. Multimeric protein interaction and complex prediction: Structure, dynamics and function. *Comput Struct Biotechnol J* **27**, 1975–1997 (2025).

Figure Legends

Figure 1: Collection of non-O157 STEC isolates (n=286) and virulence factor mapping. **A**. Distribution of different serotypes in the collection of non-O157 STEC genomes from the clinical isolates obtained from the Enterobase. **B**. Presence of different types of *stx* in these serotypes. **C**. Histogram of number of virulence factors in non-O157 STEC genomes. **D**. Subtypes of *stx* toxin genes in non-O157 STEC genomes.

Figure 2: Virulence factors presence and co-occurence in the set of isolates. **A.** Most frequent virulence factors (present in at least 50% of the genomes). **B.** Co-occurrence of virulence factors in our set of 281 isolates. The diagonal gives the absolute number of isolates that a virulence factor occurs in. The off-diagonal entries give the number of isolates containing both virulence factors (corresponding to the row and the column), given as a fraction of the total number of isolates that the virulence factor on the row occurs in. Note that this definition means that the matrix is asymmetric.

Figure 3: Co-occurrence of *stx* and *eae* genes. **A**. Distribution of intimin subtypes in *stx* subtypes. **B**. Distribution of *stx* subtypes in intimin negatives isolates.

Figure 4: Phylogenetic tree for **A**. Core genome of 286 non-O157 STEC isolates, **B**. 85 *eae*+ LEE sequences. The LEE sequence from STEC O157 Sakai is labeled as 'LEE' in the phylogeny. **C.** Secretin C protein from the LEE of 85 *eae*+ isolates. Secretin C protein from the STEC O157 Sakai is labeled as 'LEE' in the phylogeny, and **D**. 31 unique *eae* sequences from *eae* +isolates (n=85). The leaf labels are colored according to the *eae* subtypes (eae negative:black, alpha:brown, beta:limegreen, gamma:orange, delta:magenta, rho:pink, iota:orange, delta:violet, epsilon:red, iota:wheat, theta:yellow, xi:cyan, zeta:royalblue), the bar length depicts the number of virulence factors present in the genome and the leaf node color is colored according to the Shiga Toxin subtype (*stx1*: grey, *stx2*: teal, *stx1* and *stx2*: brown).

Figure 5: Variations in the intimin amino acid sequences. **A**. Schematic of Eae domains, domain boundaries are marked. **B**. Number of variations in different intimin domains. **C**. Number and types of variations of interface residues present in Eae's D3. The number of Eae sequences with mutation is shown as orange bar and the height of blue bar denotes the types of variations observed for that particular interface residue. **D.** 15 types of variations in the beta hairpin region

of Tir, which is known to interact with Eae. **E.** Structural mapping of variations observed in Eae (salmon) and the Eae-binding region of Tir (green). The variations in Eae are colored according to their frequency; 1< frequency <=20 (brown), 21< frequency <=40 (gray), 41< frequency <=60 (teal), 61< frequency <=80 (yellow) and frequency >80 (golden sticks).

Figure 6: Structural inspection of interactions between Eae-Tir (Tir:E312 and Eae: K927 and K 928). **A**. As in the crystal structure, 1F02, the residue E312(tir) sits at the interface with Eae and interacts with K298 and K927. **B**. In the zeta subtype, E312V results in the loss of these interactions but simultaneously the gain of interactions is observed due to the other variations in *eae* namely K927D and S929N to preserve the interactions at the interface. **C.** In the alpha subtype,the co-occurring variations K927E and S929N in Eae were observed to gain interactions at the interface. **D**. In the gamma subtype, E312V, there were no notable gain in interactions.

Figure 7: Structural inspection of interactions between Eae-Tir (Tir: N300 and Eae: E930). **A**. N300 is seen to be in proximity to E930 and they may be involved in H-bond. **B.** In zeta subtype, this interaction is seen to be preserved by the formation of salt bridge between the residues N300D and E930K. **C**. In theta subtype, N300S and E930A are observed however no specific gain of interaction is observed upon the mutation. **D**. In the rho subtype, N300I is observed along with E930S. E930S may be forming a H-bond with N932 to maintain the structural stability of the Eae.

Supplementary Material

Figure S1: Co-occurrence of selected gene groups. Gene group '*aal*' includes 4 genes, '*fae*' includes 6 genes, and '*mch*' includes 2 genes, giving a total of 32 genes.

Figure S2: Pairwise percent identity between the protein sequences of 10 intimin subtypes plotted as a heatmap.

Figure S3: **A**. The phylogeny obtained using the aligned core genome sequences of the LEE obtained from the *eae*+ isolates (n = 85). Split decomposition network tree using **B**. full length intimin sequences, **C**. N-terminus region (1-700) and **D**. C-terminus region (701-939) of the intimin sequences.

Figure S4: Split decomposition network tree using **A**. N-terminus region (1-550) and **B**. C-terminus region of the intimin sequences.

Figure S5: Full length structural model of Intimin. **A**. Full length structural model for the UniProt sequence P19809 (Intimin) obtained using AlphaFold3. The sequence domains are colored: sp(yellow), LysM(orange), β-barrel(magenta), D0 and D00(blue), D1 (orange red), D2(cyan) and D3 (red). The point mutation sites are shown as black sticks. **B**. PAE (predicted aligned error) plot of the structural model obtained using AlphaFold3.

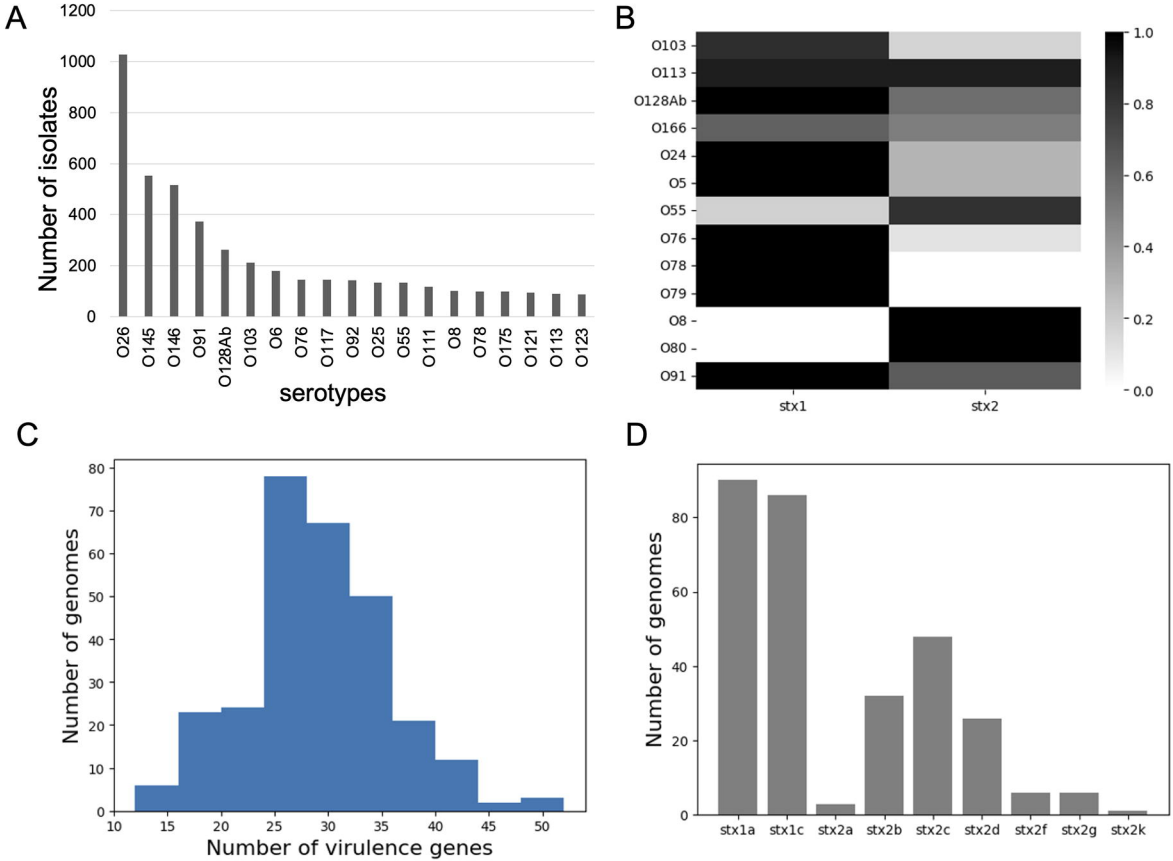
Figure S6: Different structure prediction approaches for Eae-Tir sequences. The crystal structure (PDB ID: 1F02) in all the panels is shown in green (Tir) and salmon (Eae). **A**. Structural model of the translated rho Eae-Tir (blue) complex modelled using AlphaFold3 is superposed over the crystal structure. The interface modelled for the rho subtype is different to the one observed in crystal structure. **B**. Structural model of only D3 region of rho subtype with its cognate Tir built using AlphaFold3. The modelled complex is in blue. **C.** The structural model of the theta Eae-Tir (yellow) complex modelled using AlphaFold3 is superposed over the crystal structure. The interface modelled for the theta subtype is different to the one observed in crystal structure. **D**. Structural model of only D3 region (yellow) of theta subtype with its cognate Tir built using AlphaFold3.

Table S1: Distribution of non-O157 STEC serotypes in the human isolates in the UK obtained from Enterobase.

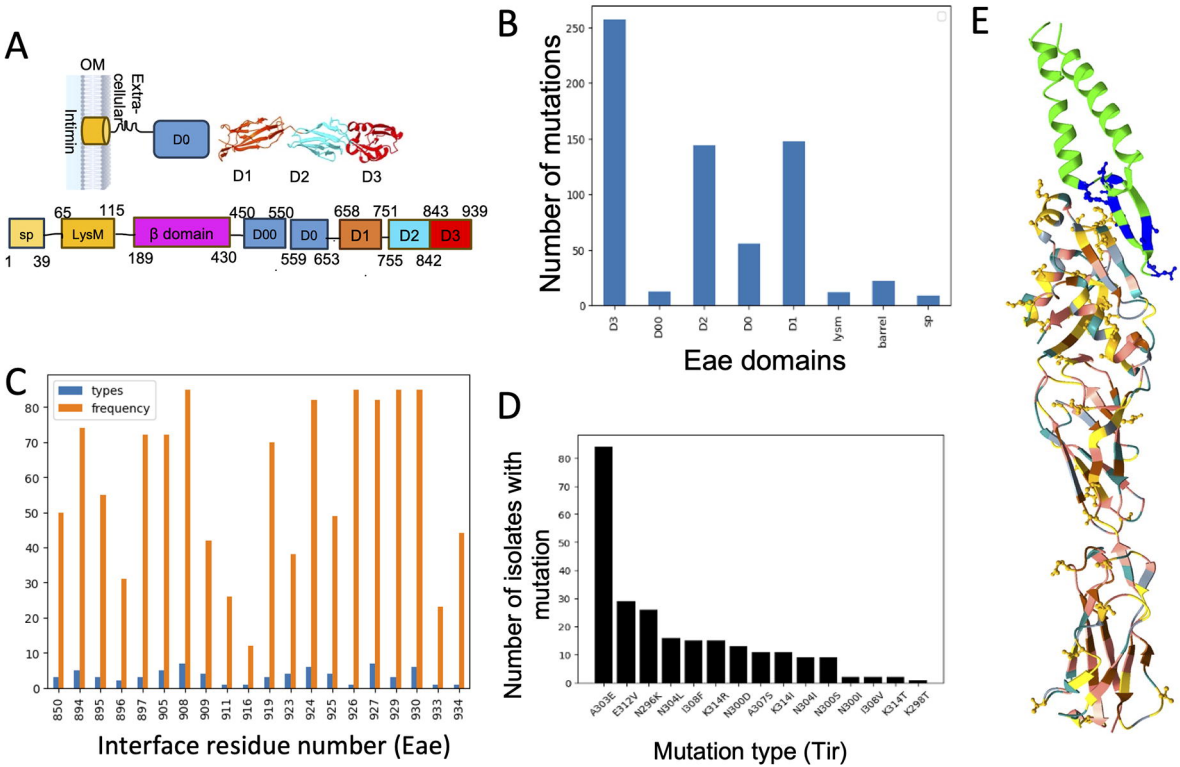
Table S2: Accession IDs (Enterobase) of the non-O157 STEC isolates used in the study.

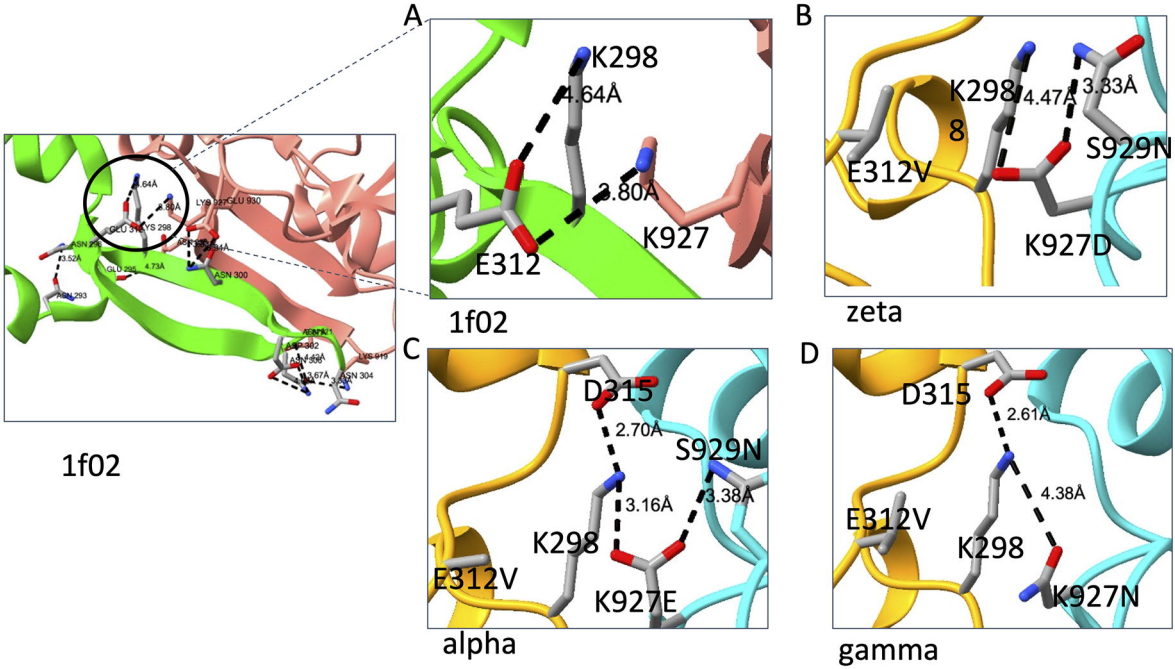
Table S3: Virulence factor genes (using 3 letter nomenclature) in non-O157 STEC isolate genomes (n=286). Their presence in intimin positive isolates, enrichment factor and Jaccard similarity is also listed.

Table S4: Virulence factor genes (using 4 letter nomenclature) in non-O157 STEC isolate genomes (n=286). Their presence in intimin positive isolates, enrichment factor and Jaccard similarity is also listed.


A
Number of isolates
1200
1000
800
600
400
200
0
O26
O145
O146
O91
O128Ab
O103
O6
O76
O117
O92
O25
O55
O111
O8
O78
O175
O121
O113
O123
serotypes
B
O103
O113
O128Ab
O166
O24
O5
O55
O76
O78
O79
O8
O80
O91
stx1
stx2
1.0
0.8
0.6
0.4
0.2
0.0
C
Number of genomes
80
70
60
50
40
30
20
10
0
10
15
20
25
30
35
40
45
50
Number of virulence genes
D
Number of genomes
80
60
40
20
0
stx1a
stx1c
stx2a
stx2b
stx2c
stx2d
stx2f
stx2g
stx2k

Q1
Q2
Q3
Q4
Q5
Q6
Q7
Q8
Q9
Q10
Q11
Q12
Q13
Q14
Q15
Q16
Q17
Q18
Q19
Q20
Q21
Q22
Q23
Q24
Q25
Q26
Q27
Q28
Q29
Q30
Q31
Q32
Q33
Q34
Q35
Q36
Q37
Q38
Q39
Q40
Q41
Q42
Q43
Q44
Q45
Q46
Q47
Q48
Q49
Q50
Q51
Q52
Q53
Q54
Q55
Q56
Q57
Q58
Q59
Q60
Q61
Q62
Q63
Q64
Q65
Q66
Q67
Q68
Q69
Q70
Q71
Q72
Q73
Q74
Q75
Q76
Q77
Q78
Q79
Q80
Q81
Q82
Q83
Q84
Q85
Q86
Q87
Q88
Q89
Q90
Q91
Q92
Q93
Q94
Q95
Q96
Q97
Q98
Q99
Q100
Q101
Q102
Q103
Q104
Q105
Q106
Q107
Q108
Q109
Q110
Q111
Q112
Q113
Q114
Q115
Q116
Q117
Q118
Q119
Q120
Q121
Q122
Q123
Q124
Q125
Q126
Q127
Q128
Q129
Q130
Q131
Q132
Q133
Q134
Q135
Q136
Q137
Q138
Q139
Q140
Q141
Q142
Q143
Q144
Q145
Q146
Q147
Q148
Q149
Q150
Q151
Q152
Q153
Q154
Q155
Q156
Q157
Q158
Q159
Q160
Q161
Q162
Q163
Q164
Q165
Q166
Q167
Q168
Q169
Q170
Q171
Q172
Q173
Q174
Q175
Q176
Q177
Q178
Q179
Q180
Q181
Q182
Q183
Q184
Q185
Q186
Q187
Q188
Q189
Q190
Q191
Q192
Q193
Q194
Q195
Q196
Q197
Q198
Q199
Q200
Q201
Q202
Q203
Q204
Q205
Q206
Q207
Q208
Q209
Q210
Q211
Q212
Q213
Q214
Q215
Q216
Q217
Q218
Q219
Q220
Q221
Q222
Q223
Q224
Q225
Q226
Q227
Q228
Q229
Q230
Q231
Q232
Q233
Q234
Q235
Q236
Q237
Q238
Q239
Q240
Q241
Q242
Q243
Q244
Q245
Q246
Q247
Q248
Q249
Q250
Q251
Q252
Q253
Q254
Q255
Q256
Q257
Q258
Q259
Q260
Q261
Q262
Q263
Q264
Q265
Q266
Q267
Q268
Q269
Q270
Q271
Q272
Q273
Q274
Q275
Q276
Q277
Q278
Q279
Q280
Q281
Q282
Q283
Q284
Q285
Q286
Q287
Q288
Q289
Q290
Q291
Q292
Q293
Q294
Q295
Q296
Q297
Q298
Q299
Q300
Q301
Q302
Q303
Q304
Q305
Q306
Q307
Q308
Q309
Q310
Q311
Q312
Q313
Q314
Q315
Q316
Q317
Q318
Q319
Q320
Q321
Q322
Q323
Q324
Q325
Q326
Q327
Q328
Q329
Q330
Q331
Q332
Q333
Q334
Q335
Q336
Q337
Q338
Q339
Q340
Q341
Q342
Q343
Q344
Q345
Q346
Q347
Q348
Q349
Q350
Q351
Q352
Q353
Q354
Q355
Q356
Q357
Q358
Q359
Q360
Q361
Q362
Q363
Q364
Q365
Q366
Q367
Q368
Q369
Q370
Q371
Q372
Q373
Q374
Q375
Q376
Q377
Q378
Q379
Q380
Q381
Q382
Q383
Q384
Q385
Q386
Q387
Q388
Q389
Q390
Q391
Q392
Q393
Q394
Q395
Q396
Q397
Q398
Q399
Q400
Q401
Q402
Q403
Q404
Q405
Q406
Q407
Q408
Q409
Q410
Q411
Q412
Q413
Q414
Q415
Q416
Q417
Q418
Q419
Q420
Q421
Q422
Q423
Q424
Q425
Q426
Q427
Q428
Q429
Q430
Q431
Q432
Q433
Q434
Q435
Q436
Q437
Q438
Q439
Q440
Q441
Q442
Q443
Q444
Q445
Q446
Q447
Q448
Q449
Q450
Q451
Q452
Q453
Q454
Q455
Q456
Q457
Q458
Q459
Q460
Q461
Q462
Q463
Q464
Q465
Q466
Q467
Q468
Q469
Q470
Q471
Q472
Q473
Q474
Q475
Q476
Q477
Q478
Q479
Q480
Q481
Q482
Q483
Q484
Q485
Q486
Q487
Q488
Q489
Q490
Q491
Q492
Q493
Q494
Q495
Q496
Q497
Q498
Q499
Q500
Q501
Q502
Q503
Q504
Q505
Q506
Q507
Q508
Q509
Q510
Q511
Q512
Q513
Q514
Q515
Q516
Q517
Q518
Q519
Q520
Q521
Q522
Q523
Q524
Q525
Q526
Q527
Q528
Q529
Q530
Q531
Q532
Q533
Q534
Q535
Q536
Q537
Q538
Q539
Q540
Q541
Q542
Q543
Q544
Q545
Q546
Q547
Q548
Q549
Q550
Q551
Q552
Q553
Q554
Q555
Q556
Q557
Q558
Q559
Q560
Q561
Q562
Q563
Q564
Q565
Q566
Q567
Q568
Q569
Q570
Q571
Q572
Q573
Q574
Q575
Q576
Q577
Q578
Q579
Q580
Q581
Q582
Q583
Q584
Q585
Q586
Q587
Q588
Q589
Q590
Q591
Q592
Q593
Q594
Q595
Q596
Q597
Q598
Q599
Q600
Q601
Q602
Q603
Q604
Q605
Q606
Q607
Q608
Q609
Q610
Q611
Q612
Q613
Q614
Q615
Q616
Q617
Q618
Q619
Q620
Q621
Q622
Q623
Q624
Q625
Q626
Q627
Q628
Q629
Q630
Q631
Q632
Q633
Q634
Q635
Q636
Q637
Q638
Q639
Q640
Q641
Q642
Q643
Q644
Q645
Q646
Q647
Q648
Q649
Q650
Q651
Q652
Q653
Q654
Q655
Q656
Q657
Q658
Q659
Q660
Q661
Q662
Q663
Q664
Q665
Q666
Q667
Q668
Q669
Q670
Q671
Q672
Q673
Q674
Q675
Q676
Q677
Q678
Q679
Q680
Q681
Q682
Q683
Q684
Q685
Q686
Q687
Q688
Q689
Q690
Q691
Q692
Q693
Q694
Q695
Q696
Q697
Q698
Q699
Q700
Q701
Q702
Q703
Q704
Q705
Q706
Q707
Q708
Q709
Q710
Q711
Q712
Q713
Q714
Q715
Q716
Q717
Q718
Q719
Q720
Q721
Q722
Q723
Q724
Q725
Q726
Q727
Q728
Q729
Q730
Q731
Q732
Q733
Q734
Q735
Q736
Q737
Q738
Q739
Q740
Q741
Q742
Q743
Q744
Q745
Q746
Q747
Q748
Q749
Q750
Q751
Q752
Q753
Q754
Q755
Q756
Q757
Q758
Q759
Q760
Q761
Q762
Q763
Q764
Q765
Q766
Q767
Q768
Q769
Q770
Q771
Q772
Q773
Q774
Q775
Q776
Q777
Q778
Q779
Q780
Q781
Q782
Q783
Q784
Q785
Q786
Q787
Q788
Q789
Q790
Q791
Q792
Q793
Q794
Q795
Q796
Q797
Q798
Q799
Q800
Q801
Q802
Q803
Q804
Q805
Q806
Q807
Q808
Q809
Q810
Q811
Q812
Q813
Q814
Q815
Q816
Q817
Q818
Q819
Q820
Q821
Q822
Q823
Q824
Q825
Q826
Q827
Q828
Q829
Q830
Q831
Q832
Q833
Q834
Q835
Q836
Q837
Q838
Q839
Q840
Q841
Q842
Q843
Q844
Q845
Q846
Q847
Q848
Q849
Q850
Q851
Q852
Q853
Q854
Q855
Q856
Q857
Q858
Q859
Q860
Q861
Q862
Q863
Q864
Q865
Q866
Q867
Q868
Q869
Q870
Q871
Q872
Q873
Q874
Q875
Q876
Q877
Q878
Q879
Q880
Q881
Q882
Q883
Q884
Q885
Q886
Q887
Q888
Q889
Q890
Q891
Q892
Q893
Q894
Q895
Q896
Q897
Q898
Q899
Q900
Q901
Q902
Q903
Q904
Q905
Q906
Q907
Q908
Q909
Q910
Q911
Q912
Q913
Q914
Q915
Q916
Q917
Q918
Q919
Q920
Q921
Q922
Q923
Q924
Q925
Q926
Q927
Q928
Q929
Q930
Q931
Q932
Q933
Q934
Q935
Q936
Q937
Q938
Q939
Q940
Q941
Q942
Q943
Q944
Q945
Q946
Q947
Q948
Q949
Q950
Q951
Q952
Q953
Q954
Q955
Q956
Q957
Q958
Q959
Q960
Q961
Q962
Q963
Q964
Q965
Q966
Q967
Q968
Q969
Q970
Q971
Q972
Q973
Q974
Q975
Q976
Q977
Q978
Q979
Q980
Q981
Q982
Q983
Q984
Q985
Q986
Q987
Q988
Q989
Q990
Q991
Q992
Q993
Q994
Q995
Q996
Q997
Q998
Q999
Q1000


A
OM
Extra-cellular
Intimin
D0
D1
D2
D3
sp
LysM
β domain
D00
D0
D1
D2
D3
1
39
65
115
189
430
450
550
559
653
658
751
755
842
843
939
B
Number of mutations
Eae domains
D3
D00
D2
D0
D1
lysm
barrel
sp
C
types
frequency
Interface residue number (Eae)
D
Number of isolates with mutation
Mutation type (Tir)
A303E
E312V
N296K
N304L
I308F
K314R
N300D
A307S
K314I
N304I
N300S
N300I
I308V
K314T
K298T
E
A
B
C
D


A
K298
4.64Å
3.80Å
K927
E312
1f02
B
K298
4.47Å
3.83Å
8
S929N
E312V
K927D
zeta
C
D315
2.70Å
S929N
3.16Å
3.38Å
K298
E312V
K927E
alpha
D
D315
2.61Å
4.38Å
E312V
K298
K927N
gamma
1f02