View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Apollo

OPEN BIOLOGY

rsob.royalsocietypublishing.org

Research

Cite this article: Stubenrauch CJ, Dougan G, Lithgow T, Heinz E. 2017 Constraints on lateral gene transfer in promoting fimbrial usher protein diversity and function. *Open Biol.* **7**: 170144.
http://dx.doi.org/10.1098/rsob.170144

Received: 14 June 2017
Accepted: 5 October 2017

Subject Area:
genomics/biochemistry/bioinformatics/microbiology

Keywords:
outer membrane, translocation and assembly module, fimbriae

Authors for correspondence:
Trevor Lithgow
e-mail: trevor.lithgow@monash.edu
Eva Heinz
e-mail: eva.heinz@sanger.ac.uk

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3917935.

THE ROYAL SOCIETY PUBLISHING

Constraints on lateral gene transfer in promoting fimbrial usher protein diversity and function

Christopher J. Stubenrauch[1], Gordon Dougan[2], Trevor Lithgow[1] and Eva Heinz[2]

[1]Infection and Immunity Program, Department of Microbiology, Monash University, Clayton 3800, Australia
[2]Infection Genomics Program, Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK

EH, 0000-0003-4413-3756

Fimbriae are long, adhesive structures widespread throughout members of the family Enterobacteriaceae. They are multimeric extrusions, which are moved out of the bacterial cell through an integral outer membrane protein called usher. The complex folding mechanics of the usher protein were recently revealed to be catalysed by the membrane-embedded translocation and assembly module (TAM). Here, we examine the diversity of usher proteins across a wide range of extraintestinal (ExPEC) and enteropathogenic (EPEC) *Escherichia coli*, and further focus on a so far undescribed chaperone–usher system, with this usher referred to as UshC. The fimbrial system containing UshC is distributed across a discrete set of EPEC types, including model strains like E2348/67, as well as ExPEC ST131, currently the most prominent multi-drug-resistant uropathogenic *E. coli* strain worldwide. Deletion of the TAM from a naive strain of *E. coli* results in a drastic time delay in folding of UshC, which can be observed for a protein from EPEC as well as for two introduced proteins from related organisms, *Yersinia* and *Enterobacter*. We suggest that this models why the TAM machinery is essential for efficient folding of proteins acquired via lateral gene transfer.

1. Introduction

Bacteria can acquire new phenotypes to adapt to changing environments through mutations of their genome and through the acquisition of new genes. Genes acquired through lateral gene transfer (LGT) are particularly important for the adaptation of bacterial pathogens, providing them with means to invade and conquer new niches and often to promote their virulence [1–7]. In many cases, the selectable phenotypes arising from the LGT are due to a monomeric enzyme or pump that promotes resistance to a heavy metal or antimicrobial compound. However, some phenotypes require multimeric structures, encoded on multiple genes and ultimately assembled by the host cell's assembly machinery. The success or failure to assemble complicated cellular machinery acquired through LGT would be a key hurdle in the evolutionary success of bacterial lineages adapting to changing environments. Stated simply, acquiring genes that encode a virulence factor ultimately needs to be followed by the assembly of a functional form of the virulence trait in order to effect a phenotypic outcome.

Efficient attachment to host cells is one of the key virulence factors essential for many bacterial pathogens. The 2011 outbreak of *Escherichia coli* STEAEC O104:H4, which was hallmarked by a high morbidity and mortality rate, is understood as a new configuration of known virulence factors: a combined effect on better adherence through an acquired adhesion system (Iha), which in turn provided for better delivery of the Stx toxin to host cells with devastating effect [8–10]. This is an example of how crucially important pathogen–host cell adhesion is to successfully establish infection in the human host for specific *E. coli* pathotypes.

© 2017 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Among the arsenal of adhesive structures in Gram-negative bacteria, the most important are fimbriae or pili, which are multimeric, extracellular fibres. In addition to the multiple subunits that form each fimbrial fibre, a set of membrane-embedded and periplasmic proteins form the molecular machinery to extrude the fimbriae across the bacterial outer membrane. Key examples of these molecular machines have been the subject of an impressive array of structural and functional studies [11,12], and genome sequencing studies are identifying a growing number of further, uncharacterized systems. The chaperone–usher systems are usually encoded in operons of genes, and comprise at least four subunits: a chaperone to aid assembly and transport of the fimbrial subunits within the periplasm; in most cases at least two types of fimbrial subunits, including a tip adhesin that confers binding specificity and the major fimbrial subunit that comprises the bulk of the structure; and an usher protein, which serves as the membrane conduit through which the fimbriae are translocated [13]. Their classification system is based on Greek letters (alpha-, beta-, gamma-, etc.) with the usher protein used as the basis of the classification [13]. While the fimbrial subunits require their cognate chaperone and usher for assembly [12], recent work suggests that, in turn, the usher proteins—which are beta-barrel outer membrane proteins—require the beta-barrel assembly machinery (BAM) complex and the translocation and assembly module (TAM) in order to be effectively assembled into the bacterial outer membrane [14].

The BAM complex is essential for the assembly of outer membrane proteins, and the core gene *bamA* is essential for bacterial cell viability [15–17]. The TAM is widely distributed across Gammaproteobacteria [18,19], and is involved in the biogenesis of outer membrane proteins such as autotransporter adhesins [20], inverse autotransporter adhesins [21] and fimbrial ushers [14]. However, the TAM, consisting of the protein subunits TamA and TamB [20], is not essential for cell viability. It has therefore been unclear what selective pressure is in place to have *tamA* and *tamB* maintained across the Gammaproteobacteria. The current hypothesis is that the TAM assists in the folding and assembly of proteins that have complex structures [22]. In principle, this may include alien proteins acquired from other bacteria via LGT.

Here, we assess the diversity of fimbrial usher proteins across an extensive collection of enteropathogenic *E. coli* (EPEC) and extraintestinal pathogenic *E. coli* (ExPEC), especially uropathogenic *E. coli* (UPEC) and subsets of other species from the Enterobacteriaceae. Analysis of this large collection emphasized earlier observations that the usher proteins have a non-uniform presence in *E. coli* [23]. We find that some usher proteins are highly conserved, suggesting that, for a significant time period, they have served core functions in *E. coli*. Other ushers are much more distinctly distributed, which suggests a more recent acquisition and/or a more specialized function. The complex distribution is further reflected more broadly across the Enterobacteriaceae. LGT is the most likely method of dissemination, given the highly uneven distribution not only in *E. coli* but also when considering other genera such as *Salmonella*, *Enterobacter*, *Yersinia* and *Klebsiella*. We show that the TAM machinery is important in the folding of an alien sequence from *Enterobacter asburiae* and *Yersinia enterocolitica* into the outer membrane of *E. coli*, and suggest that selective pressures favouring exchange of large surface proteins through LGT contribute to the maintenance of cellular factors such as the TAM. Large adhesins and other unusual outer membrane-embedded cell surface proteins are frequently exchanged on mobile elements and thus acquired through LGT to new host cells. Having a folding machinery that significantly contributes to the speed with which these large structures can be used for phenotypic advantages is crucial for the invasion of some niches, such as the urogenital tract [14]. We followed on this observation, and investigated the evolutionary history of the usher family, which form the basis of fimbriae, one of the best-studied group of adhesive structures, and the folding kinetics of several usher proteins in a heterogeneous host.

2. Results

While *E. coli* can be considered a model organism and is a commensal of humans, it is also a significant human pathogen. Pathotypes of *E. coli* have been shown to employ a diverse range of fimbriae to ensure infection [13,23,24], which translates to the variability in adherence to different host cells among EPEC and ExPEC (especially UPEC) strains. Although genomic insights over the recent years have blurred the lines between the different pathotypes [25], there has been a recent expansion of knowledge about the diversity of *E. coli*, including our understanding of the paraphyletic origins of EPEC [26–28]. We therefore sought to assess whether (and how) the diversity of fimbriae–usher adhesive systems is reflected in *E. coli* pathotypes.

A broad database of a selection of *E. coli* genomes ([26–29]; electronic supplementary material, table S3) was built and analysed by hidden Markov Model search using HMMER to identify all encoded usher proteins in the dataset. Initially, to classify sequences without functional annotations, we spiked our dataset with annotated sequences [23], and after removing highly similar sequences, we defined usher groups based on manual assessment of monophyletic branching with reference sequences (figure 1). This revealed several branches in the tree that had no clear association with previously described families or represented divergent branches (e.g. 'Lpf-like 2' is most similar to proteins annotated Lpf-like in UniProt, but clearly distinct from the described Lpf-like), indicating that we are only beginning to appreciate the diversity to be found in *E. coli* usher sequences. Based on this classification, we incorporated the monophyletic groups into our full dataset, to investigate the distribution of usher proteins across this large dataset (figure 2).

This analysis revealed a peculiar distribution of members of a Gamma subfamily, designated Gamma* ('Gamma star', in accordance with [23]). Within this subfamily, we identified an *E. coli* protein which we refer to as UshC (locus tag identifer ECSF_0165), which is located in an operon consisting of the fimbrial subunit *ushA*, chaperone *ushB*, usher *ushC* and fimbrial subunit *ushD* (electronic supplementary material, figure S4). UshC showed a distinct distribution in ExPEC and EPEC lineages and is found in the globally distributed and most abundant circulating multi-drug-resistant ExPEC sequence type ST131 [31,32]. Especially within the EPEC lineages [26–28], we observe a distribution and maintenance in a large group of strains containing the model strain O127:H6 E2348/69, as well as in the interspersed isolates of other pathotypes. The distribution of UshC across the tree also strongly suggests distribution via lateral transfer, given the distant relation between the EPEC groups and ST131 (figure 2).

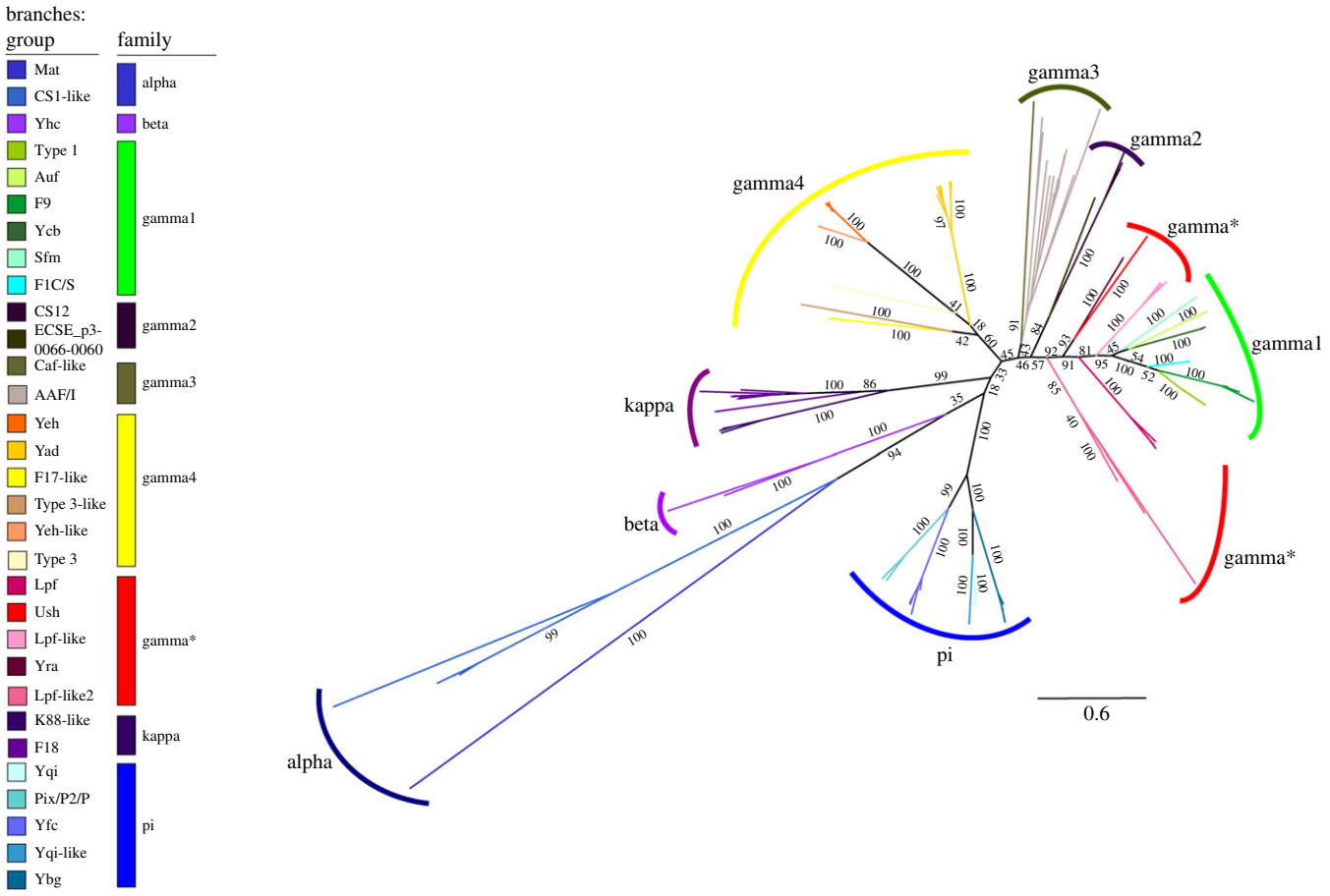

Figure 1. Phylogeny of usher proteins in a large *E. coli* collection. The *E. coli* genomes (electronic supplementary material, table S3) were searched using HMMER and the Pfam profile for usher proteins and subjected to tree calculation using RAxML. Following manual assessment visually, monophyletic groups are coloured according to their described members (electronic supplementary material, table S4); four groups without described members as in Wurpel *et al.* [23] are based on the annotation of similar sequences in UniProt (LPF-like 2, AggC (AAF/I), FedC (F18), MrkH (Type 3)).

UshC, furthermore, shows an inverse correlation with another member of this family, YraJ (figures 1 and 2), with which it is closely related and which was found in almost all other EPEC strains (figure 2).

To gain further insights into how representative the distribution is for other bacterial species, we investigated the diversity of ushers across model organisms within the Enterobacteriaceae (figure 3). This analysis emphasized that usher proteins are widely distributed, and that the different families previously based on *E. coli* representatives [13,23] can be found in a variety of organisms. However, usher proteins are not evenly distributed within the different genera. When considering selected representative species, it is clear that often unrelated genera are more similar with respect to usher families than species within the same genus (figure 3). This is in accordance with the typical distribution of adhesins [21] and other virulence factors, highlighting the importance of LGT for the dissemination of fimbriae [33].

Further analysis of the Gamma* subfamily of chaperone–usher systems closely related to UshC and YraJ revealed several monophyletic lineages for the usher proteins (figure 4*a*), emphasizing their independent acquisition in the different *E. coli* lineages (figure 1). The current model for creating diversity in the chaperone–usher systems available to a species posits that the entire chaperone–usher operon is mobile during LGT. Phyre2, a sequence/structure comparison tool [34], revealed that there are distinct adhesins present in the UshC and YraJ subgroups. While some are more similar to the LpfD adhesin [35], others are more similar to the *Pseudomonas* adhesin CupB6 ([36]; figure 4*b*; electronic supplementary material, table S5). This dichotomy is further reflected in the chaperones that assemble the adhesin subunits (electronic supplementary material, figure S2), which cluster similarly to the phylogenetic history of the usher proteins, largely consistent with an LGT event acting upon the entire chaperone–usher operon, although the ordering of the inner branches is not stable with low support values. In addition, we see several cases of pseudogenization or loss of the chaperones as well as fimbrial adhesins (electronic supplementary material, table S2), which further highlights the dynamic nature of these operons also once they have been incorporated into the chromosome. The capacity for fimbriae to carry different tip structures is well studied and often used as a surface display system [37]. However, these experiments also only report on steady-state expression levels, and not how rapidly or efficiently expression is enacted.

While the fimbrial subunits are folded with the chaperone encoded within the operon, the usher itself needs the cellular beta-barrel assembly machinery. To address the assembly mechanism for $UshC_{EPEC}$ (the UshC from EPEC O127:H6 E2348/69; UniProt: B7UIJ5), a biochemical assay was established wherein $UshC_{EPEC}$ was expressed under the control of a T7 RNA polymerase-driven promoter in *E. coli* BL21 Star^{TM} (DE3). In this system, transcription by *E. coli* RNA polymerase is repressed, and ^{35}S-labelled amino acids are incorporated into the protein of interest [14], allowing for its detection by radiography. Analysis of $UshC_{EPEC}$ assembly revealed that, relative to the levels of usher assembly seen in wild-type *E. coli*, in the absence of either *tamA* or *tamB*, there was a

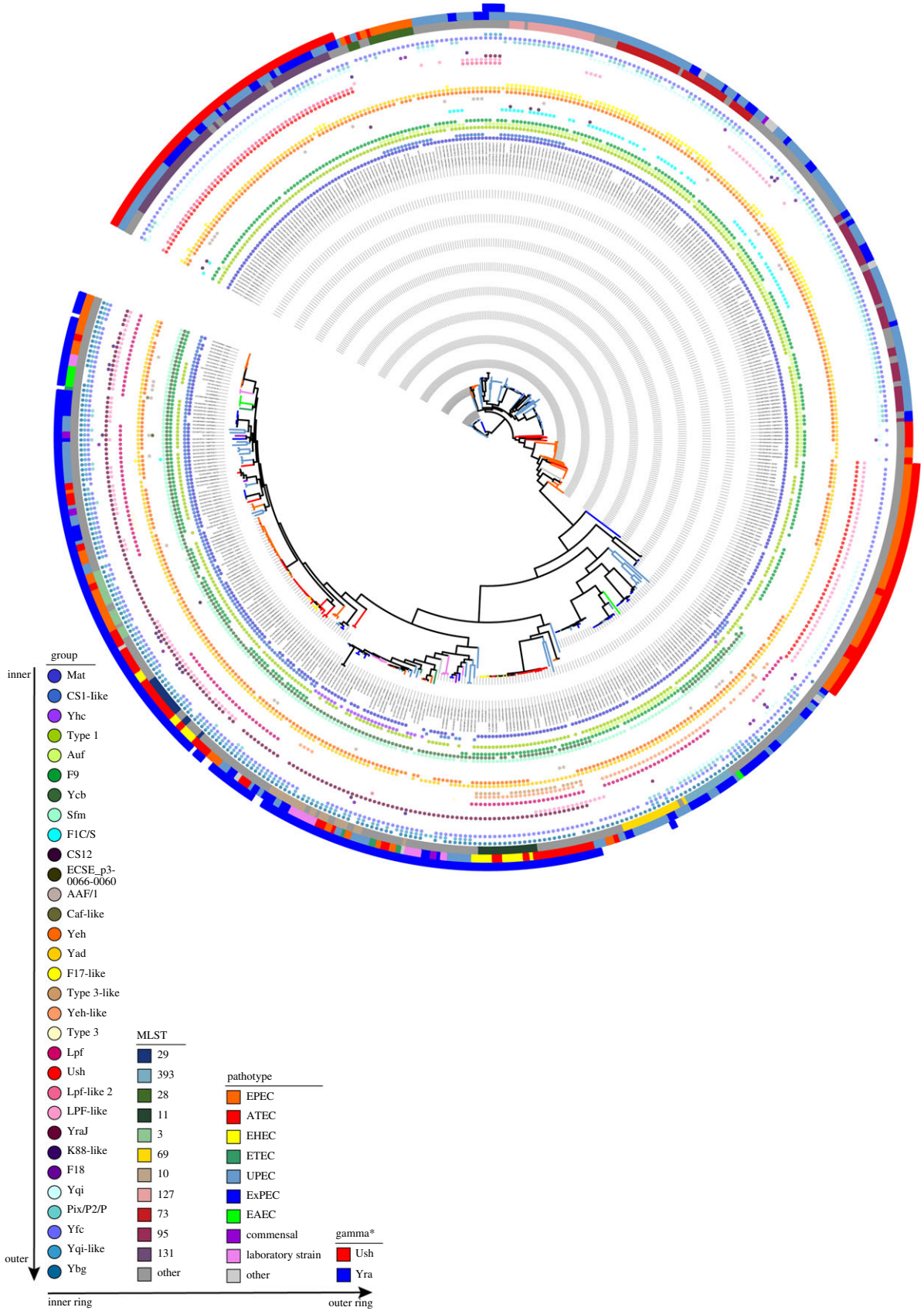

Figure 2. The distribution of ushers across the *E. coli* pangenome. Given the recent increase in publicly available EPEC/UPEC genomes [26–29], we investigated the distribution of ushers across *E. coli*. The tree is based on a core gene alignment of *E. coli* genomes with a focus on EPEC and ExPEC strains, but also including a variety of reference strains for other pathovars. The inner rings show the respective usher families, the other rings show, from inside to outside, the main sequence types according to multi-locus sequence typing (MLST), and the pathotypes. The presence of UshC and YraJ are again highlighted in the outermost ring. This highlights the uneven distribution of the two closely related usher proteins UshC and YraJ, both across the *E. coli* diversity but also within the respective pathovars; branches are coloured according to the pathovars scheme as indicated in the legend. The tree representation was performed using iTOL [30]. Pathotypes: EPEC, enteropathogenic *E. coli*; ATEC, atypical EPEC; EHEC, enterohaemorrhagic *E. coli*; ETEC, enterotoxigenic *E. coli*; UPEC, uropathogenic *E. coli*; ExPEC, extraintestinal pathogenic *E. coli*; EAEC, enteroaggregative *E. coli*; other, see details in electronic supplementary material, table S3.

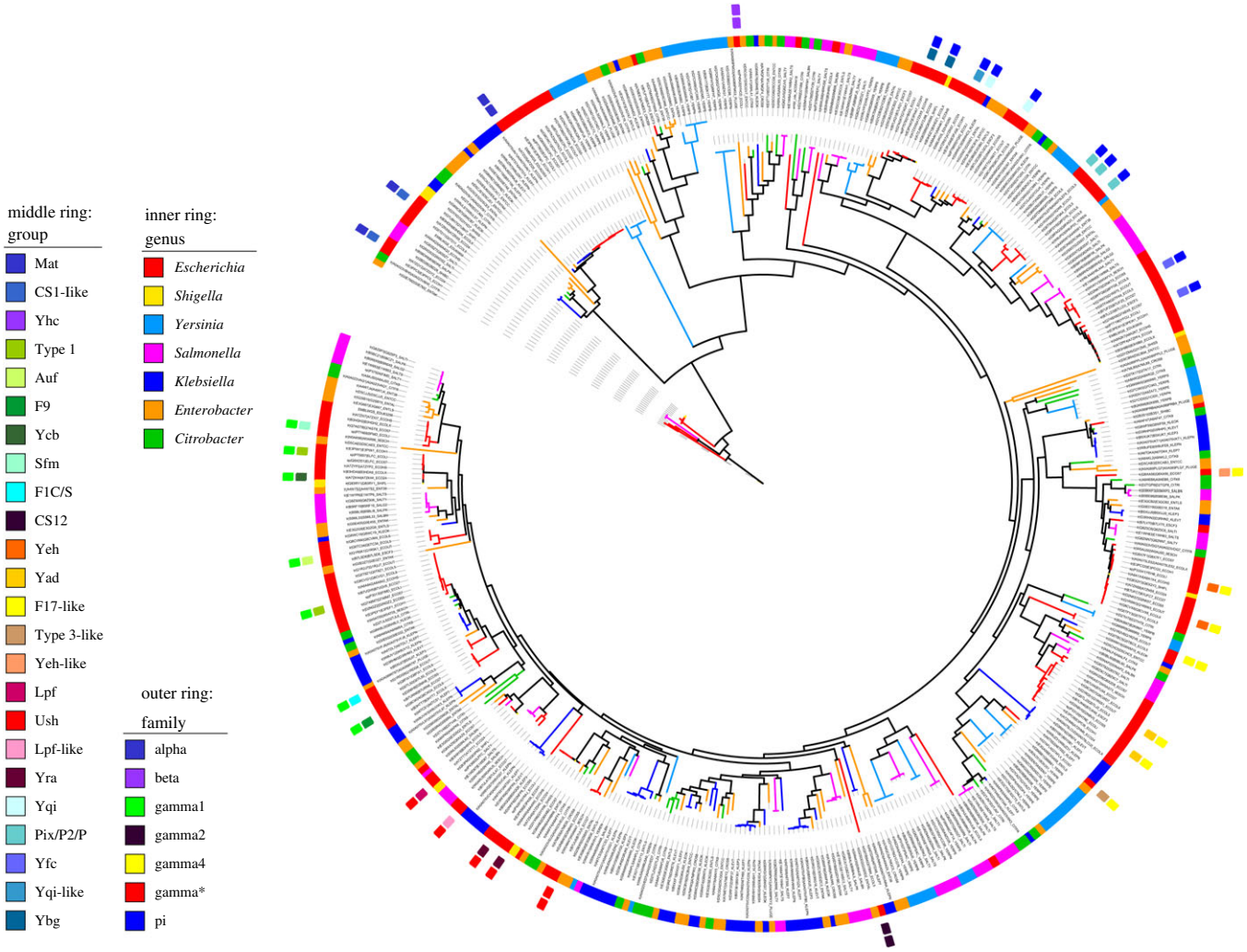

Figure 3. The phylogenetic relationships of usher proteins in Enterobacteriaceae reference strains. The tree was generated using RAxML and shows the diversity of usher proteins across several genera as given in the electronic supplementary material, table S1 and highlights that these proteins are very widely distributed across various species within the Enterobacteriaceae. The colours indicate the different genera (inset). Chaperone–usher systems are assigned as in figure 1, but only for the representative *E. coli* proteins as shown in the middle and outer ring fragments.

decrease in the amount of functionally assembled usher (figure 5*a*; electronic supplementary material, figure S5).

The irregular distribution of fimbrial operons across the taxonomic range (figures 2 and 3) and the importance and apparent frequency of LGT leading to their distribution led us to test whether the *E. coli* host machinery would be able to assemble the products of newly acquired fimbrial ushers. Homologues $UshC_{Ye}$ (from *Y. enterocolitica* LC20; UniProt W8V9V3) and $UshC_{Ea}$ (from *En. asburiae* LF7a; UniProt G2S6X5) showed a significant decrease in the amount of functionally assembled usher (figure 5*b*,*c*; electronic supplementary material, figure S5). Densitometric analyses revealed the observed rate constant for the assembly of the protease-resistant (assembled) UshC was significantly greater when catalysed by the TAM (figure 5*d*), for the recently acquired $UshC_{EPEC}$ and even more so for the alien sequences from *Y. enterocolitica* and *En. asburiae*.

3. Discussion

Genome plasticity particularly through LGT has been suggested as the key to the great variability seen in the various pathotypes of *E. coli*, by enabling constant alterations to the fitness and resultant competitiveness of individuals in specific niches [4]. Many studies in comparative genomics support this concept of *E. coli* genome plasticity; the core genome of *E. coli* K-12 substr. MG1655, EHEC O157:H7 and UPEC CFT073 encodes approximately 40% of their proteome [38]. The distribution of chaperone–usher systems contributes to this diversity in proteome and adaptive fitness in *E. coli* lineages.

Adhesion is an essential step for many human pathogens to anchor themselves in their respective niche. In the case of uropathogenic or enteric pathogens, failure to rapidly and effectively adhere impacts on colonization, given that flushing action of constant fluid movement is one of the main challenges facing bacteria in these environmental niches. Fimbriae play an essential role in the adhesion process, and a delay in their expression equates to a failure to adhere in host niches [14,39]. It is perhaps because of their importance in pathogenic lifestyles and host interactions that adhesins are (i) often shared via LGT [40] and (ii) undergo rounds of adaptation to enhance host interaction or evasion of the immune system through positive selection and/or recombination [41–43]. A high number of transposable elements associated with the usher operons were detected (electronic supplementary material, table S2), and several cases of potential pseudogenization were observed, mainly of the chaperone, through frameshifts. Fimbrial operons are a highly dynamic locus in most genomes, regarding both their occurrence/absence and precise sequence [41,42].

Figure 4. The Ush/Yra clade ushers. (*a*) Phylogenetic tree of the respective usher sequences calculated with MrBayes shows the various adhesins associated with the respective usher sequence. The nonlinear evolution of the chaperone–usher systems is apparent from the different monophyletic groups displaying a mixed distribution of associated adhesin sequences in the operons. (*b*) Similarity network of the sequences as in the electronic supplementary material, table S5, highlights two different types of adhesins associated with the different operons; one group comprises stalk-like adhesins, which can also form the tip, whereas the other group includes a second adhesion protein different to the stalk-like sequences.

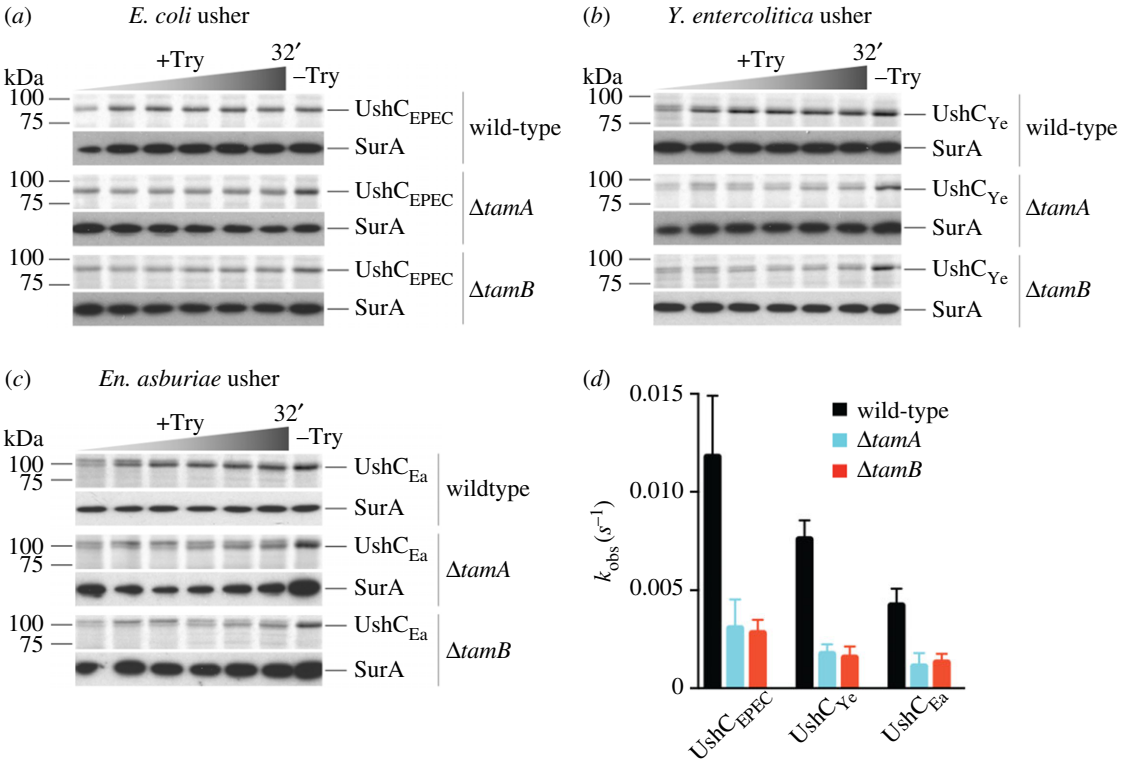

Figure 5. Usher biogenesis in *E. coli*. *Escherichia coli* cells harbouring (*a*) pCJS39, (*b*) pCJS75 or (*c*) pCJS77 were assessed by pulse chase analysis. Aliquots were taken at 10 s, 2, 4, 8, 16 and 32 min, treated with (+Try) or without (−Try, last timepoint only) 20 µg ml^{-1} trypsin. Analysis was by SDS–PAGE, storage phosphor-imaging and immunoblotting. Representative autoradiograms and immunoblots are shown, from three independent experiments ($n = 3$). The time increment is indicated as a graded triangle above the autoradiogram. SurA is a periplasmic protein used to assess the integrity of the outer membrane. (*d*) The usher densities at each timepoint (*a*–*c*) were used to calculate the observed rate constants (k_{obs}). Calculations were as per Stubenrauch *et al.* [14]. Error bars represent s.e.m. ($n = 3$), and all folding rates of mutants were significantly slower than the respective wild-type folding rate, as assessed by one-way ANOVA ($p < 0.05$).

We analysed the diversity of fimbrial usher distribution in a large collection of *E. coli* whole-genome data with a focus on EPEC and ExPEC. In many cases, a large number of fimbrial loci could be encoded within a single strain, in some cases up to 16 (figure 2). Despite this, YraJ and UshC were never simultaneously encoded by any *E. coli* lineage (figure 2). Mutual exclusion has been observed among other classes of outer membrane proteins. It has been hypothesized that this occurs as result of environmental specificity, incompatibility or functional redundancy, or to avoid interference in similar target sites [44]. It is not clear, however, how any of these factors would impact to keep a mutual exclusivity between *yraJ* and *ushC*, especially if their target adhesins have the potential to have distinct specificities [41], and if this observation remains supported when further *E. coli* sequences keep being analysed.

Genes transferred by LGT pose a potential risk to the cell, and are often initially silenced by systems such as the histone-like nucleoid structuring protein (H-NS) [45]. One such risk is that differences in codon usage between species will impact on translation rates in cells that attempt to express alien proteins acquired through LGT [46,47]. An additional risk would be the inhibition of protein assembly rates in cells that express alien proteins acquired through LGT. Studies in *E. coli* show that in the absence of the TAM, reduced assembly rates occur for model proteins like FimD [14]. The expression of adhesins is tightly regulated including complex counteracting factors [48], and several *E. coli* fimbriae–usher operons are under the control of H-NS, further highlighting the need for tight regulation and their likely role in virulence [49]. It is now also clear that for UshC, whose evolutionary history clearly indicates LGT between various lineages, the folding efficiency expressed is rate-limited by the TAM. This indicates that adhesins acquired by LGT (e.g. from *Y. enterobacter* and *En. asburiae*) would be more rapidly deployed and expressed to promote new phenotypes, provided that the TAM is present (figure 6). We suggest that this is an important function of *tamA* and *tamB* in the Proteobacteria, where they are almost universally conserved, despite their non-essential nature [18,19]. This underscores that there is no apparent limitation to the sharing of fimbrial clusters. Given the wide distribution across Enterobacteriaceae shown here, there seems to be no strictly imposed restriction to the protein sequence of the usher. This parallels the diversity of other adhesins assembled by the TAM machinery such as autotransporters and inverse autotransporters [20,21,52]. Despite its molecular complexity, the chaperone–usher system is highly adaptable for mediating bacterial adhesion and readily shared across bacterial species. A better understanding of the binding properties of the different fimbrial adhesins, combined with high-resolution sequence analysis such as shown here, provides insight into host range and tissue tropisms species of Enterobacteriaceae and will shed further light on the highly complex evolution of uro- and enteric pathogens.

4. Material and methods

4.1. Sequence analyses of reference strains

The full proteomes for the respective reference strains (electronic supplementary material, table S1) were retrieved from the UniProt database ([53]; last accessed 7 May 2015). The HMMER profile for ushers (PF00577.15) was retrieved from the Pfam

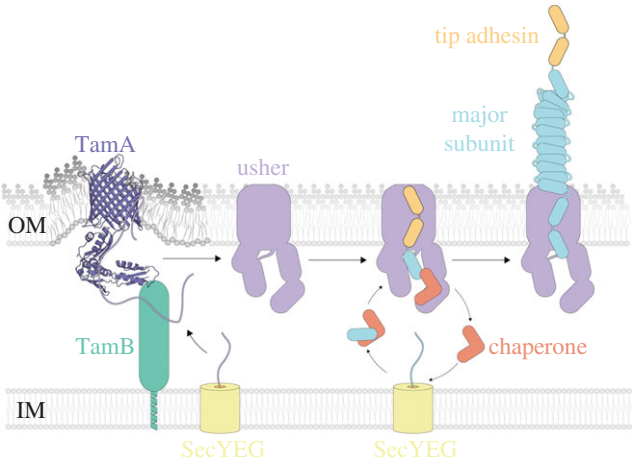

Figure 6. Schematic of fimbriae biogenesis. Nascent protein is translocated across the inner membrane (IM) via the SecYEG apparatus. The TAM is thought to promote protein insertion through destabilization of the lipid bilayer [50,51]. TamA (pdb: 4C00) acts as a lever, pushing onto TamB, to distort the outer membrane (OM). Once assembled, the fimbrial usher acts as an anchor and pore for fimbrial subunits to thread through. Initially, the dedicated chaperone transfers the tip adhesion subunit to initiate fimbrial biogenesis. The chaperone subsequently transfers hundreds to thousands of the major fimbrial subunits, allowing the growing pilus to extend from the cell surface [12].

website [54], and HMMER [55] was used to run a search (HMMER v. 3.0; hmmsearch using the -max option with all else default) against the combined file of all reference strain protein sequences. To identify non-usher contaminants from divergent usher sequences, a protein–protein similarity network was used to extract the sequences of all usher proteins following manual inspection of the formed clusters (CLANS [56]; *p*-value cut-off 1×10^{-5}; electronic supplementary material, figure S1). Sequences with less than 600 amino acids were furthermore removed to remove contaminants. One divergent *E. coli* sequence was missing from the current dataset and added manually (E3PPC5 [23]). The operons of the reference strains as shown in figure 4 and the electronic supplementary material, table S2 and figures S2 and S3 were retrieved manually from Ensembl bacteria [57], and adhesins were clustered using CLANS (*p*-value cut-off 1×10^{-10}; figure 4) to identify different types of adhesins. Alignments were performed with mafft [58] using the -linsi option, and informative sites selected using trimal with the auto-1 setting [59]. Trees were calculated using RAxML [60], MrBayes [61] or PhyloBayes [62] as indicated in the respective figure legends. Calculations for RAxML were performed with the fast bootstrap setting, and the model was set to PROTGAMMALGF with 100 bootstrap replicates; MrBayes was run for 1 million generations under the mixed amino acids model, with a burnin of 25% for the consensus tree; and PhyloBayes was run using the C20 (electronic supplementary material, figure S2*c*) or C60 (electronic supplementary material, figure S3*b*) model, and convergence was assessed manually with the bpcomp and tracecomp commands as suggested by the authors, consensus trees were calculated with 25% burnin.

4.2. Sequence analyses of enteropathogenic *Escherichia coli* diversity

For the *E. coli* diversity investigation, nucleotide sequences were retrieved from GenBank (for accession numbers, see electronic supplementary material, table S2), and to limit differences in gene/start site calling due to the different publication times and annotation software used for the included genomes, the assemblies were all annotated using Prokka [63]. The core gene alignment was generated with Roary [64], informative sites were chosen using snp_sites with default settings [65] and the tree calculation was performed using RAxML with the reversible GTR model and 100 bootstrap replicates. To find and distinguish the different usher in the dataset, a HMMER (v. 3.1 [55]) search with the Pfam profile PF00577.15 as described above was performed. All resulting hits were combined, and sequences with less than 600 amino acids removed as fragments/incomplete sequences. The remaining sequences were clustered with Uclust [66] using the usearch–cluster_fast command at a cut-off of id 0.99, and the resulting centroids were used for a tree calculation. To facilitate distinguishing the different usher proteins, the sequence set was furthermore spiked with reference sequences for the main *E. coli* usher groups as indicated in the electronic supplementary material, table S3. The sequences were then aligned using Muscle [67], and informative sites were chosen with the tcs online server [68]. The resulting reduced alignment was used as input for a tree calculation using RAxML with the LG model and empirical frequencies and 100 bootstrap replicates. The resulting tree was used to identify the centroids branching monophyletic with the different spiked usher sequences, and the presence or absence of sequences in the respective clusters is indicated in figure 2.

4.3. Functional analyses

Pulse chase analyses were performed in triplicate as described previously [14] with several modifications. Briefly, *E. coli* BL21 Star™ (DE3) wild-type, Δ*tamA* or Δ*tamB* strains were incubated to mid-log phase in LB media (37°C, 200 r.p.m. (25 mm orbit)), then transferred to M9-S media [14]. Following a 30 min incubation (37°C, 200 r.p.m. (25 mm orbit)), cells were treated for 1 h with rifampicin (200 μg ml^{-1}, 37°C, 400 r.p.m. (3 mm orbit)) and induced for 5 min with IPTG (0.2 mM, 30°C, static). Cells were then 'pulse'-labelled for 45 s with EXPRE^{35}S^{35}S, [^{35}S]-Protein Labelling Mix (30 μCi ml^{-1}, 30°C, static), containing 73% [^{35}S]-methionine and 22% [^{35}S]-cysteine (NEG072, Perkin Elmer), and then immediately subjected to centrifugation (5 min, 3000*g*, 4°C) and resuspended in M9 + S media [14]. Cells were then 'chased' for up to 32 min (30°C, static) and aliquots were taken at appropriate timepoints. Aliquots were treated with the exogenous addition of trypsin (to 20 μg ml^{-1}) for 10 min on ice, before the total protein content of the samples was TCA-precipitated. The TCA-precipitated pellets were washed with acetone and resuspended in SDS loading dye. Samples were incubated for 10 min at 100°C and analysed by 12% SDS–PAGE. After electrophoresis, proteins were transferred onto 0.45 μm nitrocellulose membranes. Radiation was captured overnight using a storage phosphor screen (GE Health Sciences) and detected using a Typhoon Trio (320 nm). Immunoblotting for the presence of the control protein SurA was performed as per Leyton *et al.* [69].

Data accessibility. The alignment and tree files are available under https://figshare.com/s/20b09a45736b6188fc32. All sequences used were retrieved from public databases, accession numbers of all used sequences are available in the electronic supplementary material, tables.

Authors' contributions. E.H. and T.L. conceived the experiments. E.H. and C.J.S. performed the experiments and analysed the data. E.H., C.J.S., T.L. and G.D. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work was supported by the Wellcome Trust (206194) and the NHMRC Program grant (1092262 to G.D. and T.L.). T.L. is an ARC Australian Laureate Fellow.

Acknowledgements. The help of the pathogen informatics team at the Wellcome Trust Sanger Institute (WTSI) is gratefully acknowledged.

References

1. Ochman H, Lawrence JG, Groisman EA. 2000 Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304. (doi:10.1038/35012500)
2. Thomas CM, Nielsen KM. 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721. (doi:10.1038/nrmicro1234)
3. Boto L. 2010 Horizontal gene transfer in evolution: facts and challenges. *Proc. R. Soc. B* **277**, 819–827. (doi:10.1098/rspb.2009.1679)
4. Leimbach A, Hacker J, Dobrindt U. 2013 *E. coli* as an all-rounder: the thin line between commensalism and pathogenicity. *Curr. Top. Microbiol. Immunol.* **358**, 3–32. (doi:10.1007/82_2012_303)
5. Roberts AP, Kreth J. 2014 The impact of horizontal gene transfer on the adaptive ability of the human oral microbiome. *Front. Cell. Infect. Microbiol.* **4**, 124. (doi:10.3389/fcimb.2014.00124)
6. Ruzzini AC, Clardy J. 2016 Gene flow and molecular innovation in bacteria. *Curr. Biol.* **26**, R859–R864. (doi:10.1016/j.cub.2016.08.004)
7. Baharoglu Z, Garriss G, Mazel D. 2013 Multiple pathways of genome plasticity leading to development of antibiotic resistance. *Antibiotics* **2**, 288–315. (doi:10.3390/antibiotics2020288)
8. Bielaszewska M, Mellmann A, Zhang W, Kock R, Fruth A, Bauwens A, Peters G, Karch H. 2011 Characterisation of the *Escherichia coli* strain associated with an outbreak of haemolytic uraemic syndrome in Germany, 2011: a microbiological study. *Lancet Infect. Dis.* **11**, 671–676. (doi:10.1016/S1473-3099(11)70165-7)
9. Mellmann A *et al*. 2011 Prospective genomic characterization of the German enterohemorrhagic *Escherichia coli* O104:H4 outbreak by rapid next generation sequencing technology. *PLoS ONE* **6**, e22751. (doi:10.1371/journal.pone.0022751)
10. Tarr PI, Bilge SS, Vary Jr JC, Jelacic S, Habeeb RL, Ward TR, Baylor MR, Besser TE. 2000 Iha: a novel *Escherichia coli* O157:H7 adherence-conferring molecule encoded on a recently acquired chromosomal island of conserved structure. *Infect. Immun.* **68**, 1400–1407. (doi:10.1128/IAI.68.3.1400-1407.2000)
11. Hospenthal MK *et al.* 2016 Structure of a chaperone-usher pilus reveals the molecular basis of rod uncoiling. *Cell* **164**, 269–278. (doi:10.1016/j.cell.2015.11.049)
12. Busch A, Phan G, Waksman G. 2015 Molecular mechanism of bacterial type 1 and P pili assembly. *Phil. Trans. R. Soc. A* **373**, 20130153. (doi:10.1098/rsta.2013.0153)
13. Nuccio SP, Baumler AJ. 2007 Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol. Mol. Biol. Rev.* **71**, 551–575. (doi:10.1128/MMBR.00014-07)
14. Stubenrauch C *et al*. 2016 Effective assembly of fimbriae in *Escherichia coli* depends on the translocation assembly module nanomachine. *Nat. Microbiol.* **1**, 16064. (doi:10.1038/nmicrobiol.2016.64)
15. Doerrler WT, Raetz CR. 2005 Loss of outer membrane proteins without inhibition of lipid export in an *Escherichia coli* YaeT mutant. *J. Biol. Chem.* **280**, 27 679–27 687. (doi:10.1074/jbc.M504796200)
16. Werner J, Misra R. 2005 YaeT (Omp85) affects the assembly of lipid-dependent and lipid-independent outer membrane proteins of *Escherichia coli*. *Mol. Microbiol.* **57**, 1450–1459. (doi:10.1111/j.1365-2958.2005.04775.x)
17. Dunstan RA *et al.* 2015 Assembly of the secretion pores GspD, Wza and CsgG into bacterial outer membranes does not require the Omp85 proteins BamA or TamA. *Mol. Microbiol.* **97**, 616–629. (doi:10.1111/mmi.13055)
18. Heinz E, Selkrig J, Belousoff MJ, Lithgow T. 2015 Evolution of the translocation and assembly module (TAM). *Genome Biol Evol.* **7**, 1628–1643. (doi:10.1093/gbe/evv097)
19. Heinz E, Lithgow T. 2014 A comprehensive analysis of the Omp85/TpsB protein superfamily structural diversity, taxonomic occurrence, and evolution. *Front. Microbiol.* **5**, 370. (doi:10.3389/fmicb.2014.00370)
20. Selkrig J *et al*. 2012 Discovery of an archetypal protein transport system in bacterial outer membranes. *Nat. Struct. Mol. Biol.* **19**, 506–510. (doi:10.1038/nsmb.2261)
21. Heinz E, Stubenrauch CJ, Grinter R, Croft NP, Purcell AW, Strugnell RA, Dougan G, Lithgow T. 2016 Conserved features in the structure, mechanism, and biogenesis of the inverse autotransporter protein family. *Genome Biol. Evol.* **8**, 1690–1705. (doi:10.1093/gbe/evw112)
22. Selkrig J, Leyton DL, Webb CT, Lithgow T. 2014 Assembly of beta-barrel proteins into bacterial outer membranes. *Biochim. Biophys. Acta* **1843**, 1542–1550. (doi:10.1016/j.bbamcr.2013.10.009)
23. Wurpel DJ, Beatson SA, Totsika M, Petty NK, Schembri MA. 2013 Chaperone-usher fimbriae of *Escherichia coli*. *PLoS ONE* **8**, e52835. (doi:10.1371/journal.pone.0052835)
24. Del Canto F *et al*. 2016 Chaperone-usher pili loci of colonization factor-negative human enterotoxigenic *Escherichia coli*. *Front. Cell. Infect. Microbiol.* **6**, 200. (doi:10.3389/fcimb.2016.00200)
25. Robins-Browne RM, Holt KE, Ingle DJ, Hocking DM, Yang J, Tauschek M. 2016 Are *Escherichia coli* pathotypes still relevant in the era of whole-genome sequencing? *Front. Cell. Infect. Microbiol.* **6**, 141. (doi:10.3389/fcimb.2016.00141)
26. Hazen TH *et al*. 2016 Genomic diversity of EPEC associated with clinical presentations of differing severity. *Nat. Microbiol.* **1**, 15014. (doi:10.1038/nmicrobiol.2015.14)
27. Ingle DJ *et al*. 2016 Evolution of atypical enteropathogenic *E. coli* by repeated acquisition of LEE pathogenicity island variants. *Nat. Microbiol.* **1**, 15010. (doi:10.1038/nmicrobiol.2015.10)
28. Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA. 2013 Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **110**, 12 810–12 815. (doi:10.1073/pnas.1306836110)
29. Salipante SJ, Roach DJ, Kitzman JO, Snyder MW, Stackhouse B, Butler-Wu SM, Lee C, Cookson BT, Shendure J. 2015 Large-scale genomic sequencing of extraintestinal pathogenic *Escherichia coli* strains. *Genome Res.* **25**, 119–128. (doi:10.1101/gr.180190.114)
30. Letunic I, Bork P. 2016 Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245. (doi:10.1093/nar/gkw290)
31. Ben Zakour NL, Alsheikh-Hussain AS, Ashcroft MM, Khanh Nhu NT, Roberts LW, Stanton-Cook M, Schembri MA, Beatson SA. 2016 Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *MBio* **7**, e00347-16. (doi:10.1128/mBio.00347-16)
32. Mathers AJ, Peirano G, Pitout JD. 2015 The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant *Enterobacteriaceae*. *Clin. Microbiol. Rev.* **28**, 565–591. (doi:10.1128/CMR.00116-14)
33. Baumler AJ, Gilde AJ, Tsolis RM, van der Velden AW, Ahmer BM, Heffron F. 1997 Contribution of horizontal gene transfer and deletion events to development of distinctive patterns of fimbrial operons during evolution of *Salmonella* serotypes. *J. Bacteriol.* **179**, 317–322. (doi:10.1128/jb.179.2.317-322.1997)
34. Kelley LA, Sternberg MJ. 2009 Protein structure prediction on the Web: a case study using the Phyre server. *Nat. Protoc.* **4**, 363–371. (doi:10.1038/nprot.2009.2)
35. Coppens F, Iyyathurai J, Ruer S, Fioravanti A, Taganna J, Vereecke L, De Greve H, Remaut H. 2015 Structural and adhesive properties of the long polar fimbriae protein LpfD from adherent-invasive *Escherichia coli*. *Acta Crystallogr. D Biol. Crystallogr.* **71**, 1615–1626. (doi:10.1107/S1399004715009803)

36. Rasheed M, Garnett J, Perez-Dorado I, Muhl D, Filloux A, Matthews S. 2016 Crystal structure of the CupB6 adhesive tip from the chaperone-usher family of pili from *Pseudomonas aeruginosa*. *Biochim. Biophys. Acta* **1864**, 1500–1505. (doi:10.1016/j.bbapap.2016.07.010)
37. Klemm P, Schembri MA. 2000 Fimbrial surface display systems in bacteria: from vaccines to random libraries. *Microbiology* **146**, 3025–3032. (doi:10.1099/00221287-146-12-3025)
38. Welch RA *et al.* 2002 Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17 020–17 024. (doi:10.1073/pnas.252529799)
39. Staerk K, Khandige S, Kolmos HJ, Moller-Jensen J, Andersen TE. 2016 Uropathogenic *Escherichia coli* express type 1 fimbriae only in surface adherent populations under physiological growth conditions. *J. Infect. Dis.* **213**, 386–394. (doi:10.1093/infdis/jiv422)
40. Ong CL, Beatson SA, Totsika M, Forestier C, McEwan AG, Schembri MA. 2010 Molecular analysis of type 3 fimbrial genes from *Escherichia coli*, *Klebsiella* and *Citrobacter* species. *BMC Microbiol.* **10**, 183. (doi:10.1186/1471-2180-10-183)
41. Kisiela DI *et al*. 2012 Evolution of *Salmonella enterica* virulence via point mutations in the fimbrial adhesin. *PLoS Pathog.* **8**, e1002733. (doi:10.1371/journal.ppat.1002733)
42. Paul S, Linardopoulou EV, Billig M, Tchesnokova V, Price LB, Johnson JR, Chattopadhyay S, Sokurenko EV. 2013 Role of homologous recombination in adaptive diversification of extraintestinal *Escherichia coli*. *J. Bacteriol.* **195**, 231–242. (doi:10.1128/JB.01524-12)
43. Lamelas A *et al.* 2014 Emergence of a new epidemic *Neisseria meningitidis* serogroup A clone in the African meningitis belt: high-resolution picture of genomic changes that mediate immune evasion. *MBio* **5**, e01974-14. (doi:10.1128/mBio.01974-14)
44. Zhang X, Kupiec M, Gophna U, Tuller T. 2011 Analysis of coevolving gene families using mutually exclusive orthologous modules. *Genome Biol. Evol.* **3**, 413–423. (doi:10.1093/gbe/evr030)
45. Dorman CJ. 2004 H-NS: a universal regulator for a dynamic genome. *Nat. Rev. Microbiol.* **2**, 391–400. (doi:10.1038/nrmicro883)
46. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009 Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* **324**, 255–258. (doi:10.1126/science.1170160)
47. Park C, Zhang J. 2012 High expression hampers horizontal gene transfer. *Genome Biol. Evol.* **4**, 523–532. (doi:10.1093/gbe/evs030)
48. Engstrom MD, Mobley HL. 2016 Regulation of expression of uropathogenic *Escherichia coli* nonfimbrial adhesin TosA by PapB Homolog TosR in conjunction with H-NS and Lrp. *Infect. Immun.* **84**, 811–821. (doi:10.1128/IAI.01302-15)
49. Korea CG, Badouraly R, Prevost MC, Ghigo JM, Beloin C. 2010 *Escherichia coli* K-12 possesses multiple cryptic but functional chaperone-usher fimbriae with distinct surface specificities. *Environ. Microbiol.* **12**, 1957–1977. (doi:10.1111/j.1462-2920.2010.02202.x)
50. Selkrig J *et al*. 2015 Conserved features in TamA enable interaction with TamB to drive the activity of the translocation and assembly module. *Sci. Rep.* **5**, 12905. (doi:10.1038/srep12905)
51. Shen HH *et al*. 2014 Reconstitution of a nanomachine driving the assembly of proteins into bacterial outer membranes. *Nat. Commun.* **5**, 5078. (doi:10.1038/ncomms6078)
52. Celik N *et al*. 2012 A bioinformatic strategy for the detection, classification and analysis of bacterial autotransporters. *PLoS ONE* **7**, e43245. (doi:10.1371/journal.pone.0043245)
53. UniProt C. 2015 UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212. (doi:10.1093/nar/gku989)
54. Finn RD *et al.* 2016 The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285. (doi:10.1093/nar/gkv1344)
55. Eddy SR. 2011 Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195. (doi:10.1371/journal.pcbi.1002195)
56. Frickey T, Lupas A. 2004 CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* **20**, 3702–3704. (doi:10.1093/bioinformatics/bth444)
57. Kersey PJ *et al.* 2016 Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* **44**, D574–D580. (doi:10.1093/nar/gkv1209)
58. Katoh K, Standley DM. 2013 MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780. (doi:10.1093/molbev/mst010)
59. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009 trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973. (doi:10.1093/bioinformatics/btp348)
60. Stamatakis A, Hoover P, Rougemont J. 2008 A rapid bootstrap algorithm for the RAxML Web servers. *Syst. Biol.* **57**, 758–771. (doi:10.1080/10635150802429642)
61. Ronquist F *et al.* 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542. (doi:10.1093/sysbio/sys029)
62. Lartillot N, Lepage T, Blanquart S. 2009 PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25**, 2286–2288. (doi:10.1093/bioinformatics/btp368)
63. Seemann T. 2014 Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069. (doi:10.1093/bioinformatics/btu153)
64. Page AJ *et al.* 2015 Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693. (doi:10.1093/bioinformatics/btv421)
65. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016 SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056. (doi:10.1099/mgen.0.000056)
66. Edgar RC. 2010 Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461. (doi:10.1093/bioinformatics/btq461)
67. Edgar RC. 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797. (doi:10.1093/nar/gkh340)
68. Chang JM, Di Tommaso P, Notredame C. 2014 TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. *Mol. Biol. Evol.* **31**, 1625–1637. (doi:10.1093/molbev/msu117)
69. Leyton DL *et al*. 2014 A mortise-tenon joint in the transmembrane domain modulates autotransporter assembly into bacterial outer membranes. *Nat. Commun.* **5**, 4239. (doi:10.1038/ncomms5239)