Autotransporters of *Escherichia coli*: a sequence-based characterization

Timothy J. Wells, Makrina Totsika and Mark A. Schembri

Correspondence

Mark A. Schembri
m.schembri@uq.edu.au

Centre for Infectious Disease Research, School of Chemistry and Molecular Biosciences,
University of Queensland, Brisbane, Queensland 4072, Australia

Autotransporter (AT) proteins are found in all *Escherichia coli* pathotypes and are often associated with virulence. In this study we took advantage of the large number of available *E. coli* genome sequences to perform an in-depth bioinformatic analysis of AT-encoding genes. Twenty-eight *E. coli* genome sequences were probed using an iterative approach, which revealed a total of 215 AT-encoding sequences that represented three major groups of distinct domain architecture: (i) serine protease AT proteins, (ii) trimeric AT adhesins and (iii) AIDA-I-type AT proteins. A number of subgroups were identified within each broad category, and most subgroups contained at least one characterized AT protein; however, seven subgroups contained no previously described proteins. The AIDA-I-type AT proteins represented the largest and most diverse group, with up to 16 subgroups identified from sequence-based comparisons. Nine of the AIDA-I-type AT protein subgroups contained at least one protein that possessed functional properties associated with aggregation and/or biofilm formation, suggesting a high degree of redundancy for this phenotype. The Ag43, YfaL/EhaC, EhaB/UpaC and UpaG subgroups were found in nearly all *E. coli* strains. Among the remaining subgroups, there was a tendency for AT proteins to be associated with individual *E. coli* pathotypes, suggesting that they contribute to tissue tropism or symptoms specific to different disease outcomes.

Received 11 February 2010

Revised 30 April 2010

Accepted 4 May 2010

INTRODUCTION

The type V secretion pathway is the most widespread secretion pathway for the transport of molecules across the outer membrane of Gram-negative bacteria. This system comprises proteins secreted by the autotransporter (AT) pathway (also known as AT-1 or type Va), the two-partner secretion pathway (also known as type Vb) and the Oca system [also known as AT-2, type Vc or the trimeric AT adhesin (TAA) pathway] (Henderson *et al.*, 2004). AT proteins represent the largest group of type V secreted proteins. All AT proteins have several common features: an N-terminal signal sequence, a passenger (α) domain that often encodes a virulence function and is either anchored to the cell surface or released into the external milieu, and a translocation (β) domain that resides in the outer membrane (Jose *et al.*, 1995). AT proteins were originally thought to possess structural properties that facilitate their

independent transport across the bacterial membrane system and final routing to the cell surface (Desvaux *et al.*, 2004; Henderson *et al.*, 2004). However, this classical view has recently been called into question, as accessory factors such as the Bam complex (also known as the YaeT or Omp85 complex) as well as periplasmic chaperones such as SurA, Skp and DegP are required for the secretion of some AT proteins (Ieva & Bernstein, 2009; Ruiz-Perez *et al.*, 2009; Sauri *et al.*, 2009; Wagner *et al.*, 2009).

AT proteins from the type Va subgroup are generally similar with respect to the structure and length of the translocation module, whereas the passenger module can differ substantially and determines the functional characteristics of the protein. The translocation domain of AT proteins generally consists of 250–300 amino acid residues that are predicted to form 14 antiparallel β -strands (Loveless & Saier, 1997; Yen *et al.*, 2002) and inserts into the outer membrane as a β -barrel integral protein. The AT passenger domain confers the functional properties of the protein. The sequence of the passenger domain varies between different AT proteins, reflecting the diverse range of functions that AT proteins can confer, including adhesion (Benz & Schmidt, 1989), serum resistance (Fernandez & Weiss, 1994), haemagglutination, protease activity (Otto *et al.*, 1998), biofilm formation (Kjaergaard *et al.*, 2000) and toxin activity (Reyrat *et al.*, 1999). Although divergent in sequence, length and function, the

Abbreviations: AT, autotransporter; DEC, diarrhoeagenic *E. coli*; EAEC, enteroaggregative *E. coli*; EHEC, enterohaemorrhagic *E. coli*; EIEC, enteroinvasive *E. coli*; EPEC, enteropathogenic *E. coli*; ETEC, enterotoxigenic *E. coli*; ExPEC, extra-intestinal pathogenic *E. coli*; SPATE, serine protease autotransporters of *Enterobacteriaceae*; TAA, trimeric autotransporter adhesin; UPEC, uropathogenic *E. coli*.

Three supplementary tables, listing the composition of the SPATE group, the composition of the trimeric AT group and the composition of the AIDA-I AT group, are available with the online version of this paper.

passenger domain of almost all AT proteins is predicted to possess a β -helical structure, as exemplified by the well-characterized pertactin AT protein (Emsley *et al.*, 1994, 1996).

The type Vc AT proteins are also known as trimeric autotransporter adhesins (TAAs) (Cotter *et al.*, 2005; Linke *et al.*, 2006). All characterized TAAs possess a function associated with adhesion. TAAs are defined by the presence of a short 70–100 aa C-terminal membrane anchor domain that forms a trimer to create a full-sized β -barrel pore and facilitates the translocation of the passenger domain to the cell surface (Roggenkamp *et al.*, 2003; Surana *et al.*, 2004). The passenger region of TAAs is composed of two separate domains. The N-terminal head domain is primarily involved in the adhesion properties of the protein (Nummelin *et al.*, 2004; Yeo *et al.*, 2004). The stalk domain is repetitive, fibrous and highly divergent in length, with its primary function being to extend the head domain away from the surface of the bacterium (Linke *et al.*, 2006). It can, however, confer other functional properties upon the protein such as serum resistance (Roggenkamp *et al.*, 2003). The variable number of repeats in the stalk domain can lead to great differences in the size of TAAs. For example, BadA from *Bartonella henselae* is more than 3000 amino acids in length (Riess *et al.*, 2004), while YadA from *Yersinia enterocolitica* is only 422 amino acids (Hoiczky *et al.*, 2000).

Many of the AT proteins that have been characterized from *Escherichia coli* are associated with virulence. For example, the surface-associated AIDA-I and TibA AT proteins mediate epithelial cell adhesion (Benz & Schmidt, 1992), with TibA additionally being able to promote invasion (Elsinghorst & Weitz, 1994). Another *E. coli* AT protein, antigen 43 (Ag43), promotes biofilm formation (Hasman *et al.*, 1999; Kjaergaard *et al.*, 2000) as well as long-term colonization of the murine bladder (Ulett *et al.*, 2007). Despite extensive amino acid sequence differences in their passenger domain, the AIDA-I, TibA and Ag43 proteins all promote cell-to-cell aggregation (Klemm *et al.*, 2006; Sherlock *et al.*, 2004, 2005). A few TAA proteins have been characterized from *E. coli*, including UpaG from CFT073, which promotes cell-to-cell aggregation, biofilm formation and adhesion to fibronectin, laminin and bladder epithelial cells (Valle *et al.*, 2008), and the Eib group of proteins, which are present in some *E. coli* strains and mediate serum resistance (Sandt & Hill, 2000, 2001).

The rapid increase in genome sequence data available in the public domain enables the use of bioinformatic tools to identify genes that encode AT proteins. At the time of preparing this manuscript there were eight complete *E. coli* genomes and many more in draft format representing commensal strains, intestinal pathogens (i.e. diarrhoeagenic *E. coli*; DEC) and extra-intestinal pathogens (extra-intestinal pathogenic *E. coli*; ExPEC). There are six main subgroups of DEC: enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enteroinvasive *E. coli* (EIEC), enteroaggregative *E. coli* (EAEC), diffusely adhering

E. coli (DAEC) and enterohaemorrhagic *E. coli* (EHEC). ExPEC subgroups include uropathogenic *E. coli* (UPEC) and meningitis-associated *E. coli* (MNEC). In this study, 28 complete/draft *E. coli* genomes (representing commensal, DEC and ExPEC groups) were selected to search for genes encoding putative AT proteins. Three major AT groups of distinct domain architecture were identified and further divided into subgroups based on protein sequence comparisons.

METHODS

Identification of putative AT proteins. The BLAST program from NCBI was used to search 28 *E. coli* genomes (Table 1) using full-length *E. coli* AT protein sequences previously described in the literature. Protein sequences initially used were AIDA-I, Ag43, EaaA, EatA, EhaA-D, EibA-F, EspCP, Pet, PicU, RpeA, Sat, TibA, Tsh, UpaA-E and UpaG. Putative sequences were analysed using Pfam (<http://pfam.sanger.ac.uk/>) for the presence of either an Autotransporter or a YadA domain (PF03797 and PF03895, respectively). If the protein sequence contained one of these domains as well as an N-terminal signal sequence (as predicted by Signal 3.0) (Bendtsen *et al.*, 2004), they were deemed to be full-length AT protein sequences. All newly identified AT sequences were used to reprobe the *E. coli* genomes until no new sequences were found. Internal repeats were detected by Rapid Automatic Detection and Alignment of Repeats (RADAR; <http://www.ebi.ac.uk/Tools/Radar/index.html>).

Due to the repetitive nature of some of the uncharacterized AT protein sequences, positional orthologues from complete and incomplete genomes were aligned to deduce in a rudimentary fashion whether differences could be due to repeat collapses. As resequencing each AT protein is unviable, the conclusions drawn from these analyses were strongly based on complete genome AT protein sequences or took into account the possibility of repeat collapses.

Multiple sequence alignment and phylogenetic tree construction. Protein sequences were aligned using CLUSTAL_X (Thompson *et al.*, 1997) and edited using BioEdit. Phylogenetic distances were estimated using the neighbour-joining method of Saitou & Nei (1987), applying a distance matrix. The phylogenetic tree was drawn using MEGA 4 with incorporation of bootstrap values that were obtained involving 1000 replicates. AT proteins were divided into subgroups within each broad category of AT proteins; AT proteins within each subgroup contained greater than 70 % similarity in their primary amino acid sequence.

RESULTS AND DISCUSSION

Identification of AT-encoding sequences from selected *E. coli* genomes

Putative AT protein sequences were identified using an increasingly selective methodology. Initially, protein sequences of characterized *E. coli* ATs were obtained (see Methods). These were then used to probe the 28 selected *E. coli* genomes (this included associated plasmids; Table 1) using protein BLAST. Sequences identified by BLAST searches were selected only if they possessed an Autotransporter Pfam (PF03797) or a YadA Pfam domain (PF03895), corresponding to the translocation domain of type Va and

Table 1. Numbers of genes in each *E. coli* genome that encode putative AT proteins

<i>E. coli</i> strain	Total number of putative full-length AT proteins	No. of SPATEs	No. of TAAs	No. of AIDA-I-type ATs	Source or reference
EAEC strains					
101-1	4	0	1	3	TIGR*
55989	10	2	1	7	Touchon <i>et al.</i> (2009)
EIEC strain					
53638	7	0	1	6	TIGR
ETEC strains					
B7A	8	1	1	6	TIGR
E24377A	9	1	1	7	TIGR
EPEC strains					
B171	10	0	1	9	TIGR
E110019	7	0	1	6	TIGR
E22	10	2	1	7	TIGR
O127:H6 E2348/69	4	1	1	2	Iguchi <i>et al.</i> (2009)
EHEC strains					
O157:H7 EDL933	9	1	1	7	Perna <i>et al.</i> (2001)
O157:H7 Sakai	9	1	1	7	Makino <i>et al.</i> (1999)
Laboratory strains					
K12 MG1655	5	0	0	5	Blattner <i>et al.</i> (1997)
K12 W3110	5	0	0	5	Nara Institute of Science and Technology
Commensal or environmental strains					
ATCC 8739	6	0	0	6	DOE Joint Genome Institute
B	5	0	0	5	International <i>E. coli</i> B consortium
ED1a	10	2	5	3	Touchon <i>et al.</i> (2009)
HS	7	0	1	6	TIGR
IAI1	5	0	1	4	Touchon <i>et al.</i> (2009)
SE11	8	0	1	7	Kitasato University
SMS-3-5	12	1	3	8	TIGR
UPEC strains					
536	8	1	1	6	Hochhut <i>et al.</i> (2006)
CFT073	11	3	1	7	Welch <i>et al.</i> (2002)
F11	8	1	1	6	TIGR
UTI89	6	0	1	5	Chen <i>et al.</i> (2006)
ExPEC strains					
APEC01	9	2	2	5	Johnson <i>et al.</i> (2006)
IAI39	6	0	1	5	Touchon <i>et al.</i> (2009)
S88	7	1	1	5	Touchon <i>et al.</i> (2009)
UMN026	10	1	1	8	Touchon <i>et al.</i> (2009)

type Vc (TAA) ATs, respectively. An iterative approach was employed such that as new sequences were identified they were used to reprobe the *E. coli* genomes until no new sequences were found. All sequences obtained were then further analysed to ensure they contained an N-terminal signal sequence, thus removing partial and truncated sequences. This analysis also identified many sequences that were mis-annotated, the most likely start codon being further upstream on the chromosome. These mis-annotated sequences were manually fixed and added to the AT protein set. Finally, the selected sequences were examined for additional Pfam domains and motifs, with the vast majority containing the Pertactin Pfam (PF03212) motif.

In total, 215 AT proteins were identified from the 28 *E. coli* genome sequences (Table 1). The majority of AT proteins belonged to the type Va subgroup. Due to the highly repetitive nature of the sequence encoding the predicted passenger domain of some of these proteins, we cannot rule out the possibility that repeat collapses (i.e., where a series of repeats is arranged as a single sequence) may result in shortening of some genes or misassembly. This is especially the case for multiple tandem repeats, where it can be difficult to distinguish an incorrectly collapsed repeat from true polymorphisms (Eichler, 2001; Salzberg & Yorke, 2005). However, the presence of homologues with the same predicted amino acid length from multiple

genomes suggests that it is unlikely that repeat collapses have taken place.

Multiple sequence alignment and grouping of AT proteins

Using the full-length protein sequences of all ATs obtained, we constructed a multiple sequence alignment and an unrooted phylogram (Fig. 1). The neighbour-joining

analysis revealed three broad groups that reflected the three domain architecture-based subfamilies described in the literature: (i) the serine protease autotransporters of the *Enterobacteriaceae* (SPATEs) (Henderson *et al.*, 1998) (Supplementary Table S1), (ii) the TAAs (Cotter *et al.*, 2005) (Supplementary Table S2) and (iii) the AIDA-I-type AT proteins (Henderson *et al.*, 2004) (Supplementary Table S3). The three groups have distinct domain organizations. The TAAs have the most distinct motif arrangement

Fig. 1. (a) Unrooted phylogram of 230 AT proteins, 215 identified from 28 *E. coli* genomes and 16 from other *E. coli* strains. Proteins are separated into three major groups based on protein domains: SPATE (red shading), TAA (yellow shading) and the remaining proteins in the broad group of AIDA-I-type AT proteins (blue shading). Subgroups are named after the best-characterized representative from that group. Values in parentheses indicate the number of proteins in each group. Clades with no characterized members are given numbers and further details are given in Table 3. Bootstrap values from 1000 replicates are shown for important nodes. For a detailed list of all protein sequences used in the analysis refer to Supplementary Tables S1, S2 and S3. (b) General Pfam domain organization found in the three major groups of *E. coli* AT proteins.

due to the possession of a shorter C-terminal domain (Pfam: YadA in place of Pfam: Autotransporter). The SPATE proteins contain a full-length Autotransporter Pfam domain, as well as a Pertactin Pfam domain; all SPATE proteins contain a Peptidase_S6 Pfam domain in the passenger module (Fig. 1b). The remainder of the AT proteins fell into the broad and loosely defined subfamily of AIDA-I-type AT proteins. These proteins, typified by the well-characterized AIDA-I and Ag43 proteins, contain an Autotransporter Pfam domain and a Pertactin Pfam domain (Henderson *et al.*, 2004). Our analysis identified 16 subgroups containing AT proteins with greater than 70 %

amino acid identity within the AIDA-I subfamily (Fig. 1, Supplementary Table S3). In the majority of cases, the subgroups within each of the three broadly defined AT groups contained at least one AT protein that had been previously characterized and this was used to identify the subgroup (Table 2). Where no characterized AT proteins were present, the subgroup was named after the provisional locus tag given to the corresponding gene in *E. coli* MG1655. There were, however, seven new subgroups that did not clade closely to any characterized or annotated proteins (Table 3). TibA and AIDA-I were not found in any of the genomes screened.

Table 2. Functional properties of AT proteins characterized from *E. coli*

AT	Strain	Characterized virulence properties	References
SPATEs			
EatA	H10407	Serine protease activity	Patel <i>et al.</i> (2004)
EpeA	O111:H21 EH41	Protease and mucinase activity	Leyton <i>et al.</i> (2003)
EspC	E2368/69	Enterotoxin	Mellies <i>et al.</i> (2001); Stein <i>et al.</i> (1996)
EspP	O157:H7	Proteolytic toxin	Brunder <i>et al.</i> (1997); Djafari <i>et al.</i> (1997)
Pet	O42	Proteolytic toxin	Eslava <i>et al.</i> (1998)
PicU	CFT073	Mucinase	Parham <i>et al.</i> (2004)
RpeA	O15:H – 83/89	Colonization factor	Leyton <i>et al.</i> (2007)
Sat	CFT073	Proteolytic toxin	Guyer <i>et al.</i> (2000, 2002)
Tsh	APECO1	Haemagglutinin/haemoglobin binding	Otto <i>et al.</i> (1998); Provence & Curtiss (1994)
TAAAs			
EibA-G	ECOR-9	Immunoglobulin binding and serum resistance	Sandt & Hill (2000, 2001)
Saa	O113:H21 98NK2	Adhesion to Hep-2 cells, autoaggregation	Paton <i>et al.</i> (2001)
UpaG	CFT073	Adhesion to T24 cells, biofilm formation, autoaggregation, adhesion to ECM	Valle <i>et al.</i> (2008)
AIDA-I-type ATs			
AatA	APECO1	Adhesion to chicken embryo fibroblast cells, contribution to APEC virulence	Li <i>et al.</i> (2010)
AIDA-I	O126:H27	Adhesion, autoaggregation, biofilm formation, self-recognition	Benz & Schmidt (1989); Sherlock <i>et al.</i> (2004)
Ag43	MG1655	Biofilm formation, autoaggregation, self-recognition, phase-variable	Kjaergaard <i>et al.</i> (2000); Klemm <i>et al.</i> (2004); Torres <i>et al.</i> (2002); Ulett <i>et al.</i> (2006)
EhaA	EDL933	Adhesion to primary bovine cell lines, autoaggregation, biofilm formation	Wells <i>et al.</i> (2008)
EhaB/UpaC	EDL933/CFT073	Biofilm formation, adhesion to laminin and collagen I	Parham <i>et al.</i> (2004); Wells <i>et al.</i> (2008)
EhaD/YpjA	EDL933	Biofilm formation	Wells <i>et al.</i> (2008)
EhaD	EDL933	Biofilm formation	Wells <i>et al.</i> (2008)
UpaB	CFT073	Not yet characterized	Parham <i>et al.</i> (2004)
UpaE	CFT073	Not yet characterized	Parham <i>et al.</i> (2004)
UpaH	CFT073	Biofilm formation	Allsopp <i>et al.</i> (2010)
TibA	O78:H11	Adhesion/invasion, autoaggregation, biofilm formation, self-recognition	Lindenthal & Elsinghorst (2001); Sherlock <i>et al.</i> (2005b)
YcgV	MG1655	Biofilm formation	Roux <i>et al.</i> (2005)
YdeK	MG1655	Psuedogene in K-12. Not characterized	
YejO	MG1655	Not yet characterized	Roux <i>et al.</i> (2005)
YfaL/EhaC	MG1655/EDL933	YfaL increases biofilm formation	Parham <i>et al.</i> (2004); Wells <i>et al.</i> (2008); Roux <i>et al.</i> (2005)

Table 3. Putative new AT proteins that do not clade with previously characterized AT proteins

Number	NCBI locus tag	Characteristics
1	EcSMS35_3187	369 aa total length, 28 aa signal peptide, one HIM domain, YadA domain
2	ECED1_2077	517 aa total length, 25 aa signal peptide, one HIM and one Hep_hag domain, YadA domain
3	ECED1_1141	359 aa total length, 26 aa signal peptide, YadA domain
4	ECUMN_0533	600 aa total length, 29 aa signal peptide, YadA domain
5	APEC01_1523, ECIAI39_0467, EcSMS35_0533, E2348C_0425,	338 aa total length, 23 aa signal peptide, YadA domain
6	E2348C_2704	1029 aa total length, 23 aa signal peptide, Pertactin domain and Autotransporter domain
7	EC55989_1263, ECSE_1215, ECIAI39_1904, EcSMS35_1980	882 aa total length, 23 aa signal peptide, Pertactin domain and Autotransporter domain

SPATE AT proteins

SPATE proteins are a subgroup of AT proteins that possess a consensus serine protease domain (Henderson *et al.*, 1998). They have been identified in many pathogenic *E. coli* and are strongly associated with virulence. SPATE proteins are among some of the best-characterized AT proteins of *E. coli*, and possess a range of functional properties including cytotoxicity (Brunder *et al.*, 1997; Eslava *et al.*, 1998; Guyer *et al.*, 2002; Mellies *et al.*, 2001), mucinase activity (Parham *et al.*, 2004), haemoglobin degradation (Otto *et al.*, 1998; Provence & Curtiss, 1994) and promotion of intestinal colonization (Leyton *et al.*, 2007) (Table 3). Some SPATE proteins are located on plasmids; examples include Pet (Eslava *et al.*, 1998) and EspP (Brunder *et al.*, 1997). All of the SPATE proteins identified in this analysis demonstrated greater than 70 % similarity to previously characterized SPATE proteins.

TAAAs

UpaG, Saa and the Eib proteins are the only TAAAs that have previously been characterized from *E. coli* (Lu *et al.*, 2006; Paton *et al.*, 2001; Sandt & Hill, 2000, 2001; Valle *et al.*, 2008). These proteins all possess a common C-terminal region consisting of the Pfam: YadA domain, but vary greatly in length. UpaG consists of 1778 amino acids, while the Saa and Eib TAAAs range between 392 and 535 residues.

The majority of TAA-encoding genes identified from the *E. coli* genomes formed one group with the characterized UpaG. UpaG was recently characterized from UPEC strain CFT073 and mediates cell aggregation, biofilm formation and adhesion to fibronectin, laminin and human bladder epithelial cells (Valle *et al.*, 2008). A putative UpaG-encoding gene was found in 21 of the 28 *E. coli* genomes examined. It was absent in the two K-12 strains (MG1655 and W3110), two commensal/environmental strains (commensal B and ATCC 8739) and the EPEC O127:H6 strain E2348/69; the gene was truncated by an internal stop codon in ExPEC strains IAI39 and UMN026. In the commensal *E. coli* B REL606 strain, UpaG is a

pseudogene; however, as this genome is incomplete it may be due to repeat collapsing. The UpaG homologues formed three distinct clades that essentially reflect UPEC, DEC and commensal groups. One exception was UpaG from EIEC strain 53638, which shared the highest sequence similarity with the homologues from commensal strains. The major difference among representatives of each of the UpaG clades was the order, number and type of repeats in the passenger domain. For example, UpaG from UPEC strain CFT073 had almost 200 extra amino acids (mainly associated with repeats) in its passenger domain compared with the homologous protein in EHEC strain EDL933.

Genes encoding the Eib TAA proteins were originally identified in a small selection of *E. coli* strains from the ECOR collection (Sandt *et al.*, 1997). There are several Eib paralogues, many of which are encoded within the same *E. coli* strain (Sandt & Hill, 2000). Eib proteins bind immunoglobulin and mediate serum resistance (Lu *et al.*, 2006; Sandt & Hill, 2000, 2001). They possess the conserved YadA domain but contain a short passenger domain which accounts for their distinct clustering from UpaG (Fig. 1). Two genes that encoded an EibA homologue were identified in the *E. coli* ED1a genome; however, no Eib homologues were found in any of the other genome-sequenced *E. coli* strains. The *saa* gene is encoded on a megaplasmid in the locus of enterocyte effacement (LEE)-negative EHEC O113:H21 strain 98NK2 (Paton *et al.*, 2001). Although it has been identified in other LEE-negative EHEC strains, it was not found in any of the genome-sequenced strains examined in this study.

This analysis also identified four novel TAA-encoding genes that did not clade closely to any of the characterized TAA proteins (Table 3). Two of these genes were from *E. coli* ED1a, with the other two found in *E. coli* SMS-3-5 and UMN026. Interestingly, the analysis also identified a novel group of four TAA-encoding genes from *E. coli*, APEC01, IAI39, SMS35 and E2348/69, that shared 96 % amino acid identity (Group 5, Table 3). These TAA proteins may represent novel adhesins.

AIDA-I-type AT proteins

The largest and most diverse collection of *E. coli* AT proteins is contained within the AIDA-I group (Henderson *et al.*, 2004). The passenger domain of the AIDA-I group of AT proteins does not contain a common Pfam domain and displays a wide distribution of lengths, ranging from 776 to 2771 amino acids (Fig. 2). Despite this, one common feature of all AIDA-I AT proteins is the presence of short internal repeats in the passenger domain that are between 16 and 24 amino acids in length. To determine whether all AIDA-I-type AT proteins had internal repeats, RADAR software was utilized. The program identified short internal repeats in the passenger domain from all of the AT proteins examined; however, these repeat sequences were not conserved among AIDA-I AT proteins from different subgroups. Some of the repeats are recognized by Pfam as small motifs, such as Chlam PMP, Fil_haemagg, Dicty_cad and Hep_Hag; however, even within most groups these Pfam motifs were not consistent. Only one group (YfaL) had consistent Pfam motifs in the passenger domain with all sequences containing five Chlam PMP regions.

The precise function of these repetitive motifs has not been investigated; however, repetitive sequence motifs are often associated with proteins that have adhesive functions (Henderson & Nataro, 2001). These repetitive motifs may also contribute to the tertiary structure of the AT proteins. Although no tertiary structure has been deduced for any of these AT proteins, the structure for AIDA-I, TibA and Ag43 has been modelled based on their similarity to P.69 pertactin from *Bordetella pertussis*. Pertactin has regions of small repeats that form a right-handed parallel β -helix

(Emsley *et al.*, 1996), and it is likely that Ag43, AIDA-I and TibA share this structure (Kajava *et al.*, 2001; Klemm *et al.*, 2004, 2006).

Common functions of AIDA-I-type proteins

The passenger module of AT proteins within the AIDA-I subgroups displayed extensive sequence diversity, despite the fact that some of these proteins share common functions. For example, the full-length AIDA-I, TibA and Ag43 proteins share approximately 25 % similarity (8.3 % identity) at the amino acid level, yet have previously been grouped together as self-associating autotransporters (SAATs) due to their ability to mediate cell-to-cell aggregation and biofilm formation (Klemm *et al.*, 2006). EhaA, EhaB, EhaD and UpaH also possess properties associated with cell aggregation or biofilm formation (Table 2) (Allsopp *et al.*, 2010; Wells *et al.*, 2008). YfaL (EhaC homologue), YpjA (EhaD homologue) and YcgV increase biofilm formation when overexpressed in *E. coli* MG1655 (Roux *et al.*, 2005). Many of the AT protein subgroups with no known function (UpaB, YdeK, UpaE) contain the Pfam repeat motifs Fil_haemagg or Dicty_cad, both of which are associated with proteins that aggregate (Loomis & Fuller, 1990; Relman *et al.*, 1989), and thus these AT proteins may possess similar functions. Only one subgroup (YejO) has no characterized or Pfam motif-based association with aggregation or biofilm formation. Previous work has also shown that Ag43 can recognize AIDA-I and TibA to form mixed-cell aggregates (Klemm *et al.*, 2006; Sherlock *et al.*, 2004). Thus, it is possible that heterologous interactions may occur with other AT proteins within the AIDA-I subgroups.

Fig. 2. Domain organization and length of a representative member from each AIDA-I-type AT protein subgroup.

Although many of the characterized AT proteins described above show common functions, including biofilm formation and aggregation, it is important to note that these proteins often possess additional functions. For example, TibA, AIDA-I, Ag43, EhaA, EhaB and AatA all mediate adhesion to epithelial cells (Benz & Schmidt, 1989; Henderson & Owen, 1999; Li *et al.*, 2010; Lindenthal & Elsinghorst, 2001; Wells *et al.*, 2008) or extracellular matrix (ECM) proteins (Wells *et al.*, 2009). All of the *E. coli* genomes examined in this study have multiple AIDA-I-type AT-encoding genes (between four and 10), and functional redundancy of these proteins may contribute to adhesion and colonization of epithelial cells.

AIDA-I-type AT groups associated with specific *E. coli* pathotypes

A closer examination of the AIDA-I subgroups revealed that some AT protein sequences were present in the majority of the *E. coli* genomes investigated, while others were only found in a select number of genomes. The most common AT protein sequences identified from the 28 *E. coli* genomes were YfaL/EhaC (28 strains), EhaB/UpaC (23 strains) and Ag43 (19 strains). The proteins within these subgroups displayed significant sequence variation in their passenger domain, with ExPEC/UPEC and DEC homologues tending to clade separately (Fig. 1, Supplementary Table S2). In the case of Ag43, there were also differences in the copy number of the Ag43-encoding *flu* gene; for example, the K-12 strain MG1655 had one copy, whereas the EPEC strain B171 had four copies (Supplementary Table S2). In the genome-sequenced UPEC strain CFT073, two variant *flu* genes are present. The genes are located within distinct pathogenicity islands and possess different functional properties with respect to aggregation, biofilm formation and long-term colonization of the mouse urinary tract (Klemm *et al.*, 2004; Ulett *et al.*, 2007). These findings were recently corroborated by comparative prevalence analysis of the *flu* gene variants in clinical urinary tract infection (UTI) strains (Luthje & Brauner, 2010). The UpaH group displayed the most sequence diversity; however, this could be due to repeat collapses in the genome assembly of the strains examined (Allsopp *et al.*, 2010).

The remaining subgroups of AIDA-I-type AT proteins had representatives in only a portion of the *E. coli* genomes. Interestingly, many of these subgroups were unique to either DEC strains or UPEC strains. The YcgV, YejO, EhaD, YdeK and UpaB homologue subgroups all displayed some degree of association with *E. coli* pathotypes and had conserved passenger domain sequences. The YcgV-encoding gene was found in the majority of commensal and DEC genomes (it was truncated in both EHEC strains); however, the entire gene was missing from every single UPEC genome. Similarly, the YejO group of AT proteins was found in all commensal and DEC genomes but was truncated in all UPEC genomes. The EhaD-encoding gene

was also absent from UPEC genomes. A full-length YdeK-encoding gene was identified in fewer than half of the genomes; *ydeK* is a pseudogene in MG1655 and EHEC strains and was absent from all UPEC strains. Conversely, the UpaB group (encoded by the *upaB* gene) was present in all UPEC genomes (as well as several commensal and ExPEC strains) but was either missing or truncated in all DEC genomes. It may be that the pathotype-associated AT proteins have functions specific to their environment, although a larger number of *E. coli* genomes would need to be investigated to establish a strong link to pathogenesis. Two AT protein groups, however, do lend support to this hypothesis, namely EhaA and AatA.

The subgroup represented by EhaA was identified in nine strains from the commensal and diarrhoeagenic *E. coli* types, but not in any UPEC/ExPEC strains. The *ehaA* gene resides in a unique genomic island and this sequence context was conserved in all of the *ehaA*-positive strains examined. The EhaA primary amino acid sequence was highly conserved, suggesting a similar function for all EhaA proteins. EhaA from EHEC EDL933 promotes adhesion to primary epithelial cells of the bovine terminal rectum (Wells *et al.*, 2008), the principal site of colonization by *E. coli* O157:H7 (Naylor *et al.*, 2003). This function is specific for *E. coli* O157:H7 strains and may explain the absence of the gene in ExPEC genomes. Similarly, AatA was only identified in the avian pathogenic *E. coli* (APEC) strain APEC01 (Li *et al.*, 2010). AatA enhances adhesion of APEC to chicken embryo fibroblast cells and contributes to APEC virulence (Li *et al.*, 2010).

Conclusions

The 28 genome-sequenced *E. coli* strains examined in this study possess multiple (between four and 12) AT protein-encoding genes; most strains have at least one TAA and multiple AIDA-I-type AT-encoding genes. Many *E. coli* strains also contain a gene encoding a member of the virulence-associated SPATE group. Four AT subgroups represented by Ag43, YfaL/EhaC, EhaB/UpaC and UpaG were found in most of the strains examined and thus may possess functions conserved among all *E. coli* strains. In contrast, other AT proteins were only found in particular *E. coli* pathotypes and thus may have more specific functions. Seven new AT subgroups that contained no previously described proteins were also identified. Further research is now required to ascertain the function of many of these AT proteins in *E. coli* pathogenesis.

ACKNOWLEDGEMENTS

This work was supported by grants from the Australian Research Council (DP1097032) and the University of Queensland. T.J.W. is the recipient of a University of Queensland Graduate Scholarship.

REFERENCES

- Allsopp, L. P., Totsika, M., Tree, J. J., Ulett, G. C., Mabbett, A. N., Wells, T. J., Kobe, B., Beatson, S. B. & Schembri, M. A. (2010). UpaH is a newly identified autotransporter protein that contributes to biofilm formation and bladder colonization by uropathogenic *Escherichia coli* CFT073. *Infect Immun* **78**, 1659–1669.
- Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* **340**, 783–795.
- Benz, I. & Schmidt, M. A. (1989). Cloning and expression of an adhesin (Aida-I) involved in diffuse adherence of enteropathogenic *Escherichia coli*. *Infect Immun* **57**, 1506–1511.
- Benz, I. & Schmidt, M. A. (1992). Isolation and serologic characterization of Aida-I, the adhesin mediating the diffuse adherence phenotype of the diarrhea-associated *Escherichia coli* strain-2787 (O126:H27). *Infect Immun* **60**, 13–18.
- Blattner, F. R., Plunkett, G., III, Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K. & other authors (1997). The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1462.
- Brunner, W., Schmidt, H. & Karch, H. (1997). EspP, a novel extracellular serine protease of enterohaemorrhagic *Escherichia coli* O157:H7 cleaves human coagulation factor V. *Mol Microbiol* **24**, 767–778.
- Chen, S. L., Hung, C.-S., Xu, J., Reigstad, C. S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R. R. & other authors (2006). Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc Natl Acad Sci U S A* **103**, 5977–5982.
- Cotter, S. E., Surana, N. K. & St Geme, J. W. (2005). Trimeric autotransporters: a distinct subfamily of autotransporter proteins. *Trends Microbiol* **13**, 199–205.
- Desvaux, M., Parham, N. J. & Henderson, I. R. (2004). The autotransporter secretion system. *Res Microbiol* **155**, 53–60.
- Djafari, S., Ebel, F., Deibel, C., Kramer, S., Hudel, M. & Chakraborty, T. (1997). Characterization of an exported protease from Shiga toxin-producing *Escherichia coli*. *Mol Microbiol* **25**, 771–784.
- Eichler, E. E. (2001). Segmental duplications: what's missing, misassigned, and misassembled – and should we care? *Genome Res* **11**, 653–656.
- Elsinghorst, E. A. & Weitz, J. A. (1994). Epithelial-cell invasion and adherence directed by the enterotoxigenic *Escherichia coli* *tib* locus is associated with a 104-kilodalton outer-membrane protein. *Infect Immun* **62**, 3463–3471.
- Emsley, P., McDermott, G., Charles, I. G., Fairweather, N. F. & Isaacs, N. W. (1994). Crystallographic characterization of pertactin, a membrane-associated protein from *Bordetella pertussis*. *J Mol Biol* **235**, 772–773.
- Emsley, P., Charles, I. G., Fairweather, N. F. & Isaacs, N. W. (1996). Structure of *Bordetella pertussis* virulence factor P.69 pertactin. *Nature* **381**, 90–92.
- Eslava, C., Navarro-Garcia, F., Czeckulin, J. R., Henderson, I. R., Cravioto, A. & Nataro, J. P. (1998). Pet, an autotransporter enterotoxin from enteroaggregative *Escherichia coli*. *Infect Immun* **66**, 3155–3163.
- Fernandez, R. C. & Weiss, A. A. (1994). Cloning and sequencing of a *Bordetella pertussis* serum resistance locus. *Infect Immun* **62**, 4727–4738.
- Guyer, D. M., Henderson, I. R., Nataro, J. P. & Mobley, H. L. T. (2000). Identification of Sat, an autotransporter toxin produced by uropathogenic *Escherichia coli*. *Mol Microbiol* **38**, 53–66.
- Guyer, D. M., Radulovic, S., Jones, F. E. & Mobley, H. L. T. (2002). Sat, the secreted autotransporter toxin of uropathogenic *Escherichia coli*, is a vacuolating cytotoxin for bladder and kidney epithelial cells. *Infect Immun* **70**, 4539–4546.
- Hasman, H., Chakraborty, T. & Klemm, P. (1999). Antigen-43-mediated autoaggregation of *Escherichia coli* is blocked by fimbriation. *J Bacteriol* **181**, 4834–4841.
- Henderson, I. R. & Nataro, J. P. (2001). Virulence functions of autotransporter proteins. *Infect Immun* **69**, 1231–1243.
- Henderson, I. R. & Owen, P. (1999). The major phase-variable outer membrane protein of *Escherichia coli* structurally resembles the immunoglobulin A1 protease class of exported protein and is regulated by a novel mechanism involving Dam and OxyR. *J Bacteriol* **181**, 2132–2141.
- Henderson, I. R., Navarro-Garcia, F. & Nataro, J. P. (1998). The great escape: structure and function of the autotransporter proteins. *Trends Microbiol* **6**, 370–378.
- Henderson, I. R., Navarro-Garcia, F., Desvaux, M., Fernandez, R. C. & Ala'Aldeen, D. (2004). Type V protein secretion pathway: the autotransporter story. *Microbiol Mol Biol Rev* **68**, 692–744.
- Hochhut, B., Wilde, C., Balling, G., Middendorf, B., Dobrindt, U., Brzuszkiewicz, E., Gottschalk, G., Carniel, E. & Hacker, J. (2006). Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic *Escherichia coli* strain 536. *Mol Microbiol* **61**, 584–595.
- Hoiczky, E., Roggenkamp, A., Reichenbecher, M., Lupas, A. & Heesemann, J. (2000). Structure and sequence analysis of *Yersinia* YadA and *Moraxella* UspAs reveal a novel class of adhesins. *EMBO J* **19**, 5989–5999.
- Ieva, R. & Bernstein, H. D. (2009). Interaction of an autotransporter passenger domain with BamA during its translocation across the bacterial outer membrane. *Proc Natl Acad Sci U S A* **106**, 19120–19125.
- Iguchi, A., Thomson, N. R., Ogura, Y., Saunders, D., Ooka, T., Henderson, I. R., Harris, D., Asadulghani, M., Kurokawa, K. & other authors (2009). Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. *J Bacteriol* **191**, 347–354.
- Johnson, T. J., Wannemuehler, Y. M., Scaccianoce, J. A., Johnson, S. J. & Nolan, L. K. (2006). Complete DNA sequence, comparative genomics, and prevalence of an IncHI2 plasmid occurring among extraintestinal pathogenic *Escherichia coli* isolates. *Antimicrob Agents Chemother* **50**, 3929–3933.
- Jose, J., Jahng, F. & Meyer, T. F. (1995). Common structural features of Iga1 protease-like outer-membrane protein autotransporters. *Mol Microbiol* **18**, 378–380.
- Kajava, A. V., Cheng, N., Cleaver, R., Kessel, M., Simon, M. N., Willery, E., Jacob-Dubuisson, F., Locht, C. & Steven, A. C. (2001). Beta-helix model for the filamentous haemagglutinin adhesin of *Bordetella pertussis* and related bacterial secretory proteins. *Mol Microbiol* **42**, 279–292.
- Kjaergaard, K., Schembri, M. A., Ramos, C., Molin, S. & Klemm, P. (2000). Antigen 43 facilitates formation of multispecies biofilms. *Environ Microbiol* **2**, 695–702.
- Klemm, P., Hjerrild, L., Gjermansen, M. & Schembri, M. A. (2004). Structure–function analysis of the self-recognizing antigen 43 autotransporter protein from *Escherichia coli*. *Mol Microbiol* **51**, 283–296.
- Klemm, P., Vejborg, R. M. & Sherlock, O. (2006). Self-associating autotransporters, SAATs: functional and structural similarities. *Int J Med Microbiol* **296**, 187–195.

- Leyton, D. L., Sloan, J., Hill, R. E., Doughty, S. & Hartland, E. L. (2003). Transfer region of pO113 from enterohemorrhagic *Escherichia coli*: similarity with R64 and identification of a novel plasmid-encoded autotransporter, EpeA. *Infect Immun* **71**, 6307–6319.
- Leyton, D. L., Adams, L. M., Kelly, M., Sloan, J., Tauschek, M., Robins-Browne, R. M. & Hartland, E. L. (2007). Contribution of a novel gene, *rpeA*, encoding a putative autotransporter adhesin to intestinal colonization by rabbit-specific enteropathogenic *Escherichia coli*. *Infect Immun* **75**, 4664–4669.
- Li, G., Feng, Y., Kariyawasam, S., Tivendale, K. A., Wannemuehler, Y., Zhou, F., Logue, C. M., Miller, C. L. & Nolan, L. K. (2010). AatA is a novel autotransporter and virulence factor of avian pathogenic *Escherichia coli*. *Infect Immun* **78**, 898–906.
- Lindenthal, C. & Elsinghorst, E. A. (2001). Enterotoxigenic *Escherichia coli* TibA glycoprotein adheres to human intestine epithelial cells. *Infect Immun* **69**, 52–57.
- Linke, D., Riess, T., Autenrieth, I. B., Lupas, A. & Kempf, V. A. J. (2006). Trimeric autotransporter adhesins: variable structure, common function. *Trends Microbiol* **14**, 264–270.
- Loomis, W. F. & Fuller, D. L. (1990). A pair of tandemly repeated genes code for gp24, a putative adhesion protein of *Dictyostelium discoideum*. *Proc Natl Acad Sci U S A* **87**, 886–890.
- Loveless, B. J. & Saier, M. H. (1997). A novel family of channel-forming, autotransporting, bacterial virulence factors. *Mol Membr Biol* **14**, 113–123.
- Lu, Y., Iyoda, S., Satou, H., Satou, H., Itoh, K., Saitoh, T. & Watanabe, H. (2006). A new immunoglobulin-binding protein, EibG, is responsible for the chain-like adhesion phenotype of locus of enterocyte effacement-negative, Shiga toxin-producing *Escherichia coli*. *Infect Immun* **74**, 5747–5755.
- Luthje, P. & Brauner, A. (2010). Putative link between the virulence-associated *fluA* gene and fluoroquinolone resistance in uropathogenic *Escherichia coli*. *J Clin Microbiol* **48**, 675–676.
- Makino, K., Yokoyama, K., Kubota, Y., Yutsudo, C. H., Kimura, S., Kurokawa, K., Ishii, K., Hattori, M., Tatsuno, I. & other authors (1999). Complete nucleotide sequence of the prophage VT2-Sakai carrying the verotoxin 2 genes of the enterohemorrhagic *Escherichia coli* O157:H7 derived from the Sakai outbreak. *Genes Genet Syst* **74**, 227–239.
- Mellies, J. L., Navarro-Garcia, F., Okeke, I., Frederickson, J., Nataro, J. P. & Kaper, J. B. (2001). *espC* pathogenicity island of enteropathogenic *Escherichia coli* encodes an enterotoxin. *Infect Immun* **69**, 315–324.
- Naylor, S. W., Low, J. C., Besser, T. E., Mahajan, A., Gunn, G. J., Pearce, M. C., McKendrick, I. J., Smith, D. G. & Gally, D. L. (2003). Lymphoid follicle-dense mucosa at the terminal rectum is the principal site of colonization of enterohemorrhagic *Escherichia coli* O157:H7 in the bovine host. *Infect Immun* **71**, 1505–1512.
- Nummelin, H., Merckel, M. C., Leo, J. C., Lankinen, H., Skurnik, M. & Goldman, A. (2004). The *Yersinia* adhesin YadA collagen-binding domain structure is a novel left-handed parallel β -roll. *EMBO J* **23**, 701–711.
- Otto, B. R., van Dooren, S. J. M., Nuijens, J. H., Luirink, J. & Oudega, B. (1998). Characterization of a hemoglobin protease secreted by the pathogenic *Escherichia coli* strain EB1. *J Exp Med* **188**, 1091–1103.
- Parham, N. J., Srinivasan, U., Desvaux, M., Foxman, B., Marrs, C. F. & Henderson, I. R. (2004). PicU, a second serine protease autotransporter of uropathogenic *Escherichia coli*. *FEMS Microbiol Lett* **230**, 73–83.
- Patel, S. K., Dotson, J., Allen, K. P. & Fleckenstein, J. M. (2004). Identification and molecular characterization of EatA, an autotransporter protein of enterotoxigenic *Escherichia coli*. *Infect Immun* **72**, 1786–1794.
- Paton, A. W., Srimanote, P., Woodrow, M. C. & Paton, J. C. (2001). Characterization of Saa, a novel autoagglutinating adhesin produced by locus of enterocyte effacement-negative Shiga-toxigenic *Escherichia coli* strains that are virulent for humans. *Infect Immun* **69**, 6999–7009.
- Perna, N. T., Plunkett, G., III, Burland, V., Mau, B., Glasner, J. D., Rose, D. J., Mayhew, G. F., Evans, P. S., Gregor, J. & other authors (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* **409**, 529–533.
- Provence, D. L. & Curtiss, R. (1994). Isolation and characterization of a gene involved in hemagglutination by an avian pathogenic *Escherichia coli* strain. *Infect Immun* **62**, 1369–1380.
- Relman, D. A., Domenighini, M., Tuomanen, E., Rappuoli, R. & Falkow, S. (1989). Filamentous hemagglutinin of *Bordetella pertussis*: nucleotide sequence and crucial role in adherence. *Proc Natl Acad Sci U S A* **86**, 2637–2641.
- Reyrat, J. M., Pelicic, V., Papini, E., Montecucco, C., Rappuoli, R. & Telford, J. L. (1999). Towards deciphering the *Helicobacter pylori* cytotoxin. *Mol Microbiol* **34**, 197–204.
- Riess, T., Andersson, S. G., Lupas, A., Schaller, M., Schäfer, A., Kyme, P., Martin, J., Wälzlein, J. H., Ehehalt, U. & other authors (2004). *Bartonella* adhesin a mediates a proangiogenic host cell response. *Exp Med* **200**, 1267–1278.
- Roggenkamp, A., Ackermann, N., Jacobi, C. A., Truelzsch, K., Hoffmann, H. & Heesemann, H. (2003). Molecular analysis of transport and oligomerization of the *Yersinia enterocolitica* adhesin YadA. *J Bacteriol* **185**, 3735–3744.
- Roux, A., Beloin, C. & Ghigo, J. M. (2005). Combined inactivation and expression strategy to study gene function under physiological conditions: application to identification of new *Escherichia coli* adhesins. *J Bacteriol* **187**, 1001–1013.
- Ruiz-Perez, F., Henderson, I. R., Leyton, D. L., Rossiter, A. E., Zhang, Y. H. & Nataro, J. P. (2009). Roles of periplasmic chaperone proteins in the biogenesis of serine protease autotransporters of *Enterobacteriaceae*. *J Bacteriol* **191**, 6571–6583.
- Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406–425.
- Salzberg, S. L. & Yorke, J. A. (2005). Beware of mis-assembled genomes. *Bioinformatics* **21**, 4320–4321.
- Sandt, C. H. & Hill, C. W. (2000). Four different genes responsible for nonimmune immunoglobulin-binding activities within a single strain of *Escherichia coli*. *Infect Immun* **68**, 2205–2214.
- Sandt, C. H. & Hill, C. W. (2001). Nonimmune binding of human immunoglobulin A (IgA) and IgG Fc by distinct sequence segments of the EibF cell surface protein of *Escherichia coli*. *Infect Immun* **69**, 7293–7303.
- Sandt, C. H., Wang, Y. D., Wilson, R. A. & Hill, C. W. (1997). *Escherichia coli* strains with nonimmune immunoglobulin-binding activity. *Infect Immun* **65**, 4572–4579.
- Sauri, A., Soprov, Z., Wickstrom, D., de Gier, J.-W., Van der Schors, R. C., Smit, A. B., Jong, W. S. P. & Luirink, J. (2009). The Bam (Omp85) complex is involved in secretion of the autotransporter haemoglobin protease. *Microbiology* **155**, 3982–3991.
- Sherlock, O., Schembri, M. A., Reisner, A. & Klemm, P. (2004). Novel roles for the AIDA adhesin from diarrheagenic *Escherichia coli*: cell aggregation and biofilm formation. *J Bacteriol* **186**, 8058–8065.
- Sherlock, O., Vejborg, R. M. & Klemm, P. (2005). The TibA adhesin/invasin from enterotoxigenic *Escherichia coli* is self recognizing and

induces bacterial aggregation and biofilm formation. *Infect Immun* 73, 1954–1963.

Stein, M., Kenny, B., Stein, M. A. & Finlay, B. B. (1996). Characterization of EspC, a 110-kilodalton protein secreted by enteropathogenic *Escherichia coli* which is homologous to members of the immunoglobulin A protease-like family of secreted proteins. *J Bacteriol* 178, 6546–6554.

Surana, N. K., Cutter, D., Barenkamp, S. J. & St Geme, J. W. (2004). The *Haemophilus influenzae* Hia autotransporter contains an unusually short trimeric translocator domain. *J Biol Chem* 279, 14679–14685.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 25, 4876–4882.

Torres, A. G., Perna, N. T., Burland, V., Ruknudin, A., Blattner, F. R. & Kaper, J. B. (2002). Characterization of Cah, a calcium-binding and heat-extractable autotransporter protein of enterohaemorrhagic *Escherichia coli*. *Mol Microbiol* 45, 951–966.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C. & other authors (2009). Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* 5, e1000344.

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

Ulett, G. C., Mabbett, A. N., Fung, K. C., Webb, R. I. & Schembri, M. A. (2007). The role of F9 fimbriae of uropathogenic *Escherichia coli* in biofilm formation. *Microbiology* 153, 2321–2331.

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

Wagner, J. K., Heindl, J. E., Gray, A. N., Jain, S. & Goldberg, M. B. (2009). Contribution of the periplasmic chaperone Skp to efficient presentation of the autotransporter IcsA on the surface of *Shigella flexneri*. *J Bacteriol* 191, 815–821.

Welch, R. A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E. L., Liou, S. R., Boutin, A. & other authors (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A* 99, 17020–17024.

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

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

Yen, M. R., Peabody, C. R., Partovi, S. M., Zhai, Y. F., Tseng, Y. H. & Saier, M. H. (2002). Protein-translocating outer membrane porins of Gram-negative bacteria. *Biochim Biophys Acta* 1562, 6–31.

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

Edited by: D. W. Ussery