Complete Genome Sequence and Comparative Metabolic Profiling of the Prototypical Enteroaggregative *Escherichia coli* Strain 042


**Abstract**

*Background:* *Escherichia coli* can experience a multifaceted life, in some cases acting as a commensal while in other cases causing intestinal and/or extraintestinal disease. Several studies suggest enteroaggregative *E. coli* are the predominant cause of *E. coli*-mediated diarrhea in the developed world and are second only to *Campylobacter* sp. as a cause of bacterial-mediated diarrhea. Furthermore, enteroaggregative *E. coli* are a predominant cause of persistent diarrhea in the developing world where infection has been associated with malnourishment and growth retardation.

*Methods:* In this study we determined the complete genomic sequence of *E. coli* 042, the prototypical member of the enteroaggregative *E. coli*, which has been shown to cause disease in volunteer studies. We performed genomic and phylogenetic comparisons with other *E. coli* strains revealing previously uncharacterised virulence factors including a variety of secreted proteins and a capsular polysaccharide biosynthetic locus. In addition, by using BiologTM Phenotype Microarrays we have provided a full metabolic profiling of *E. coli* 042 and the non-pathogenic lab strain *E. coli* K-12. We have highlighted the genetic basis for many of the metabolic differences between *E. coli* 042 and *E. coli* K-12.

*Conclusion:* This study provides a genetic context for the vast amount of experimental and epidemiological data published thus far and provides a template for future diagnostic and intervention strategies.

**Introduction**

The predominant facultative anaerobe resident in the human colon is the Gram-negative motile bacillus *Escherichia coli* [1]. *E. coli* colonises the infant gut within hours of birth. However, *E. coli* has a dichotomous existence; while the majority of *E. coli* strains exist within the mammalian intestinal tract as harmless commensals, paradoxically several evolutionary lineages have deviated from this harmless lifestyle to become pathogens. Current dogma suggests that such latter strains of *E. coli* have acquired additional genetic elements, encoding specific virulence factors, which enable the organism to cause disease when infecting an otherwise healthy individual. The resulting clinical syndromes include extraintestinal infections, such as urinary tract infections, septicemia and meningitis, and intestinal infections mediating diarrhea. Those strains causing intestinal infections can be divided into six separate and major categories or pathotypes viz. enteroaggregative *E. coli* (EAEC), enteroinvasive (EIEC), enteropathogenic *E. coli* (EPEC), enterotoxigenic *E. coli* (ETEC), enterohemorrhagic *E. coli* (EHEC) and diffuse adhering *E. coli* (DAEC) [2,3]. The pathotype to which a particular strain belongs is defined by the clinical manifestation of disease, the repertoire of virulence factors, epidemiology and phylogenetic profiles [4].

EPEC were recognised as pathogens almost half a century ago [2]. However, it was not until much later that EPEC, EAEC and DAEC were distinguished from each other on the basis of their patterns of
adherence to HEp-2 cells. Unlike the localised “microcolony-forming” pattern of adherence associated with EPEC, or the diffuse adherence pattern associated with EAEC, EAEC display a characteristic aggregative or “stacked-brick” pattern of adherence associated with DAEC, EAEC display a forming” pattern of adherence associated with EPEC, or the diffuse adherence to HEp-2 cells. Unlike the localised “microcolony-discovery of EAEC as a category of pathogenic case control study of children in Chile [5]. Immediately following the discovery of EAEC as a category of pathogenic Campylobacter jejuni two large prospective surveillance studies in the UK and USA associated with (i) endemic diarrhea in infants in developing and industrialised nations. Thus, EAEC have been significantly associated with (i) endemic diarrhea in infants in developing and industrialised nations, particularly persistent diarrhea, (ii) persistent diarrhea in HIV-positive patients, (iii) traveller’s diarrhea, (iv) food/water-borne outbreaks and (v) sporadic cases of diarrhea [6]. Indeed, two large prospective surveillance studies in the UK and USA identified EAEC amongst the most commonly isolated bacterial species from individuals with diarrhea, with isolation rates similar to Campylobacter jejuni and greater than Salmonella sp. [8,9]. Furthermore, a meta-analysis of the previously published case control studies from a variety of different geographical regions clearly supports a role for EAEC in mediating diarrheal disease, and the increasing number of reports in which EAEC is implicated as the agent mediating diarrhea suggest that this is an important emerging pathogen [10].

The clinical features of EAEC-mediated diarrheal infection are normally a watery-mucoid stool, which is only occasionally bloody, with low-grade fever and little or no vomiting [4]. In patients with an active infection, EAEC elicits intestinal inflammation as determined by the presence of proinflammatory cytokines and fecal lactoferrin [11–14]. Studies using human intestinal tissue grown in vitro have indicated that EAEC has the ability to adhere to both ileal and colonic mucosa and that it can cause mucosal toxicity defined by crypt dilation, histopathology has not been observed in natural active human infections [15]. However, EAEC do form an aggregative biofilm embedded in a thick mucous blanket and it is this biofilm that may be related to its capacity to cause disease [6,12].

As not all strains of EAEC elicited diarrhea, the EAEC strain 042 which caused diarrhea in the volunteer study became the prototypical EAEC strain for the study of virulence factors and EAEC pathogenicity [7]. The first virulence factors to be identified were the aggregative adherence fimbriae or AAFs, followed by a wide array of functionally distinct factors including the Aab invasion [11]; a secreted protein called dispersin, which binds LPS thereby neutralising the negative charge of the cell [16]; the plasmid-encoded Pet and EAST-1 toxins [17,18]; the chromosomally encoded ShET1 toxin [19]; Pic, a mucinase widely associated with pathogenic E. coli and Shigella sp. [19], and more recently a novel type VI secretion system [10]; the genes conserved in all E. coli sequence diverged among the sequenced E. coli genomes, these islands are designated regions of difference (ROD) (Fig. 1; Table S1). The overall size of these RODs is 1.26 Mb (24% of the chromosome). The RODs encode virulence determinants, metabolic proteins, proteins with no obvious functions and mobile elements such as prophage and a conjugative transposon. The conjugative transposon Tn21 (within ROD66) is highly similar to Tn21 and carries a variety of genes encoding antibiotic resistance (Fig. 2). The functional significance of these genes is discussed below. Nine prophage regions, designated 042p1–042p9, were identified in the EAEC 042e genome (Table 2). Four of the prophage were lambdoid in nature (042p2, 042p3, 042p4 and 042p6) and were highly similar to each other however only three (042p1, 042p3 and 042p6) appeared to carry cargo genes (see Fig. S1 and File S1). The content of the remaining ROD are discussed in detail later. On the basis of nucleotide sequence homology, the plasmid pAA belongs to the IncFIA family. The plasmid includes 152 CDS, of which 32 are pseudogenes. Of the remainder, there are 7 that encode hypothetical proteins with no match in the database, 23 encode conserved hypothetical proteins with no predicted function, 55 have transfer, replication or plasmid maintenance functions, there are 18 mobile element-derived genes that encode transposons, and the remaining 17 CDS have demonstrated or predicted roles in virulence (Table 1 and Fig. S2). Insertions in the plasmid include genes encoding many of the well-characterised EAEC 042 virulence factors and include the cytopathic toxin Pet, the AAF/II aggregative fimbiae, the AggR transcriptional regulator, dispersin and its cognate secretion machinery Aat, and operons encoding a putative iron transport system and a polysaccharide biosynthesis pathway all of which are discussed later.

E. coli core genome and pangenome. The EAEC 042 genome is largely colinear with that of the previously sequenced E. coli genomes except for a few inversions and insertions/deletions (Fig. S3). A box-plot showing the estimated core genome size (i.e. the genes conserved in all E. coli strains), as a function of the number of genomes sequenced for 100 randomly selected strain combinations is shown in Fig. S4. An exponential decay curve was

<table>
<thead>
<tr>
<th>Table 1. Major features of the E. coli 042e genome.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Chromosome</strong></td>
</tr>
<tr>
<td>Size (bp)</td>
</tr>
<tr>
<td>Predicted CDSs</td>
</tr>
<tr>
<td>G+C content (%)</td>
</tr>
<tr>
<td>Coding regions (%)</td>
</tr>
<tr>
<td>Average CDS length (bp)</td>
</tr>
<tr>
<td>tRNAs</td>
</tr>
<tr>
<td>rRNA</td>
</tr>
<tr>
<td>Pseudogenes</td>
</tr>
</tbody>
</table>

[10:13771/journal.pone.0008801]
fit using the R function nlrq [21], and gave a predicted core genome size of 2356 genes (Table S2). This is larger than the previous estimate of 2200 [22,23], possibly due to our inclusion of genes that are present but unannotated in some strains. The predicted core genome size is close to the number of genes conserved across all the genomes included in this study, suggesting that the number of possible gene deletions is close to saturation, and that further E. coli genome sequencing projects are unlikely to identify many novel gene deletions. The analysis indicates an open E. coli pangenome, as has been found in previous studies [22,24], with an estimated 360 new genes being identified with each additional genome sequenced (Fig. S5). The E. coli core genome was further compared with the non-coli Escherichia albertii and Escherichia fergusonii, with 2173 genes found to be conserved (Table S2). Comparisons with the other available intact enterobacterial genomes showed that 967 genes were conserved across the family (Table S2).

E. coli phylogeny. A phylogeny was constructed based on the concatenated sequences of 2173 genes that are conserved in all E. coli strains and in E. albertii and E. fergusonii, which were included as outgroup sequences. The results are shown in Fig. S6. The established E. coli sub-groups (A, B1, B2, D and E) are all

Figure 1. Circular representation of the E. coli O42 chromosome. From the outside in, the outer circle 1 marks the position of regions of difference (mentioned in the text) including prophage (light pink) fimbrial operons (Dark green) as well as regions differentially present in other E. coli strains: blue (Present in 0157:H7 & absent/divergent in UPEC CFT073); Light Green (Present in 0157:H7 absent/divergent in UPEC CFT073). Circle 2 shows the position in bps. Circles 3 and 4 show the position of CDSs transcribed in a clockwise and anticlockwise direction, respectively (for colour codes see below); circle 4 to 13 show the position of E. coli O42 genes which have orthologues (by reciprocal FASTA analysis) in other E. coli strains (see methods): Sakai (0157:H7; red), UT189 (UPEC; dark blue), CFT073 (UPEC; light blue), 536 (UPEC; orange), APEC 01 (APEC; dark pink), E2348/69 (EPEC; black), H10407 (ETEC; salmon pink), E24377A (ETEC; pale pink), HS (grey), and K-12 MG1655 (green). Circle 14 sows the position of genes unique to E. coli O42 unique (red). Circle 15 shows a plot of G+C content (in a 10 Kb window). Circle 16 shows a plot of GC skew ([G−C]/[G+C]; in a 10 Kb window). Genes in circles 3 and 4 are colour coded according to the function of their gene products: dark green = membrane or surface structures, yellow = central or intermediary metabolism, cyan = degradation of macromolecules, red = information transfer/cell division, cerise = degradation of small molecules, pale blue = regulators, Salmon pink = pathogenicity or adaptation, black = energy metabolism, orange = conserved hypothetical, pale green = unknown, brown = pseudogenes. doi:10.1371/journal.pone.0008801.g001
monophyletic with the exception of group D, which is divided by the root. E. coli strains SECEC SMS-3-5 and IAI39 cluster with group B2, which includes many extraintestinal pathogenic E. coli strains, whereas strains EAEC 042 and UMN026 cluster with groups A, B1, E and the Shigella strains. This corresponds with the conclusions drawn in a recent MLST study, where it was proposed to classify strains such as SMS-3-5 and IAI38 in a new group F [25]. However, we prefer to designate the two groups as D1 and D2, to retain compatibility with the previous nomenclature and to follow the precedent of group B.

**Metabolic Profiling**

E. coli K-12 strains, such as E. coli MG1655, have been used to characterise many of metabolic pathways we understand today. However, recent publications have described what most E. coli biologists have known for some time; due to prolonged laboratory passage and a variety of treatments to remove λ-phage and the F plasmid, E. coli K-12 strains are not archetypal strains representing the biology of the genus [26]. The genotype of the E. coli K-12 strain MG1655 (F-, λ-, ilvG, rfb-50, rph-1) given by the E. coli stock Center, reflects only some of the differences between E. coli MG1655 and other E. coli strains. These differences extend beyond the additional virulence factors carried by pathogenic strains and include central metabolic functions carried by other E. coli strains but lost by E. coli K-12 strains [26]. To reveal a more representative metabolic profile for E. coli strains BioLog Phenotype Microarrays (PMs) were performed on EAEC 042 and compared with similar analyses of E. coli MG1655 (Table S3 and S4). The genetic basis accounting for significant differences between the strains are described. The major differences between the strains can be summarized into two main categories: resistance to antimicrobials and differences in nutrient utilization.

**Antibiotic resistance/drug resistance.** Soon after the discovery of the EAEC pathovar it was noted that many clinical isolates of EAEC displayed multiple antibiotic resistance [27]. Antibiotic resistance among EAEC strains is typically higher than among other diarrheagenic pathovars, perhaps accounting for the increasing isolation of EAEC from epidemiologic studies [28]. A variety of studies from geographically distinct areas have reported high

<table>
<thead>
<tr>
<th>ROD</th>
<th>Position</th>
<th>Size</th>
<th>Site of insertion</th>
<th>Cargo genes</th>
<th>Position of cargo genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROD15</td>
<td>877116..917926</td>
<td>40811</td>
<td>15bp direct repeat</td>
<td>gtnA8 and Eo42-0853 encoding bactoprenol glucosyl transferase involved in O-antigen modification</td>
<td>914835..917583</td>
</tr>
<tr>
<td>ROD18</td>
<td>1396781..1450149</td>
<td>53369</td>
<td>Imperfect repeat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROD21</td>
<td>1533876..1584276</td>
<td>50401</td>
<td>15bp direct repeat</td>
<td>sitA8BCD encoding iron transport proteins</td>
<td>1580324..1585773</td>
</tr>
<tr>
<td>ROD26</td>
<td>1763886..1811200</td>
<td>47315</td>
<td>Palindromic repeat</td>
<td>Eo42-1724 encoding putative exonuclease</td>
<td>1807070..1809541</td>
</tr>
<tr>
<td>ROD27</td>
<td>1832521..1843545</td>
<td>11025</td>
<td>Intragenic in Eo42-1748</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROD30</td>
<td>2214943..2256641</td>
<td>41699</td>
<td>tRNA-Arg: 31-mer direct repeat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROD30</td>
<td>2256892..2273265</td>
<td>15494</td>
<td>tRNA-Met</td>
<td>Eo42-2203 encoding putative exonuclease VIII</td>
<td>2268605..2271046</td>
</tr>
<tr>
<td>ROD34</td>
<td>2551911..2560079</td>
<td>8169</td>
<td>tRNA-Pro</td>
<td>Eo42-2429, similar to proQ, structural element that influences osmotic activation of proP at posttranslational level; Eo42-2430, PerC-like protein, similar to transcriptional activator of LEE in EPEC/EHEC</td>
<td>2557981..2558798</td>
</tr>
<tr>
<td>ROD60</td>
<td>4198051..4218142</td>
<td>20092</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROD61</td>
<td>4236703..4247732</td>
<td>11030</td>
<td>tRNA SelC(p)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

doi:10.1371/journal.pone.0008801.t002
levels of resistance to tetracycline, spectinomycin, streptomycin, trimethoprim-sulfamethoxazole and ampicillin [29–31]. The antibiotic resistance profile of *E. coli* 042 derived from PMs revealed resistance to sulphonamides, chloramphenicol, aminoglycosides and tetracyclines that was not exhibited by *E. coli* MG1655 (Table S3). This is consistent with the presence on the *E. coli* 042 chromosome of *Tn2411*, a *Tn21*-like transposon (Fig. 2). *Tn2411* possesses genes encoding resistance to chloramphenicol (*cat*) and tetracycline (*tetA*) and also includes a class 1 integron In2 that carries antibiotic resistance cassettes *aadA1* (streptomycin and spectinomycin), *sul* (sulfonamide) and *emrE* (ethidium bromide) (Fig. 2). Interestingly, the PM data revealed there was no difference between the ability of *E. coli* MG1655 and *E. coli* 042 to grow in the presence of ethidium bromide even though *E. coli* MG1655 does not possess the *Tn2411* element possessing *emrE*. This can be explained by the fact that *E. coli* MG1655 possesses both the *emrE* (phage associated) and the predicted multidrug efflux system *emrTK* on the chromosome, genes that are absent in the equivalent sections of the *E. coli* 042 genome. In addition, the *Tn2411* element also possesses genes for mercury resistance (*merRTPCAD*). However, the PMs do not include an assay for growth in mercuric chloride. Nevertheless, the high identity between the *mer* genes in the *Tn2411* element on the *E. coli* 042 chromosome and *Tn21* strongly suggests that the *E. coli* 042 *mer* resistance is functional [32].

The PMs revealed *E. coli* 042 is more resistant to arsenic and antimyco chloride than *E. coli* MG1655. This phenotype is most likely the result of the *E. coli* MG1655 *ars* operon lacking *arsD* (coding for the catalytic subunit of the ATP-driven arsenite/antimonite pump) or *arsD* (the trans-acting transcriptional repressor protein) genes (Fig. S7). Previous work has shown that in the absence of the ArsA ATPase subunit, ArsB confers only partial arsenite resistance by translocating these ions into the periplasm using energy derived either from the proton pumping respiratory chain or from F$_{0}$F$_{1}$ ATPase [33].

The PMs revealed that *E. coli* MG1655 is more resistant to acriflavine than *E. coli* 042. However, both *E. coli* 042 and *E. coli* MG1655 possess *arsAB* (*Ec042-0500/0501) and *tolC* (*Ec042-3326*) the gene products of which act in concert to form an efflux system that confers acriflavine resistance [34,35]. To determine whether *E. coli* MG1655 was more efficient than *E. coli* 042 at effluxing compounds such as acriflavine, the accumulation of Hoescht 33342 was determined. The accumulation of Hoescht 33342 reached a significantly higher steady state within *E. coli* 042 than *E. coli* MG1655 (Fig. 3). The addition of the efflux pump inhibitor PA$\text{N}$ increased accumulation of the dye by both *E. coli* MG1655 and *E. coli* 042 although the effect was much greater for the latter (Fig. 3). This indicates that the greater amount of this compound accumulated by *E. coli* 042 relative to *E. coli* MG1655 is likely to be a result of increased permeability of *E. coli* 042 rather than lack of efflux activity. Similar results were obtained with the use of carbonyl cyanide-m-chlorophenyl hydrazone (CCCP) a proton motive force inhibitor which inhibits the action of other efflux systems (Fig. 3). Nevertheless, the increase in dye accumulation in the presence of PA$\text{N}$ and CCCP demonstrates that active efflux systems are present in *E. coli* 042. The higher resistance of *E. coli* MG1655 to acriflavine is thus most likely due to decreased uptake and is perhaps unsurprising as the related dye acridine orange was used to select K-12 derivatives lacking the F plasmid [26].

To validate the genome sequence and PM data the minimum inhibitory concentration (MIC) for acriflavine and a panel of other antimicrobials was determined (Table 3). The PM data was predictive of MIC although this correlation was not absolute. *E. coli* 042 was significantly more resistant to chloramphenicol, tetracycline, streptomycin and spectinomycin than *E. coli* MG1655, most likely as a result of the carriage of the specific resistance genes. There were no other significant (i.e. two dilutions or greater) differences in susceptibility to any of the other antimicrobials tested between *E. coli* 042 and *E. coli* MG1655.

Interestingly, susceptibility testing revealed *E. coli* 042 was fourfold more susceptible than *E. coli* MG1655 to nalidixic acid. This correlated with the PM data which indicated a reduced ability of *E. coli* 042 to grow in the presence of nalidixic acid compared to...
Iron acquisition. Iron is an essential nutrient for bacterial growth and is a major limitation to successful colonisation within the mammalian host. Mammals have high-affinity iron-binding proteins such as transferrin and lactoferrin that ensure free iron availability is maintained at extremely low levels. Counteracting these protective measures, pathogens have evolved a variety of high-affinity iron scavenging systems such as siderophore production, haem and haemoglobin uptake transporters [35]. We have reviewed affinity iron scavenging systems such as siderophore production, increased uptake of the antibiotics, rifampicin, macrolides, and some other antimicrobials is more resistant than EAEC 042 to a variety of E. coli antibiotics, as described above, or may reflect the increased ability (Table S4). This could be due to increased uptake of the antibiotics, MG1655, and EAEC 042 and these are summarised in Table S5. Previous reports have shown that E. coli MG1655 grows and differs between strains [44,45]. The PM data shows that EAEC 042 possesses homologues of the S. Typhimurium \
doxKFPQ gene products [40]. EAEC 042 possesses 13.7 pmoles iron/mg protein compared to 8.15 ± 0.976 pmoles iron/mg protein in E. coli MG1655. Thus, E. coli 042 contains 1.68-fold more iron than E. coli MG1655; a finding consistent with other pathogenic strains of E. coli [M. Goldberg, unpublished].

Recent studies by Goldberg and Lund (personal communication) have shown that the enhanced ability of pathogenic strains to scavenge iron can prove disadvantageous under oxidising conditions. Oxidative stress damages iron-binding proteins which results in an increase in free iron levels in the cytoplasm, triggering Fenton reactions [38,39]. The enhanced free-radical production resulting from this reaction can overwhelm the cells’ free-radical scavenging systems, causing further damage to iron binding proteins including the Fur repressor, leading to increased iron uptake and even greater oxidative damage. Cells containing larger amounts of intracellular iron are therefore more likely to succumb to oxidative stress than cells containing less. We found that E. coli 042 was more susceptible to the readox-cycling compound menadione (MIC = 1 mg/ml) than E. coli MG1655 (MIC = 2 mg/ml) indicating EAEC 042 is more susceptible to oxidative stress than E. coli MG1655. The PM data indicated that E. coli 042 is more susceptible to gyrase inhibitors (e.g. Nalidixic acid) and β-lactam antibiotics (e.g. Oxacillin, Penicillin) than E. coli MG1655 (Table S4 and S5). The ability of EAEC 042 to sequester higher levels of iron, in conjunction with the increased susceptibility to the oxidising reagent menadione, strongly suggests the increased susceptibility of EAEC 042 to these antibiotics is due to the triggering of Fenton reactions in a manner previously described by Kohanski and colleagues [39].

Carbon source utilization. Bacteria require a sufficient supply of carbon to feed their metabolic pathways. In their native environments heterotrophic organisms encounter limited amounts of complex mixtures of carbon sources that are often present at low concentrations. As a result microbial cells have developed multiple different systems to utilise a wide array of different substrates as carbon sources. Such differences are utilised in diagnostic tests to differentiate between particular species and strains of bacteria. PMs for sole carbon source utilization showed that EAEC 042 can utilise 2-Deoxy-D-Ribose more effectively than E. coli MG1655 (Table S5). Previous reports have shown that E. coli K-12 cannot use this carbon source, but Salmonella enterica serovar Typhimurium can, using the doxKFPQ gene products [40]. EAEC 042 possesses homologues of the S. Typhimurium doxKFPQ( Ec042:4753–4756).

N-acetyl-D-galactosamine and N-acetyl-D-glucosamine are components of intestinal mucus, as well as peptidoglycan. The PM screening demonstrated that EAEC 042 utilises N-acetyl-D-galactosamine and N-acetyl-D-glucosamine better than E. coli MG1655 (Table S3), confirming previous work showing that K-12 strains are unable to use these substrates as sole carbon sources [41]. In E. coli O157:H7 the genes for N-acetyl-D-galactosamine and N-acetyl-D-glucosamine utilisation were identified as the agaZWE-FASIBCD gene cluster. In E. coli MG1655 a portion of this locus is missing due to site-specific recombination between agaW and agaA [42], however this locus is present and intact in EAEC 042 (Fig. S8). This observation is consistent with the recent demonstration that EAEC 042 can utilise intestinal mucus as a carbon source [43].

L-sorbose utilization by pathogenic E. coli and Shigella differs between strains [44,45]. The PM data shows that EAEC 042 is significantly better at utilising L-sorbose than E. coli MG1655 (Table S3). Genome comparison between the two strains (Fig. S9) shows that EAEC 042, like other E. coli and Shigella pathotypes carries the sorEMABFDC operon (Ec042:4384–4390), located between ybdC and yuhE. E. coli MG1655 does not carry this operon. BLAST analysis of the CDS in this region in EAEC 042 confirmed that the genes have high identity to previously described sor genes and thus are presumed to be functional.

Conversely, PM data showed that E. coli MG1655 grows and metabolizes D-Serine, Mucic acid (D-galactarate), β-D-Allose and D-Xylose more effectively than EAEC 042 (Table S4). A previous report has highlighted extensive genomic variability in the argW-
**Host-Pathogen Interactions**

Pathogenic bacteria produce a wide array of virulence factors which allow the organism to colonise a specific niche within the host and to mediate disease. The majority of these factors are proteins. These proteins are either secreted factors, which interact directly with the host, or form part of the machinery required for translocation of the secreted molecules to the exterior of the bacterial cell [58]. To ascertain if specific virulence factors are associated with the clinical characteristics of EAEC-mediated diarrhea, the repertoire of genes from three EAEC strains were compared with the commensal E. coli strain HS (Fig. 5). E. coli HS was chosen as a base for comparison rather than the lab adapted E. coli strain K-12; current dogma suggests genes responsible for pathogenesis are absent from commensal isolates. These analyses revealed 3497 CDS (72.7%) in EAEC 042 that are common to all four sequenced E. coli genomes, and 210 (4.4%) EAEC-specific
genes (Table S6). However, detailed analyses of the EAEC-specific genes revealed that 120 were mobile elements, 53 were present in the lab strain E. coli K-12, 18 were involved in O-antigen/colonial acid biosynthesis and 6 were pseudogenes. Only 13 CDS represented genes which might be labelled virulence factors viz. the putative polysaccharide biosynthetic shy locus (Ec042-4770–4772), the iron recruitment for locus (Ec042-4776–4782) and three genes of a previously described Type VI secretion locus (Ec042-4562, Ec042-4563 and Ec042-4571); these are discussed in detail later. The paucity of conserved virulence factors reflects the previously described phylogenetic heterogeneity of EAEC [59]. The 551 EAEC 042-specific genes represent many of the previously described virulence factors discussed below.

Protein Transport

EAEC 042 possesses the major inner membrane translocation machines viz. Tat (Ec042-4216–18) and Sec, for export of proteins to the periplasm. It is not surprising to find within EAEC 042 a complete repertoire of E. coli Sec components as many are essential for survival (e.g. SecY) and the non-essential components (e.g. SecB) are required for efficient functioning of the cell. In contrast, not all bacteria possess the Tat system and none of the Tat system components are essential for the survival of E. coli [58,60]. Screening of all EAEC 042 CDS using TatP and TatFind identified a number of substrate molecules present within EAEC 042. However, none were unique to EAEC 042, and all had homologues within the non-pathogenic laboratory strain E. coli K-12, suggesting EAEC 042 does not possess any virulence factors which require the Tat system for secretion.

All members of the major protein secretion systems (Type 1–6 and the Chaperone-Usher pathways) required for translocation of proteins to the exterior of the cell are represented in EAEC 042 [58]. The Type 4 secretion system encodes the plasmid conjugation system (Fig S2) and will not be discussed further.

**Chaperone-usher/fimbriae.** E. coli 042 possesses 11 chaperone-usher family fimbrial operons on the chromosome and an additional system on the plasmid (Fig S13). The number of these operons in EAEC 042 is similar to those of other E. coli strains. Although the chromosomally-encoded fimbrial operons are found in at least one of the previously sequenced E. coli strains, they display significant sequence divergence, particularly between the fimbrial adhesins and fimbrial structural subunits, presumably due to immunological pressures. No definitive explanation exists for such a large repertoire of fimbrial adhesins however it has been suggested for Salmonella, which possesses a similar diversity in fimbrial systems, that the differing systems reflect adaptation to colonizing different hosts; an alternative explanation may be that they are required for colonising different sites within the same host [61].

The AAF/II plasmid encoded fimbrial system is one of the better characterized virulence factors of EAEC 042. The fimbriae belong to the Dr family of fimbrial adhesins. Expression of the fimbrial genes is regulated by the AraC class of transcriptional activator, which also regulates a variety of chromosomally encoded genes. The fimbriae are essential for biofilm formation and mediating which also regulates a variety of chromosomally encoded genes.

**Type 1 secretion systems.** In addition to the secreted substrate protein, normal Type 1 protein secretion systems (T1SSs) of Gram-negative bacteria are composed of three basic subunits viz. a ToIC-like outer membrane pore-forming protein (OMP) and two inner membrane associated proteins, respectively termed the membrane-fusion protein (MFP), which contacts with the ToIC-like protein, and the ATP-binding cassette (ABC) protein, which transduces energy to the system [30]. EAEC 042 possesses three T1SSs: the plasmid-encoded aat locus and two chromosomally-encoded systems (Ec042-0526–0531 and Ec042-3196–3202) [16]. The aap gene (Ec042-pAa005) encodes Dispersin, the secreted substrate protein for the plasmid encoded T1SS. In contrast to the normal T1SSs, the aat-encoded dispersin secretion system, while containing MFP (AatD; Ec042-pAa011), ABC (AatC and AatP; Ec042-pAa007 and 010) and OMP (AatA; Ec042-pAa008) proteins, also contains an additional component (AatB; Ec042-pAa009) whose function is unknown [16].

The EAEC 042-0526–0531 locus is homologous to a synaptic locus in E. coli O157:H7 and possesses genes encoding an RTX toxin, a ToLC-like OMP, an MFP, an ABC protein and a CueR-like transcriptional regulator (Fig. S14). The substrate molecule for both systems resembles an RTX family exoprotein however in both systems it is frameshifted; in EAEC 042 it is represented by CDS Ec042-0527–0530. The Ec042-3196–3202 locus encodes genes for the synthesis and secretion of the H47 ribosomally-encoded peptide microcin antibiopic [63]. While this locus encodes an MFP and an ABC protein for secretion of the peptide, in contrast to the RTX toxin it does not encode a ToLC-like OMP. Previous studies have demonstrated that such systems can commandeer the chromosomally encoded ToLC protein (Ec042-3326) and this is supported by the demonstrable promiscuity of ToLC interaction with other MFP and ABC proteins responsible for efflux of a wide variety of substrate molecules e.g. AcrAB [34].

**Type 2 secretion systems.** CDS Ec042-3242–3255 encode a complete Type 2 secretion system (T2SS). Recent investigations have demonstrated that this locus is transcribed from a promoter upstream of Ec042-3255 resulting in a polycistronic mRNA coding for all components of the T2SS [64]. Comparisons with the other sequenced E. coli genomes reveal that the location and sequences of the genes encoding this T2SS are conserved in many pathogenic strains of E. coli (Fig. S15). E. coli K-12 possesses yghJ-gspO,pppF(1),gspC,yghP and the distal gspL-gspM genes but not the remainder of the genes suggesting that this locus has undergone genetic attrition in certain strains and acquisition of the genes predates divergence of genus E. coli (Fig. S15). The locus was previously described for ETEC H10407 and is essential for secretion of heat-labile enterotoxin (LT) [65]. However, neither EAEC 042 nor many of the other strains possess LT-toxin suggesting that this T2SS has alternate substrate molecules. Analysis of Ec042-3250-3255 (yghJ) reveals homology to AcfD, a colonization factor of Vibrio cholerae that is part of the Tcp T2SS responsible for secretion of cholera toxin, a toxin homologous to E. coli LT [66]. AcfD, like YghJ, is a lipoprotein suggested to be involved in the normal functioning of the T2SS. However, the lipoprotein sorting signal is similar to the secreted T2SS substrate molecule PulA from Klebsiella oxytoca [67] suggesting that YghJ is in fact a secreted extracellular lipoprotein and the substrate of this T2SS, though of unknown function.

**Type 3 secretion systems.** Certain E. coli rely on the T3SS to produce flagellar systems and to inject effector proteins into host cells [68]. EAEC 042 possesses the genes for an intact flagellar system encoding the H18 serotype. The EAEC 042 genome also contains an additional flagellar gene cluster, Flag-2, that potentially encodes a second flagellar system, similar to lateral flagellar systems in other bacteria. However, a frame shift disruption in an essential flagellar gene lfgC has probably rendered this system non-functional in EAEC 042 [69].

As reported previously, the EAEC 042 genome contains two non-flagellar T3SS gene clusters, ETT2 (EC042-3044–3075) and ETT3 (EC042-3172–3180). The latter contains an additional component (AatB; Ec042-pAa009) whose function is unknown [16].

The EAEC 042 genome contains two non-flagellar T3SS gene clusters, ETT2 (EC042-3044–3075) and ETT3 (EC042-3172–3180). The latter contains an additional component (AatB; Ec042-pAa009) whose function is unknown [16].
to the absence of obvious ablating mutations [70,71]. We have identified several CDS that encode known and potential effectors that could be secreted via these systems (Table S7). Most of these are orthologues of EHEC and ETEC genes but few are unique to EAEC 042 relative to all the analysed E. coli genomes (Table S7).

The gene product of EC042-3075, encoded at one end of the main ETT2 gene cluster, shows homology to glucoamylases and has been tentatively identified as a potential ETT2 effector [72]. Searches for T3SS effector genes outside the ghU and sidC islands failed to identify any of the phage-encoded effectors that predominate in EHEC. However, it has been suggested that non-phage-encoded effector genes in E. coli encode currently or formerly active ETT2 effectors. Consistent with this view, we found that 15 of the 19 non-phage-encoded effector genes from the EHEC O157 genomes have positional orthologs in the EAEC 042 genome (Table S7). Plus the EAEC 042 genome houses one additional non-phage-encoded effector gene: Ec042-1240, which encodes leucine-rich repeats and is a relative of the ipaH gene family from Shigella spp. At least four previously undescribed putative effector genes disrupted by frame shifts in EHEC are apparently intact in the EAEC 042 genome (Ec042-1600, Ec042-4064, Ec042-4074, Ec042-4075), adding weight to the idea that the ETT2 T3SS is still active or was recently so. Experiments to demonstrate roles for these novel effectors are under way.

The T3SS of EPEC has been demonstrated to interact with intimin in a Tir dependent manner to promote intimate attachment to the host cell. While no Tir like molecules could be identified in EAEC 042, several intimin-like proteins are present viz. Ec042-0533, Ec042-2220 and three outer membrane proteins, Ec042-2711–2713, similar to proteins on CS54 island of S. Typhimurium. The mechanism by which these molecules are secreted remains enigmatic, however it is clear they do not exit via the T3SS and recent publications suggest a secretion mechanism analogous to the T5SS [73].

**Type 5 secretion systems.** Based on differences in the modes of biogenesis the Type 5 secretion system (T5SS) has been divided into three subclasses termed the autotransporters (AT-1; T3aSS), the two-partner secretion system (TPS; T5bSS) and the trimeric autotransporters (AT-2; T5cSS) [58,74,75]. EAEC 042 appears to possess the full complement of T5SS encoding genes including 14 AT-1 systems, one TPS system and two AT-2 systems (Fig. S16).

Members of the classical AT-1 autotransporter family possess a conserved architecture of five domains [58]. The two major domains are the functional secreted passenger, or α, domain, which may be released into the extracellular milieu or remain attached to the cell surface, and the β-domain, an integral OMP which mediates secretion of the passenger domain across the outer membrane [58]. EAEC 042 possesses 14 genes encoding AT-1 proteins with a repertoire almost identical to that of E. coli O157:H7 (Fig. S16). Two genes encode the previously described mucinase Pec (Ec042-4593) and the plasmid encoded toxin Pet (Ec042-pAA035), both of which are serine proteases secreted into the extracellular milieu; Pec cleaves mucin and confers fitness for intestinal colonization whereas Pet is a toxin discussed later in recruiting β-barrel OMPs into the OM [80]. It exists in complex with a number of accessory factors and consists of a periplasmically located repetitive domain comprising five POTRA repeat motifs and a C-terminal integral OM β-barrel domain [80,81].

Like YaT, the TPS proteins also consist of a periplasmically located POTRA domain (two POTRA repeats) and an OM β-barrel domain. However, in contrast to the YaT proteins, the TPS proteins do not form heterooligomeric complexes and while YaT is promiscuous for OMPs in general, mediating insertion of OMPs into the OM, TPS proteins are selective for their genetically linked TpsA protein and allow for translocation across the OM [58]. Phylogenetic analyses reveal that YtfM is related to both YaT and TpsB proteins but forms a distinct clade with a group of homologous proteins [82].

**PAECC Genome Sequence**

YtfM possesses three POTRA domains; the first POTRA could not be modeled against known structures (Fig. 7). Immediately, downstream of ytfM is ytfN (Ec042-4703). The genetic linkage of these genes is universally conserved in E. coli and a wide range of Gram-negative bacteria (Fig. S16). The conserved linkage of these molecules, and the identical phenotypes derived for deletion of each of these molecules suggests YtfN is a substrate molecule for YtfM analogous to the TpsA protein. Further support for the theory that YtfM/N form a TPS system is derived from the demonstration that both possess signal sequences for location to the periplasm. YtfM does not function in a manner analogous to another gene is more disease [78]. Two of these genes (Ec042-2242 and 4511) are closely related (87% identity) while the other gene is more divergent (72 and 65% identity, respectively). Interrogation of Genbank reveals that multiple alleles of agn43 can occur within a single strain and such occurrences are not limited to any one branch of the E. coli phylogeny [78]. No absolute correlation exists between the clinical disease manifested by a particular strain of E. coli and the presence or absence of a particular allele of agn43. As recently reported, the dichotomy in the grouping can be explained by a region of significant diversity, encompassing the C-terminus of α14 and the N-terminus of β15, overlapping the point of cleavage between the two domains. Thus, while Ec042-2242 and Ec042-4511 possess the empirically determined cleavage site Ec042-4803 does not. Nevertheless, no defect in the processing of the Ag43 protein derived from Ec042-4803 was detected and expression of Ec042-4803 promoted biofilm formation and cell-cell aggregation in a manner similar to expression of the other alleles (Fig 6). Both allelic groups possess promoters containing three GATC sites with similar spacing, suggesting that members of both families undergo reversible phase variation in a deoxyadenosine methyltransferase- and OxyR-dependent fashion [78].

Four of the CDS (Ec042-1219, -1258, -2590 and -1642) appear to encode non-functional AT-1 proteins while the remaining 6 CDS are variably present in pathogenic and non-pathogenic E. coli strains suggesting that these genes may be important for the ability of E. coli to colonise the gut. A reasonable explanation for the varying nature of the AT-1 genes (as intact or pseudogenes) across the breadth of the E. coli phylogeny may be functional redundancy; many of these proteins have been shown to perform the same in vitro functions i.e. cell-cell aggregation and biofilm formation, thus loss of one gene may be compensated for by the presence of another [77].

In contrast to the AT-1s discussed above, the TPS system consists of two separate proteins; an outer membrane β-barrel protein (TpsB) and a cognate substrate protein (TpsA) both containing signal sequences to mediate inner membrane transloction. Interrogation of the EAEC 042 genome reveals two genes encoding proteins bearing similarity to the TpsB proteins i.e. yaeT (Ec042-0175) and ytfM (Ec042-4702), a gene of unknown function [79,80]. YaT is an essential protein which plays an unknown role in recruiting β-barrel OMPs into the OM [80]. It exists in complex with a number of accessory factors and consists of a periplasmically located repetitive domain comprising five POTRA repeat motifs and a C-terminal integral OM β-barrel domain [80,81]. Like YaT, the TPS proteins also consist of a periplasmically located POTRA domain (two POTRA repeats) and an OM β-barrel domain. However, in contrast to the YaT proteins, the TPS proteins do not form heterooligomeric complexes and while YaT is promiscuous for OMPs in general, mediating insertion of OMPs into the OM, TPS proteins are selective for their genetically linked TpsA protein and allow for translocation across the OM [58]. Phylogenetic analyses reveal that YtfM is related to both YaT and TpsB proteins but forms a distinct clade with a group of homologous proteins [82].

Tertiary structural predictions indicate that YtfM adopts a similar structure to YaeT and TpsB proteins but forms a distinct cluster with a group of homologous proteins [82].

Tertiary structural predictions indicate that YtfM adopts a similar structure to YaeT and TpsB proteins but forms a distinct cluster with a group of homologous proteins [82].
YaeT as it is not essential for viability and the levels of OMPs present in the OM are not affected [79]. However, in contrast to the TpsA molecules, YtfN does not possess a conserved TPS motif suggesting the targeting of this molecule to YtfM is different to the TPS systems characterized to date. Furthermore, analysis of YtfN with the BetaWRAP program reveals that YtfN possesses only a partial β-helix structure, a structural organization which appears to be a feature of TpsA proteins. While the function of this locus remains elusive, this locus deserves significantly more attention as several studies using a variety of Gram-negative bacteria with mutations in either YtfM or YtfN have shown those organisms to be severely attenuated in infectious disease models suggesting use as a target for therapeutic or preventive interventions [83–85].

EAEC 042 possesses genes encoding two trimeric AT-2 proteins (Ec042-0536 and 3912). Ec042-3912 is orthologous to the recently characterized UpaG protein of UPEC CFT073 [86]. This gene transcends the E. coli phylogeny and is present in E. albertii suggesting acquisition before divergence of the E. coli species. Like many of the classical AT-1 proteins, UpaG mediates cell-cell aggregation, biofilm formation and adherence to epithelial cells. Furthermore, like other members of the AT-2 family, it binds extracellular matrix proteins [86]. The presence of UpaG in the commensal E. coli HS suggests that UpaG does not play a significant role in disease per se but may be required for efficient colonization of the gut. However, recent studies have suggested that immunization with UpaG mediates a protective effect against challenge with virulent extraintestinal isolates of E. coli in mouse models of infection [73]. The literature provides no clues to the function of Ec042-0536.

**Type 6 secretion systems.** Within the genome of 042, we found three predicted type VI secretion systems (T6SS). Two of the T6SSs identified (Ec042-4524–4555 and approximately 4 kb downstream Ec042-4562–4577) have been previously characterized with the latter locus under the control of AggR [88]. Interestingly, Ec042-4562, Ec042-4563 and Ec042-4571 are amongst the few genes conserved within EAEC strains 042, 101-1 and 55989. The third T6SS (Ec042-0210–0226) encoded within the EAEC 042 genome is inserted at the aqV tRNA locus. This T6SS is more widely distributed than the other two, and the ca. 23 kb region encoding these genes is >95% identical on the nucleotide level to genomes of E. coli urinary tract isolates UMN026, 536, and UTB9, avian isolates APEC O1 and BEN2908, the cerebro-spinal fluid isolate S88, and fecal isolates ED1a and EAEC 55989 [89–92]. There is no experimental evidence that these genes are expressed. This locus is also the only one of the three that encodes a putative Forkhead-associated domain protein (Ec042-0219), which is proposed in the T6SS of Pseudomonas aeruginosa [93] to be phosphorylated by a serine-threonine kinase in response to an environmental cue, resulting in activation of secretion. However, no kinase or phosphatase capable of modulating a phosphorylation event was found encoded within this putative T6SS.

**Toxins**

The pAA plasmid carries the genes for two toxins, the plasmid-encoded toxin Pet, and EAST-1 (Ec042-pAA040) (Fig. S2). The

---

**Figure 6. Functional characteristics of the EAEC 042 agn43 alleles.** The three Antigen 43 encoding genes (Ec042-2242, Ec042-4511 and Ec042-4803) were cloned into a high copy number vector (pCRII-TOPO) and expressed in E. coli TOP10. (A) Western immunoblot of E. coli outer membrane fractions probed with anti-agn43 antisera demonstrating all three alleles produced cross reacting species. As expected from nucleotide sequence analysis Ec042-4803 produces a smaller passenger domain than Ec042-2242, Ec042-4511 and Antigen 43 from E. coli K-12. Autoaggregation assays (B) and biofilm assays (C) of E. coli expressing each agn43 allele demonstrated that each gene product was capable of promoting autoaggregation or biofilm in a manner similar to that previously described. There was no appreciable difference between the proteins in their capacity to induce autotaggregation or biofilm formation. A strain lacking a functional agn43 allele (E. coli deltaflu) failed to autoaggregate and failed to form a biofilm. doi:10.1371/journal.pone.0008801.g006
well characterized Pet is a serine protease autotransporter (SPATE) secreted by EAEC 042 which induces dilation of crypt openings and rounding and extrusion of enterocytes in human tissue explants [15,18]. Once secreted Pet exerts its toxic effects by being internalized into host cells where it cleaves the host cytoskeletal protein spectrin [94,95]. While Pet may play a role in EAEC mediated disease it is only present in a minority of strains [59]. In contrast, EAST-1, encoded by the astA gene adjacent to pet, is present in a wide variety of strains and different pathovars [59]. EAST-1 induces an increase in short circuit current in Ussing chambers indicative of a net anion secretion [96,97]. However, the presence of EAST-1 in EAEC 17-2, a strain which did not cause diarrhea in volunteers, suggests that EAST-1 is not the sole mediator of diarrhea, an observation confirmed by studies which demonstrated EAST-1 from EAEC 042 in in vitro models of toxicity [7,97].

In addition to Pet and EAST-1, EAEC 042 chromosomally encodes at least two additional toxins, ShET1 and HlyE. HlyE (Ec042-1231) is a 34-kDa, predominantly α-helical, protein which oligomerises into higher order structures to form a pore-forming toxin mediating cytolytic and cytopathic effects on cultured human cells [96]. Several lines of evidence, including a complex regulatory circuit and recognition by convalescent antisera, indicate a role in disease [99,100]. However, the occurrence of hlyE amongst non-pathogenic bacteria [101] suggests that if HlyE plays a role in mediating disease then it is a minimal role. In contrast to HlyE, the ShET1 toxin is a subunit toxin encoded by setA and setB (Ec042-4593a,b), which are thought to form an oligomeric toxin consisting of a single 20-kDa SetA protein associated with a pentamer of SetB subunits [19]. ShET1 appears to induce intestinal secretion via cAMP and cGMP, however the precise mechanism of action and detailed biochemistry remains elusive. Unusually, the setAB genes are encoded within the pic gene but on the complementary strand and thus have the same prevalence characteristics and disease associations as pic [19].

Surface Polysaccharides

Polysaccharides are major determinants of virulence in Gram-negative bacteria. Surface exposed polysaccharides play a dual role mediating interaction of the bacterium with the environment whilst creating a barrier to noxious substances [102]. EAEC 042 possesses four major loci involved in polysaccharide synthesis which distinguishes it from E. coli K-12 laboratory strains. These are Ec042-2270–99, responsible for synthesis of the O44 serotype O-antigen of lipopolysaccharide, two copies of the Shf locus encoding a myristoyl transferase for modifying the lipid A moiety of LPS (Ec042-4769–72 and Ec042-pAA021–023) and a locus encoding capsular polysaccharide (Ec042-3230–40). Analysis of the CDS from Ec042-3230 to Ec042-3240 revealed the presence of a Group 2 capsule gene cluster with conserved regions 1 and 3 flanking a central variable region 2 encoding for the biosynthesis of the particular capsular polysaccharide (Fig. S18). Genes Ec042-3230–3235 encode proteins that were 99% identical to the KpsF, E, D, U, C and S proteins encoded by region 1 of group 2 capsule gene clusters involved in polysaccha-
ride export. Likewise E.coli is considered to be virulent in adult volunteers. The sequenced strain was obtained from the original stock kept at the Center for Vaccine Development and was subjected to minimal laboratory passages. The whole genome was sequenced to a depth of 0 x coverage from pUC19 (insert size 2.8-5 kb) and pMAQ1b (insert size 5.5-10 kb) small insert libraries using dye terminator chemistry on ABI3700 automated sequencers. End sequences from larger insert plasmid (pBACe3.6, 20-30 kb insert size) libraries were used as a scaffold. The sequence was assembled and finished as described previously [3].

Gene Prediction, Annotation and Comparative Analysis

CDSs were first identified using GeneHacker followed by manual inspection of start codons and ribosome binding sequences of each CDS. Intergenic regions of >150 bp were further reviewed for the presence of small CDS encoding proteins with significant homology to known proteins. Functional annotation of the CDSs was made on the basis of results of homology searches against the public non-redundant protein database (http://www.ncbi.nlm.nih.gov/BLAST). Genes for tRNAs, tmRNAs, rRNAs and other small RNAs were identified using the Rfam database through the Rfam website (http://www.sanger.ac.uk/Software/Rfam/index.shtml). We also searched the EAEC 042 genome for all the RNA genes that have been identified in E. coli K-12 and Sakai by BLASTN. The annotated genome sequences of EAEC 042 have been deposited in the public databases database (accession numbers: N554766 for EAEC 042 complete genome and FN554767 for the EAEC 042 plasmid pAA).

A search for type-III-secretion effectors in the genome was performed using a previously described set of effectors as the input to BLASTP searches of EAEC 042 CDS predictions [106]. In addition, comparisons of the EAEC 042 and EHEC genomes via the xBASE facility were used to identify positional orthologs of known EHEC effectors [107].

E. coli Core Genome and Pangenome

To assess the size of the core genome and pangenome of E. coli a set of 24 complete or almost complete E. coli and Shigella genomes was selected. These included the complete genome sequences of E. coli strains K-12 MG1655 (GenBank accession U00096), O157:H7 EDL933 (AE005174 and AF074613), CFT073 (AE014075), S. flexneri 3a 2457T, and E. coli strains MG1655, EDL933 and 2457T, respectively. The sequenced strain was obtained from the original stock kept at the Center for Vaccine Development and was subjected to minimal laboratory passages. The whole genome was sequenced to a depth of 0 x coverage from pUC19 (insert size 2.8-5 kb) and pMAQ1b (insert size 5.5-10 kb) small insert libraries using dye terminator chemistry on ABI3700 automated sequencers. End sequences from larger insert plasmid (pBACe3.6, 20-30 kb insert size) libraries were used as a scaffold. The sequence was assembled and finished as described previously [3].

Materials and Methods

Bacterial Strain and Sequencing

The EAEC O44:H18 strain 042 was isolated from a child with diarrhea in the course of an epidemiologic study in Lima, Peru, in 1983 and was subsequently shown to cause diarrhea in adult volunteers. The sequenced strain was obtained from the original stock kept at the Center for Vaccine Development and was subjected to minimal laboratory passages. The whole genome was sequenced to a depth of 0 x coverage from pUC19 (insert size 2.8-5 kb) and pMAQ1b (insert size 5.5-10 kb) small insert libraries using dye terminator chemistry on ABI3700 automated sequencers. End sequences from larger insert plasmid (pBACe3.6, 20-30 kb insert size) libraries were used as a scaffold. The sequence was assembled and finished as described previously [3].

EAEC Genome Sequence

The chromosomal shf locus consists of four monocistronically transcribed genes previously designated shf, shfU, shfK and msbB2. In contrast, the plasmid-encoded locus contains only shf, shfU and shfK but displays 98% nucleotide sequence identity with the chromosomal locus over 3155 bp (Fig. S19). Both loci possess the previously identified PhoP/Q, magnesium and temperature regulated promoter with no interruptions in the CDS suggesting both loci can be transcribed to produce the three proteins [104]. The function of the locus has remained enigmatic. A recent investigation of the locus revealed an insertion in the plasmid copy of shf diminished biofilm formation of EAEC 042 but that deletion of the downstream genes did not [105]. This observation is difficult to reconcile since the locus is monocistronic and since a second copy of the gene exists on the chromosome. However, the demonstrated role of RbU and MsbB2 in LPS modification suggests that this locus is likely to be involved in altering the cell surface through modification of various domains of LPS but only under specific environmental conditions. A variety of studies have demonstrated that this locus is widespread amongst EAEC isolates, and given its important role in the ability of other organisms to cause disease it warrants further study in relation to EAEC pathogenesis [59].

Conclusion

EAEC is an increasingly recognized enteric pathogen, implicated in diverse clinical and epidemiologic scenarios. Full understanding of the pathogenesis and epidemiology of this organism has been hampered by considerable genomic diversity of clinical isolates. The heterogeneity of virulence was clearly demonstrated in volunteers, but the basis of this heterogeneity has not been characterized. Here, we present the first comprehensive genomic analysis of the prototype strain 042, which was shown to be virulent in adult volunteers. Though our analyses do not yet suggest the basis of this enhanced pathogenicity, the genome of 042 was found to possess many genetic characteristics of pathogenic Shigella, Salmonella and diarrheagenic E. coli strains. These factors include (but are not limited to) apparently complete type II, III, and VI secretion systems (including likely effectors), multiple autotransporter proteins, several proven and putative adhesins, polysaccharide and lipopolysaccharide modification loci, and iron scavenging systems. Comparative functional genomics, where the phenotype of the organism can be related to the genome, has shown that some of the differences between EAEC 042 and E. coli MG1655 can be related to other genomic differences, often due to loss of single genes, or simple mutations, but also because whole operons are not present in one strain or the other, or that EAEC 042 has acquired a mobile genetic element. This series of experiments demonstrates the feasibility of whole genome-phenome comparisons in search of roles for genes of unknown function. Our studies have provided substantial grist for further experimental studies of EAEC pathogenesis.
the query protein. To overcome potential problems associated with variations in annotation between the different genomes, the predicted protein sequences were also used as the query sequences in tblastn searches against the total set of genome sequences. If a homologue was identified in any of the genomes in the tblastn search that was not identified in the blastp search, using the same criteria as above, then an unannotated gene potentially encoding a homologous protein was considered to be present. The size of the core genome and pangenome were assessed using methods similar to those previously described [22,108].

As the E. coli K-12 MG1655 genome is thought to be the most comprehensively annotated strain, 2370 annotated genes from that genome which were conserved in all the other strains were used to represent the E. coli core genome (listed in Table S2). These genes were used as the query sequences in further tblastn searches against a database consisting of complete and almost complete non-E. coli enterobacterial genomes. The included genomes were from Escherichia albertii, Escherichia fergusonii, Citrobacter koseri, S. enterica serovars Typhi, Typhimurium and Gallinarum; Yersinia pestis strains KIM, CO92, Angola, Antiqua, Nepal516 and Pestoides F; Yersinia pseudotuberculosis strains IP 31756, IP 32953, PB1/+ and YPIII; Yersinia enterocolitica; Seratia proteamnaulans; Proteus mirabilis; Photobahbus luminescens; Klebsiella pneumoniae; Escheria tasmaniana; Enterobacter sp. 638; Enterobacter sakazaki and Pectobacterium atrosepticum. Organisms from the genera Sodalis, Wigglesworthia, Buchnera and Blochmannia were omitted from this analysis since they are endosymbionts with reduced genomes that are not representative of the genomes of free-living enterobacteria. The same criteria of >60% identity over at least 80% of the length of the query sequence were used to determine the presence or absence of homologues.

Phylogenetic Relationships within E. coli

To investigate the phylogenetic relationships among the E. coli genomes we selected a set of 2173 E. coli K-12 genes that were conserved in all the other E. coli genomes and in E. albertii and E. fergusonii, which were included as an outgroup. The homologous sequences from our representative set of E. coli genomes were compiled and aligned using ClustalW [109]. A maximum likelihood phylogeny was obtained using the general time reversible (GTR/REV) model with the CAT approximation of rate heterogeneity as implemented in RAxML version 7.0.4 [110]. Support for individual branches was assessed by conducting 100 non-parametric bootstrap replicates, using the rapid algorithm implemented in RAxML [111]. The phylogeny was displayed using MEGA version 4 [112].

Molecular biology techniques and functional assays. PCR reactions were performed using BioLine Readymix according to manufacturer's instructions. Primers for amplification of Ec042-2242, 4511 and 4803 alleles possessed the following sequences 5’-CTGAGCTCGTGAAGATTTACCGGTCG-3’ (forward primer), 5’-CAGAAAGTCCGCCGCAACCCCCGTCTTTTGACACA-3’ (2242), 5’-GGCCGGACCTCCTAGACAGAACCTGCCTGC-3’ (4511) and 5’-TTTCTCTAGTCATGACTAGTGTTGAAAGACAGG-3’ (4803). All reactions were performed with a 60°C annealing temperature. Reaction products were analysed by DNA agarose electrophoresis as previously described. Products were cloned into pCRII-TOPO (Invitrogen) according to manufacturers' instructions. Proteins were analysed by SDS-PAGE and Western immunoblotting as previously described [113].

The standard autoaggregation assay was performed as previously described; overnight cultures were allowed to stand and at various time points samples were removed from the top of the culture medium and the OD_{600} was measured [113]. Autoaggregation is noted as a decrease in the OD_{600} value. A biofilm assay was performed essentially as previously described; bacteria were grown overnight in 2 ml of LB medium in 15 ml polystyrene tubes before staining with crystal violet to visualise the biofilm [113].

Based on the sequence and phenotypic PM data the susceptibility of strains to a range of compounds was determined using the agar dilution method following guidelines of the British Society for Antimicrobial Chemotherapy [114]. To determine whether any of the changes in antimicrobial susceptibility seen between strains were due to altered membrane permeability/active efflux of antimicrobials the accumulation of the fluorescent dye Hoechst 33342 (bis-benzimide), a substrate of the major AcrAB-ToIC efflux system was measured. Experiments were repeated at least four times in the presence and absence of the efflux pump inhibitor phenyl-arginine-β-naphthylamide (PAbN) or CCCP as previously described [115].

Phenotype Microarrays

Phenotype microarray growth and respiration analyses were performed as described previously by Biolog Inc. (Hayward, California USA) [116,117]. Colonies from the test strains (E. coli 042 and E. coli K-12 MG1655) were inoculated from pregrowth R2A or LB agar plates [116] into Biolog inoculating fluid and then grown in Biolog defined minimal medium or nitrogen-, sulphur- or phosphorous-free versions of the medium in twenty PM plates, which tested nearly 2000 phenotypes. Growth at 36°C and respiration was measured every 15 minutes for 24 hours as colour changes using an Omnilog reader. Tests were performed in duplicate, and the mean signal in arbitrary units calculated for each replicate before the arithmetic difference of the mean EAEC 042 signal minus the mean of the E. coli MG1655 signal for each test well was calculated.

Supporting Information

Table S1 Regions of difference (RODS) in the EAEC 042 genome. Found at: doi:10.1371/journal.pone.0008801.s001 (0.04 MB XLS)

Table S2 List of genes conserved in all sequenced E. coli genomes. Found at: doi:10.1371/journal.pone.0008801.s002 (1.99 MB DOC)

Table S3 Comparison of EAEC 042 and E. coli MG1655 metabolism by phenotype microarray, where EAEC 042 shows greater metabolic activity. Found at: doi:10.1371/journal.pone.0008801.s003 (0.12 MB DOC)

Table S4 Comparison of EAEC 042 and E. coli MG1655 metabolism by phenotype microarray, where E. coli MG1655 shows greater metabolic activity. Found at: doi:10.1371/journal.pone.0008801.s004 (0.07 MB DOC)

Table S5 Comparison of iron transport / storage genes from MG1655, O157 [Sakai] and O42. Found at: doi:10.1371/journal.pone.0008801.s005 (0.05 MB XLS)

Table S6 List of EAEC 042 CDS conserved in the other sequenced EAEC genomes (101-1 and 55989), but absent from the genome of the commensal E. coli HS.
Figure S2 Genetic map of pAA, the large virulence plasmid of EAEC 042. (A) On the basis of nucleotide sequence homology, the plasmid pAA belongs to the IncFIIA family and carries just one identifiable replicon consisting of Ec042-pAA152 (RepA) and Ec042-pAA153 (CopB), which is in contrast to many of the F-family plasmids that have multiple replications. It possesses auxiliary stable inheritance functions including Ec042-pAA136–137 encoding the type 2 partitioning proteins ParM and ParR, a Hok/Sok post-segregational killing system (Ec042-pAA106) and the putative pair of Ec042-pAA147–148 encoding a RelE/StbE homologue. As for most IncFI plasmids, pAA appears to encode a complete F-like conjugal transfer system (Ec042-pAA066–100), most closely related to those of pUT118 and R100, and does not seem to lack any standard component which might explain why attempts to transfer a derivative tagged with an antibiotic resistance marker (IH, unpublished) have been unsuccessful. However, analysis of the predicted gene products identified two CDS with unusual features that might be worth investigating as the basis of a transfer defect: TraP (Ec042-pAA092) shows N-terminal segments differing significantly from the nearest relatives despite high overall sequence alignment; and TraN (Ec042-pAA069) contains an internal region towards the end of the protein with a significant amplification of a run of PQQP repeats which may have caused it to become non-functional. Functional analyses are needed to determine whether these features could be responsible for the Tra- phenotype. Other transfer-associated genes are: Ec042-pAA129 encoding a putative anti-restriction gene; Ec042-pAA113 encoding Ssb (single stranded DNA binding protein); Ec042-pAA109–110 encoding the SOS-induced response proteins PsiA/B associated with plasmid transfer; and Ec042-pAA101 encoding a lytic transglycosylase that helps hydrolyse cell wall in recipients prior to conjugal transfer. With respect to overall organisation it is interesting that the point at which a large amount of mobile DNA has been inserted into the plasmid is between finO (Ec042-pAA066) at the end of the transfer region and the rep region (Ec042-pAA0152–0153). In many, but not all, F-like plasmids these are contiguous and there is evidence that this region is involved in coordination between the transfer process and plasmid copy number replication. However, in F there is an IS insertion in the finO gene itself, so there is precedent for disruption of this region. Plasmid pAA is relatively unusual in having most of the transposable elements inserted into this particular region. Thus, the plasmid has many standard elements but with features that render it worthy of further study.

Table S7 Type 3 secretion system effector genes in the EAEC 042 genome.

Figure S1 Genetic similarity of the EAEC 042 prophage. Nine prophage regions, designated 042p1–042p9, were identified in the EAEC 042 genome (see Table 2 in manuscript). The figure represents homology of the phage elements; solid lines indicate complete identity, the absence of lines within the boxes reflect little or no homology and the presence of dashed lines indicates partial identity. Four of these were lambdoid in nature (042p2, 042p3, 042p4 and 042p6) and were highly similar to each other and to the lambdoid prophages of E. coli O157:H7 [1], and E. coli O127:H6 strain E2348/69 (EPEC) [2]. Some of the related lambdoid prophages in EAEC 042 and E. coli O157:H7 are integrated in corresponding genomic locations and share some sequence identity; 042p3 and Sp10, 042p4 and Sp11/Sp12, and 042p6 and Sp14 are found relative to each other in the respective bacterial genomes. The EAEC 042 lambda-like prophages also have high homology to lambdoid prophages in several of the other sequenced pathogenic E. coli genomes including avian pathogenic (APEC), uropathogenic (UPEC) and enterotoxigenic (ETEC). In addition, 042p4 has high homology to the K-12 cryptic prophage Qin and genome comparison shows where the deletions have occurred in the Qin genome to render it defective. 042p5 and 042p7 have some sequence similarity to O157:H7 Sakai prophage Sp7 which is of an unstudied type [1,3]. Half of prophage 042p1, from the non-tail end to the lysis module, is related to lambda, P22 and PPB in EPEC, but the remainder of the prophage, which encodes mainly structural proteins, is highly related to the Shigella flexneri serotype converting phage SIV [4]. 042p1 encodes bacterioprenol glucosyl transferase (GtrB) and glucose translocase (GtrA), at the very end of the prophage, after the tail genes. These proteins are involved in O-antigen modification and have homology to the sorotype-converting proteins of SIV (87% protein identity to GtrA and GtrB) and P22 (88% protein identity to GtrA and 77% to GtrB) [4,5]. It is therefore likely that the presence of 042p1 in the genome of EAEC 042 alters the sorotype of the host bacterium. 042p3 encodes an iron ABC transport system SitABCD which is carried at the tail end of the prophage and is discussed later. The only other EAEC 042 prophage for which we have identified putative cargo is 042p6 which encodes a putative exo-oxyribonuclease VIII. 042p9 is a P4-like satellite phage, and although there is no P2-like phage in the EAEC 042 genome for it to parasitise, it is possible that 042p9 is able to induce virion morphogenesis in a different helper phage [6]. 042p8 has no sequence similarity to any currently known phages and its only identified homologue is ROD-9 in ETEC H10407. Five of the EAEC 042 prophages (042p1, 042p2, 042p3, 042p4, and 042p6) appear to have all the necessary genes to produce fully functional phages. The other prophage regions, with the exception of the satellite prophage 042p9, are probably non-functional prophage-like remnants.

Table S1 Genetic similarity of the EAEC 042 prophage. Nine prophage regions, designated 042p1–042p9, were identified in the EAEC 042 genome (see Table 2 in manuscript). The figure represents homology of the phage elements; solid lines indicate complete identity, the absence of lines within the boxes reflect little or no homology and the presence of dashed lines indicates partial identity. Four of these were lambdoid in nature (042p2, 042p3, 042p4 and 042p6) and were highly similar to each other and to the lambdoid prophages of E. coli O157:H7 [1], and E. coli O127:H6 strain E2348/69 (EPEC) [2]. Some of the related lambdoid prophages in EAEC 042 and E. coli O157:H7 are integrated in corresponding genomic locations and share some sequence identity; 042p3 and Sp10, 042p4 and Sp11/Sp12, and 042p6 and Sp14 are found relative to each other in the respective bacterial genomes. The EAEC 042 lambda-like prophages also have high homology to lambdoid prophages in several of the other sequenced pathogenic E. coli genomes including avian pathogenic (APEC), uropathogenic (UPEC) and enterotoxigenic (ETEC). In addition, 042p4 has high homology to the K-12 cryptic prophage Qin and genome comparison shows where the deletions have occurred in the Qin genome to render it defective. 042p5 and 042p7 have some sequence similarity to O157:H7 Sakai prophage Sp7 which is of an unstudied type [1,3]. Half of prophage 042p1, from the non-tail end to the lysis module, is related to lambda, P22 and PPB in EPEC, but the remainder of the prophage, which encodes mainly structural proteins, is highly related to the Shigella flexneri serotype converting phage SIV [4]. 042p1 encodes bacterioprenol glucosyl transferase (GtrB) and glucose translocase (GtrA), at the very end of the prophage, after the tail genes. These proteins are involved in O-antigen modification and have homology to the sorotype-converting proteins of SIV (87% protein identity to GtrA and GtrB) and P22 (88% protein identity to GtrA and 77% to GtrB) [4,5]. It is therefore likely that the presence of 042p1 in the genome of EAEC 042 alters the sorotype of the host bacterium. 042p3 encodes an iron ABC transport system SitABCD which is carried at the tail end of the prophage and is discussed later. The only other EAEC 042 prophage for which we have identified putative cargo is 042p6 which encodes a putative exo-oxyribonuclease VIII. 042p9 is a P4-like satellite phage, and although there is no P2-like phage in the EAEC 042 genome for it to parasitise, it is possible that 042p9 is able to induce virion morphogenesis in a different helper phage [6]. 042p8 has no sequence similarity to any currently known phages and its only identified homologue is ROD-9 in ETEC H10407. Five of the EAEC 042 prophages (042p1, 042p2, 042p3, 042p4, and 042p6) appear to have all the necessary genes to produce fully functional phages. The other prophage regions, with the exception of the satellite prophage 042p9, are probably non-functional prophage-like remnants.

Figure S3 Global comparison between EAEC 042 chromosome and those of EHEC, UPEC and K-12. ACT comparison (http://www.sanger.ac.uk/Software/ACT) of amino-acid matches between the complete six-frame translations (computed using TBLASTX) of the whole genome sequences of enterohaemorrhagic E. coli O157:H7 str. Sakai (EHEC; EMBL acc: BA000007), enterohaemorrhagic E. coli O157:H7 EDL933 (EHEC; EMBL acc: AE005174), uropathogenic E. coli strain CFT073 (UPEC; EMBL acc: AE014075), and E. coli strain K-12 MG1655 (K-12; EMBL acc: U00096). Forward and reverse strands of DNA are shown for each genome (dark grey lines). The red bars between the DNA lines represent individual TBLASTX matches, with inverted matches coloured blue.

Figure S4 Box-plot showing the estimated core genome size as a function of the number of genomes sequenced. The plot was constructed based on 100 randomly selected permutations of 30 genome sequences. The curve is approaching the limit suggesting that the 2356 conserved genes identified in this study represent the minimal E. coli genome and that this number is unlikely to reduce much further with additional E. coli genome sequencing.
with those from *E. coli* K12 MG1655. NmpC (OmpD) is a pseudogene in *E. coli* K12 characterised by a C-terminal deletion. Identical residues are marked by asterisks, whereas similar residues are marked by periods and colons. The amino acid sequences of the OmpF, OmpC, PhoE and OmpN alleles from EAEC 042 and *E. coli* MG1655 are relatively well conserved (see Figure 4 in manuscript). In contrast, the OmpD proteins are more divergent and Ec042-2121 is sufficiently divergent to represent an apparently different lineage. Such divergence may be indicative of additional functions.

**Figure S13** Genetic architecture and distribution of the Chaperone-Usher systems of EAEC 042. (A) The organisation of the fimbrial loci from EAEC 042 are depicted. CDS flanking the fimbrial loci are depicted by green arrows, CDS encoding regulators are depicted by purple arrows, CDS encoding chaperones are indicated with black arrows, ushers are depicted with red arrows and the secreted fimbrial subunits by blue arrows. CDS of unknown function are depicted by orange arrows. The EAEC CDS designation for the first flanking gene is indicated along with the common name for the loci. The loci designated 0022 and 3336 are uncharacterised fimbrial loci and the numbers represent the EAEC CDS designation for the gene encoding the putative usher protein. (B) Phylogenetic distribution of the fimbrial loci amongst pathogenic and non-pathogenic lineages of *E. coli*. The loci encoding the EAEC fimbrial systems demonstrate a differential distribution amongst the *E. coli* phylogeny but with some evidence of phylogenetic clustering.

Found at: doi:10.1371/journal.pone.0008801.s020 (0.26 MB DOC)

**Figure S14** Genetic architecture of the RTX-like Type 1 secretion locus. An alignment of the genomic regions encompassing the TISS is depicted, demonstrating the absence of the locus in *E. coli* K12 and the presence of the locus in EAEC 042 and *E. coli* O157:H7. In both pathogenic strains the gene encoding the putative secreted RTX-like protein is frameshifted and would be predicted to encode a non-functional protein. The frameshifts occur at different positions within the respective genes; in EAEC 042 there are 3 frameshifts whereas in *E. coli* O157:H7 there is only one frameshift. The position of the OMP, ABC, MFP and a putative regulator are indicated.

Found at: doi:10.1371/journal.pone.0008801.s021 (0.40 MB DOC)

**Figure S15** Genomic architecture of the Type II secretion system (T2SS) apparatus of EAEC 042 and representative *E. coli* strains. The T2SS locus (gsp) appears intact in EAEC 042 and a variety of phylogenetically disparate strains of pathogenic *E. coli*. In contrast, the locus possesses a lesion in *E. coli* K12 and is absent in *E. coli* CFT073. Where the locus is intact it is always syntenic with yghJ a gene encoding a putative secreted lipoprotein.

Found at: doi:10.1371/journal.pone.0008801.s022 (0.38 MB DOC)

**Figure S16** Phylogenetic distribution of members of the Type 5 secretion system present in EAEC 042. The Autotransporters (AT-1) are differentially represented amongst the *E. coli* phylogeny, though several members of this group show phylogenetic clustering. The Two-partner secretion system (TPS) is present in all strains of *E. coli* suggesting an important physiological role. The trimeric Autotransporters (AT-2) are differentially represented amongst the *E. coli* phylogeny. Only upaG, an extracellular matrix binding protein, shows a wide distribution and is present in both
commensal and pathogenic lineages suggesting the protein is not directly responsible for the ability of E. coli to mediate disease. Found at: doi:10.1371/journal.pone.0008801.s025 (0.11 MB DOC)

Figure S17 Comparison of three type VI secretion systems (T6SSs) of EAEC 042, and a separate Vgr-encoding region. Genes depicted by the schematic are indicated by the “Ec042” numbers on the right of the figure. Genes are designated by arrows, which are colored by xBASE to correspond to GC content, except for Hcp- and Vgr-homologs, which are shown in purple and white, respectively. Red blocks connecting genes indicate significant (P<0.05) identity on the amino acid level among homologs in all three T6SSs. Blue connectors indicate homologs that were only found within two T6SSs. The locus labeled Ec042-1384–1388 depicts a chromosomal region that encodes a Vgr-homolog, but is not in proximity to the other T6SSs. One of the T6SS identified (Ec042-4562–4577) had been previously characterized [7] as a locus under the control of AggR, the master regulator of virulence in this strain. The genes are part of a 117-kb pathogenicity island that is inserted at the phoU locus. This T6SS does not encode an Hcp homolog [7], however it is speculated that the third gene in this locus, previously designated aaiC, is a functional homolog. BLASTN results of this entire locus indicate that a nearly identical 18.4 kb sequence is found in enterohagggregative E. coli (EAEC) strain 55989 [8]. This is consistent with the previous report [7] that homologs of aaiC and the first gene of this locus designated aaiA are widely distributed among strains of EAEC. Unlike Hcp from other organisms, which is found widely within T6SSs [9], homologs of AaiC are only found within putative T6SSs of the atypical EAEC strain 101-1 (33% amino acid identity; GenBank AAMK02000010) and Citrobacter youngae ATCC 29220 (30% amino acid identity; GenBank EK20055). Therefore, the function of this T6SS may be specific to the pathogenesis of EAEC and a limited number of other pathogens. Found at: doi:10.1371/journal.pone.0008801.s024 (0.21 MB DOC)

Figure S18 Genetic architecture of the capsular polysaccharide locus from EAEC 042. The locus is required for the biosynthesis of a group 2 capsular polysaccharide and appears to contain all of the functional genes necessary for capsular production. Three CDS are present in the central region 2 and are likely to be important in the biosynthesis of the particular capsular polysaccharide expressed by this strain of E. coli. The ORF Ec042-3238 had 33% identity and 53% homology over 202 amino acids to CMP-N-acetylneuraminic acid synthetase enzymes from a number of bacterial species. This enzyme catalyses the conversion of CTP and neuraminic (NeuNAc) acid to form CMP-NeuNAc a key sugar activation step for the subsequent incorporation of NeuNAc into polysaccharides. At this stage in the absence of biochemical data one cannot be unequivocal about the enzymatic activity of the encoded protein but one can be sure it is involved in the activation of a nine-carbon sugar, which may be NeuNAc. Gene Ec042-3237 is predicted to encode a large protein of 1113 amino acids and has significant homology to the HAD super family of hydrolase enzymes and a number of putative glycosyl transferases from Shigella enterica. This analysis would suggest a role in polysaccharide biosynthesis but in the absence of any structural data for the capsular polysaccharide expressed by this strain assigning a precise function to this protein is difficult. The remaining ORF in region 2 Ec042-3236 has significant homology (50% over 206 amino acids) with a putative acetyl-transferase from Neisseria meningitidis and less homology to similar acetyl-transferases from a number of other bacteria. One possibility is that this is an acetyltransferase that acylates the capsular polysaccharide of this strain. Acetylation of the K1 polysaccharide of E. coli is known to occur and has been suggested as a mechanism by which the antigenicity of the cell surface polysaccharide may be modified in a stochastic fashion [10]. Analysis of the sequences 800 bp 5’ to Ec042-3230 showed no differences to those sequences 5’ to group 2 capsule gene clusters in strains of E. coli known to encode group 2 capsules such as UT189, CFT073 and APEC 01. This conservation of the promoter and regulatory region would indicate that regulation of expression of this capsule gene cluster in Ec042 involves ShyA, H-NS, IHF and BifA as previously demonstrated for other group 2 capsule gene clusters [11]. Implicit in this is that the expression will be temperature regulated being on at 37°C but not at 20°C [12]. A similar analysis of the sequences 5’ to Ec042-3240 showed no differences to those sequences 5’ to KpsM in other group 2 capsule gene clusters. The ops site 28 bp 5’ to kpsM, which is the cis-acting site for RfaH and is essential to allow read through transcription from the region 3 promoter into region 2 [13] was also conserved 28 bp 5’ to Ec042-3240. This conservation of the region 3 promoter and regulatory sequences would indicate that this group 2 capsule gene cluster is functional. The gene 5’ to region 3 in Ec042 is the gene encoding for the putative general secretion pathway protein YghD. This is the same organization as with other group 2 capsule gene clusters. Likewise the sequence 5’ to Ec042-3230 including the ORF Ec042-3229 also flank group 2 capsule gene clusters indicating these flanking regions have been acquired with the group 2 capsule gene cluster. Found at: doi:10.1371/journal.pone.0008801.s025 (0.11 MB DOC)

Figure S19 Comparison of the EAEC 042 chromosomal- and plasmid-based copies of the shf loci. The sequences surrounding the loci are not homologous. The ca. 3 kb of nucleotide sequence encompassing shf, rfbU and virK is identical; mshB2 is absent from the plasmid copy but present in the chromosomal copy. The similarity of Shf to the Sphingobacterium spadernudis protein IcaB, which is required for exopolysaccharide modification and biofilm formation, indicated that Shf might also play a role in polysaccharide modification, a hypothesis supported by the presence of a polysaccharide deacetylase domain within Shf [14]. Recently, RfbU was discovered to be responsible for the addition of α-1,7-GlcN to the R3 core region of LPS and has been renamed WabB [15]. The R3 core is found in a variety of pathogenic E. coli and Shigella and is of significant biomedical interest. The function of VirK remains elusive. Initially described in Shigella, where it was found to be necessary for localisation of the autotransporter IcsA to the bacterial surface and subsequent intracellular spreading [16], VirK has been characterised in Salmonella where it has also been shown to be important for virulence [17]. MshB2 was recently shown to act as a myristoyl transferase which modifies the lipid A portion of LPS and acts in a manner analogous to the chromosomal gene lpxM [18]. Furthermore, several investigations have demonstrated the importance of this gene for full virulence in Shigella and E. coli. Deletion of both mshB2 and lpxM results in altered membrane fatty acid composition and susceptibility to a variety of antibiotics and detergents suggesting defects in membrane biogenesis [19]. Found at: doi:10.1371/journal.pone.0008801.s026 (0.15 MB DOC)

File S1 References for supplementary data. Found at: doi:10.1371/journal.pone.0008801.s027 (0.04 MB DOC)
References


