The population genetics of pathogenic *Escherichia coli*

Erick Denamur[1,2,3]✉, *Olivier Clermont*[1,2], *Stéphane Bonacorsi*[1,4,5] *and David Gordon*[6]

Abstract | *Escherichia coli* is a commensal of the vertebrate gut that is increasingly involved in various intestinal and extra-intestinal infections as an opportunistic pathogen. Numerous pathotypes that represent groups of strains with specific pathogenic characteristics have been described based on heterogeneous and complex criteria. The democratization of whole-genome sequencing has led to an accumulation of genomic data that render possible a population phylogenomic approach to the emergence of virulence. Few lineages are responsible for the pathologies compared with the diversity of commensal strains. These lineages emerged multiple times during *E. coli* evolution, mainly by acquiring virulence genes located on mobile elements, but in a specific chromosomal phylogenetic background. This repeated emergence of stable and cosmopolitan lineages argues for an optimization of strain fitness through epistatic interactions between the virulence determinants and the remaining genome.

Pathotypes (also known as pathovars)
Groups of organisms that have the same pathogenicity on a specified host.

[1]*Université de Paris, IAME, UMR1137, INSERM, Paris, France.*

[2]*Université Sorbonne Paris Nord, IAME, Paris, France.*

[3]*AP-HP, Laboratoire de Génétique Moléculaire, Hôpital Bichat, Paris, France.*

[4]*AP-HP, Laboratoire de Microbiologie, Hôpital Robert Debré, Paris, France.*

[5]*Centre National de Référence Escherichia coli, Hôpital Robert Debré, Paris, France.*

[6]*Ecology and Evolution, Research School of Biology, The Australian National University, Acton, ACT, Australia.*

✉*e-mail: erick.denamur@inserm.fr*

https://doi.org/10.1038/s41579-020-0416-x

Escherichia coli is a commensal member of the vertebrate gut microbiota[1] as well as an opportunistic pathogen[2,3] of mammals and birds. *E. coli* is the predominant aerobic bacterium of the gut microbiota, although it is outnumbered by anaerobic bacteria 100:1–10,000:1. In humans, its prevalence is more than 90% with a concentration per gram of faeces from 10^7 to 10^9 colony-forming units[1]. *E. coli* strains can cause both extra-intestinal pathologies (urinary tract infections (UTIs), diverse intra-abdominal, pulmonary, skin and soft tissue infections, newborn meningitis (NBM) and bacteraemia) and intestinal pathologies (various forms of diarrhoea, including haemolytic and uraemic syndrome (HUS)). These infections can be very common (UTIs)[4], associated with high morbidity (renal failure in HUS in children[5], neurologic sequelae in NBM[6]) and high mortality (~15% in bacteraemia[7,8]). The incidence of extra-intestinal infections is increasing in humans[9], and we regularly experience major HUS epidemics, such as the 2011 epidemic in Europe[10]. Furthermore, antibiotic resistance in *E. coli* is rising[11] and it now ranks third in the list of the 12 antibiotic-resistant 'priority pathogens' described by the WHO.

E. coli pathogenic strains are usually classified into pathotypes (also known as pathovars)[2,3], and they are identified using acronyms. These pathotypes have been proposed over time as specific discoveries have been made and are not unified in a meaningful way. The definition of these pathotypes can be based on various criteria, such as the target organ (for example, urinary tract and uropathogenic *E. coli* (UPEC)); the infected host (for example, bird and avian pathogenic *E. coli* (APEC)); the association with an organ and host (for example, cerebrospinal fluid in newborns and newborn meningitis *E. coli* (NMEC)); the association with the targeted organs, the presence of specific genes or the virulence in an animal model (for example, extra-intestinal pathogenic *E. coli* (ExPEC)); the pathology caused by the strains (for example, diarrhoea and intestinal pathogenic *E. coli* (InPEC)); the presence of a specific gene or genes, alone or in combination (for example, Shiga-toxin encoding *stx* gene and Shiga toxin-producing *E. coli* (STEC), intimin-encoding *eae* with or without pili-encoding *bfp* gene(s) and typical or atypical enteropathogenic *E. coli* (tEPEC or aEPEC, respectively)); or a specific ex vivo phenotype (for example, adhesion and invasion of epithelial cells and adherent-invasive *E. coli* (AIEC)). A detailed list of the most commonly used pathotypes with their main characteristics is presented in TABLE 1. Extensive knowledge of the molecular and cellular mechanisms of *E. coli* pathogenicity has accumulated over the years[2,3].

During recent years, complex hybrid pathotypes have emerged, either within the InPEC pathotypes (for example, enterohaemorrhagic *E. coli* (EHEC) and enteroaggregative *E. coli* (EAEC)) or between InPEC and ExPEC pathotypes (for example, EHEC and ExPEC) (TABLE 1), rendering the pathotype classification difficult to follow. The description of cryptic *Escherichia* clades[12] and the difficulty to identify other *Escherichia* species resulted in additional confusion. Concomitantly, the democratization of whole-genome sequencing (WGS) has led to the accumulation of genomic data that may enable phylogenomic approaches to classify pathogenic *E. coli* strains. In this context, an overview of the emergence

Table 1 | **Main characteristics of the more commonly used *Escherichia coli* pathotypes**

Pathotype[a]	Definition basis	Main strain host	Main virulence genes	Strain phylogenetic background	Main ST[WU]	Main serotypes[b]
ExPEC	Non-intestinal infection, specific genes, animal model	Human, domestic mammals, birds	Genes encoding adhesins, toxins, protectins and iron capture systems	B2	STc131	O16:H5, O25:H4
					STc73	O2:H1, O6:H1
					STc95	O1/O2/O18/ O45:K1:H7, O2:K1:H4
					STc12	O4:H1/H5
					STc14	O75:H5
				D	STc69	O17/O73/O77:H18
				C	STc88	O8/O9:H4/H9/H19, O78:H4
				F	STc62	O7:K1:H45
UPEC	Isolated from urine	Human, domestic mammals	*papGII*, *papGIII*	B2	STc131, STc73, STc95, STc12, STc14	Idem ExPEC
				D	STc69	
NMEC	Isolated from cerebrospinal fluid of neonates	Human	Genes encoding the K1 antigen, pS88 genes	B2	STc95	Idem ExPEC
				F	STc59	O1:K1:H7
					STc62	O7:K1:H45
Pneumonia-associated *E. coli*	Isolated from lung	Human	*hly*, *sfa*	B2	STc73	Idem ExPEC
					STc127	O6:H31
					STc141	O2:K1:H6
APEC	Isolated from birds	Poultry	pColV genes	B2	STc95	Idem ExPEC
				C	STc88	O8/O78:H4/H9/H19
				G	STc117	O multiple:H4
InPEC	Diarrhoeal disease	Human, domestic mammals	Various	All phylogroups	Numerous	Numerous
STEC and/or EHEC	*stx* genes	Human, cattle[c], sheep[c]	*stx*, *eae*, *ehxA*	E	STc11	O157:H7
				B1	STc29	O26:H11/H$^-$
						O111:H8/H$^-$
					ST17	O45/O103:H2
EPEC	Attaching and effacing lesions on intestinal epithelial cells	Human, domestic mammals	*eae*, *bfp*	A	ST1788 (EPEC5)	Variable
					STc10 (EPEC10)	O variable:H40
				B1	STc3 (EPEC2)	O103/O111/O114/
						O126/O128:H2
					STc328 (EPEC7)	O88:H25
						O128/O153/O?:H7
				B2	STc15 (EPEC1)	O55/O127/O142:H6
					STc28 (EPEC4)	O85:H31, O33/O119:H6
					STc5342 (EPEC8)	O55/O76:H51
					STc2346 (EPEC9)	O33/O142:H34
				E	STc335	O55:H7
					STc32	O145:H28
ETEC	Heat-stable and heat-labile enterotoxins	Human, pig, cattle	Genes encoding enterotoxins and colonization factors	A, B1, C, E	Numerous	Numerous
EIEC	Colonocyte invasion	Strictly human	*ipa*, *isc*, *vir*	A	ST6	O124:H30
			Inactivation of *nadA*, *nadB* and *cadA*	B1	ST270	O164:H7
				E	ST280	O143:H26

Table 1 (cont.) | **Main characteristics of the more commonly used *Escherichia coli* pathotypes**

Pathotype[a]	Definition basis	Main strain host	Main virulence genes	Strain phylogenetic background	Main ST[WU]	Main serotypes[b]
EAEC	Aggregative adhesion on enterocytes	Human, domestic mammals	Aggregative adherence fimbriae (*aaf/agg*) and transcriptional (*aggR*) genes	A, B1, B2, D	Numerous	Numerous
DAEC[d]	Diffuse adhesion on enterocytes	Human	Genes encoding adhesins (*afa* and *dra*)	All phylogroups	Numerous	Numerous
AIEC	Adhesion and invasion of intestinal epithelial cells	Human	Unknown	All phylogroups with a majority of B2	ST135	O83:H1
					ST73, ST95, ST127, ST131	ExPEC serotypes
Hybrid InPEC	EHEC and EAEC characteristics	Human	*stx*, *aggABCD*, *aggR*	B1	ST678	O104:H4
Hybrid InPEC–ExPEC	HUS and septicaemia	Human, cattle[e]	*stx*, *eae*, pS88 ExPEC genes	A	ST301	O80:H2

AIEC, adherent-invasive *E. coli*; APEC, avian pathogenic *E. coli*; DAEC, diffusely adherent *E. coli*; EAEC, enteroaggregative *E. coli*; EHEC, enterohemorrhagic *E. coli*; EIEC, enteroinvasive *E. coli*; EPEC, enteropathogenic *E. coli*; ETEC, enterotoxigenic *E. coli*; ExPEC, extra-intestinal pathogenic *E. coli*; HUS, haemolytic and uraemic syndrome; InPEC, intestinal pathogenic *E. coli*; NMEC, newborn meningitis *E. coli*; STc, sequencing type complex; STEC, Shiga toxin-producing *E. coli*; ST[WU], sequence type according to the Warwick University scheme; UPEC, uropathogenic *E. coli*. [a]In addition to pathotypes, ExPEC strains from pneumonia are considered. [b]Only K1 types are indicated among K antigens. O?, O unknown. [c]Asymptomatic. [d]Encompasses also UPEC strains. [e]Diarrhoea only.

of virulence in a population genetic framework seems particularly timely. Such knowledge will help design preventive and therapeutic strategies to fight *E. coli* infections. In this Review, we place *E. coli* species within the genus *Escherichia* and present the phylogeny and global population structure of *E. coli*. We also provide an overview of the general principles of the emergence of virulence before more thoroughly describing the main ExPEC, InPEC and hybrid clones. *Shigella* species, which belong to the *E. coli* species[13], are not be discussed as there is a recent review devoted to the genomic signatures of *Shigella* evolution[14].

The population structure of *E. coli*

Escherichia *genus and* E. coli *species phylogeny*. The taxonomy of the genus *Escherichia* has recently changed with the description of five cryptic *Escherichia* clades[15] and the reassignment of *Escherichia blattae*[16], *Escherichia hermanii*[17] and *Escherichia vulneris*[18] to other genera. The genus *Escherichia* is now composed of three nomen species (that is, *Escherichia albertii*, *Escherichia fergusonii* and *E. coli*) and five *Escherichia* clades labelled I–V (see Supplementary information S1 for their pathogenicity). These clades are phenotypically undistinguishable from *E. coli* but divergent to various degrees at the nucleotide level from *E. coli*. Based on average nucleotide identity[19] (Supplementary information S2), it has been proposed that clade V and clades III and IV represent two new *Escherichia* species, clades III and IV being two subspecies[12]. The name *Escherichia marmotae* has been proposed for clade V[20], although the strains of this species are not limited to marmots. Clade I and *E. coli* can be considered two subspecies of a single species[12] and it has been proposed to name this new species *E. coli sensu lato*, whereas *E. coli sensu stricto* would refer to classic *E. coli* strains[21]. This classification is corroborated by the existence of genetic exchange of core genes between clade I and classic *E. coli* strains, but not between *E. coli sensu lato* and other members of the genus[22]. Very few clade II strains have been described, and these could represent a new species (FIG. 1; Supplementary information S3).

E. coli sensu stricto has a strong phylogenetic structure representing at least eight phylogenetic groups partitioned into two main clusters: phylogroups B2, G, F and D; and phylogroups A, B1, C and E (FIG. 1). An additional group, named H[23], which seems to be related to phylogroup D, can be distinguished.

Epidemiological studies have benefited from easy and rapid PCR-based methods that enable *E. coli* phylogroups[21,24,25], clades[26] and *Escherichia* species[27–30] to be determined. This approach has been adapted to type strains in silico from their complete genomes[30]. Also, EnteroBase, an integrated software environment and database currently containing more than 100,000 assembled genomes of *Escherichia* and *Shigella* strains with metadata, provides a unique opportunity to perform comprehensive epidemiological studies of the genus[31].

Order and disorder in* E. coli *genomes. The *E. coli* genome is composed of a circular chromosome and plasmids. The genome of *E. coli* strains (excluding *Shigella*) varies from 4.2 to 6.0 Mbp, which corresponds to 3,900–5,800 genes, respectively[32–34]. This variability is the result of frequent acquisitions and deletions of fragments of DNA during *E. coli* divergence. There is a large phylogenetic component to the observed size variation: strains from phylogroups A and B1 have the smallest genomes, whereas the largest genomes are observed in phylogroup E[34,35]. All *E. coli* strains share around 2,000 genes (core genome), with the balance of genes in a strain being drawn from the pan-genome. The pan-genome (that is, the total number of genes) varies with the number of genomes analysed: from 15,000 genes in

Clades
(Also known as lineages). Groups of organisms that consist of a common ancestor and all its lineal descendants. This term has been used at different phylogenetic levels, leading to some confusion. For the cryptic clades, it corresponds to species or subspecies, whereas within the *Escherichia coli* species it designates groups of organisms composing a sequence type.

Phylogroups
Groups of organisms that belong to a large phylogenetic entity within the species. There are at least eight phylogenetic groups within the *Escherichia coli* species, named A, B1, B2, C, D, E, F and G.

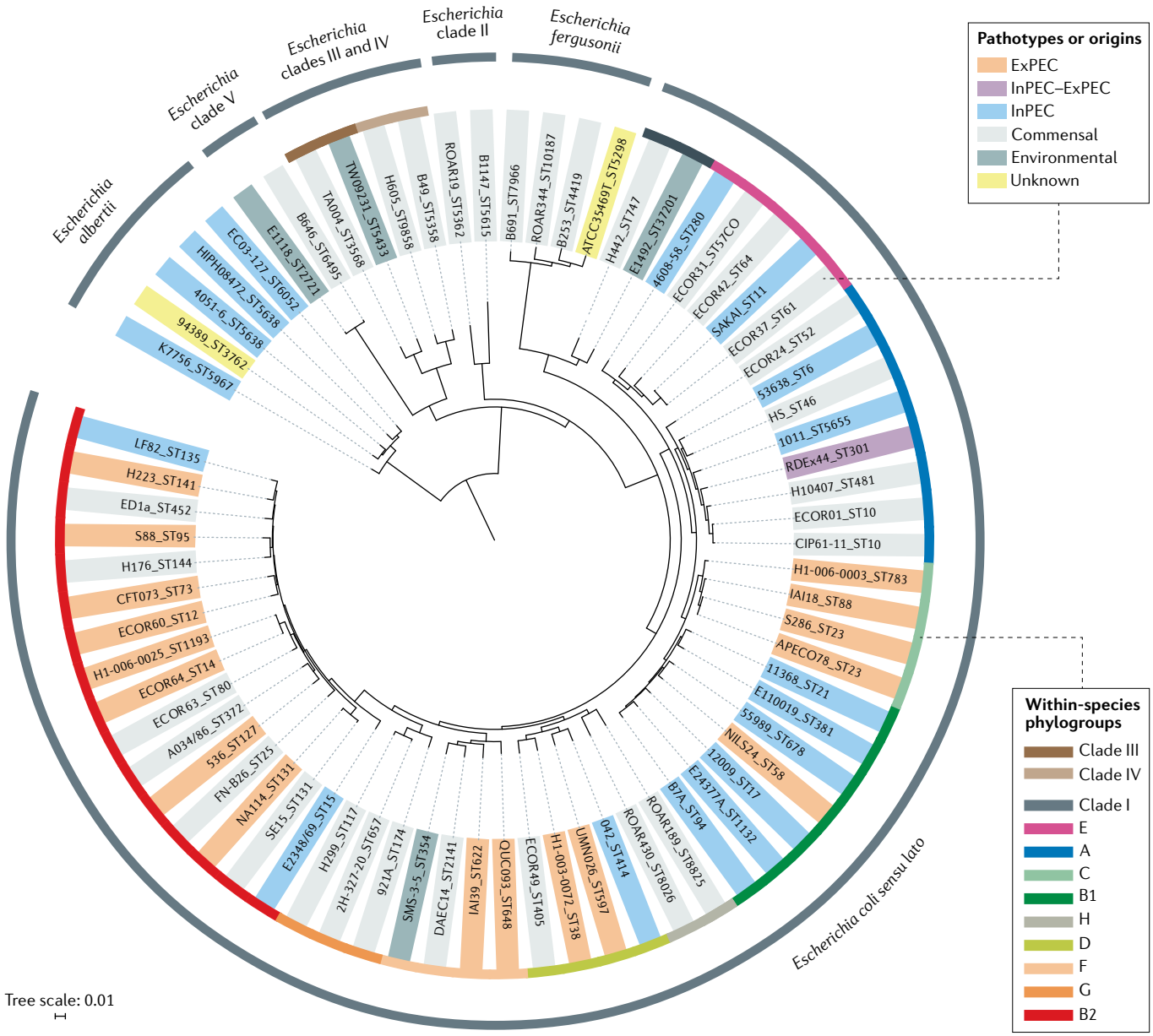


Fig. 1 | **Phylogenetic history of 72 *Escherichia* strains.** The tree was reconstructed from the SNPs ($n = 374,678$) of core genome genes ($n = 1,302$) using Roary[210] and RAxML[211] and rooted on the *Escherichia albertii* species. The strains have been chosen to represent the phylogenetic diversity of the genus. They are identified by their ID followed by the sequence type (ST) number according to the Warwick University scheme[58] and are coloured according to their pathotype and/or origin. Of note, ST597 belongs to STc69 (phylogroup D). The outer ring corresponds to the different species defined by the average nucleotide identity[19] (Supplementary information S2), whereas the inner ring represents the within-species phylogenetic groups. The five strains of *E. albertii* are representative of the five described phylogroups[28]. All of the internal nodes have bootstrap values of 100% (1000 bootstraps). Similar phylogeny was obtained using Harvest (27,564 SNPs) or when recombinant fragments were removed with Gubbins[212]. A table with the main strain characteristics is available in Supplementary information S3. ExPEC, extra-intestinal pathogenic *Escherichia coli*; InPEC, intestinal pathogenic *E. coli*.

20 genomes[32,36] to 75,000 genes in 1,500 genomes[34,37]. Twenty-six new genes are expected to be identified with each new strain sequenced[34].

A high level of homologous recombination (gene conversion) is also observed on the chromosome, which is at least as frequent as mutation, with an average length of fragments involved estimated to be between 50 bp and 2–4 kbp[32,38–41] and, rarely, large fragments over 100 kbp[41]. This wide range of reported recombined fragment lengths could be explained by the various effects of the restriction systems of the recipient cells reducing the length of the acquired DNA and the result of successive overlapping incorporations of large fragments leading to a mosaic of small segments over time[38,41]. Although scattered over the chromosome, recombination is less frequent at the terminus of replication and there are two hotspots of recombination located at the O-antigen biosynthesis gene cluster and

the *fim* operon[32,40], which correspond to the 'bastions of polymorphism' as previously described[42]. Interestingly, these recombination hotspots are also integration hotspots, which indicates that homologous recombination can be involved in the acquisition of large fragments of DNA[43]. However, recombination does not 'blur' the phylogenetic signal and a meaningful phylogeny can be reconstructed, probably because of the small length of recombined fragments[1,31,32] (FIG. 1).

Moreover, almost all *E. coli* isolates carry plasmids, typically two to four plasmids per strain[44–46]. The size of these plasmids varies with their mobility characteristics (conjugative plasmids being larger (up to 300 kbp) and mobilizable, and non-transferable plasmids being smallest (<30 kbp)) and their structure is highly mosaic[46–50]. The type of plasmid is species-specific[46,49,50], as exemplified by the frequent presence of incompatibility group IncF and IncI plasmids in *E. coli*, and plasmid content seems to vary within species according to the strain phylogeny[44,47].

Defining a clone in the genomic era. *E. coli* has a clonal population structure, meaning that there is strong non-random association of alleles (linkage disequilibrium) and frequent recovery of only a few of all the possible multilocus genotypes[51,52]. This is due, as stated above, to the rate and specific pattern of recombination (horizontal transfer) involving small fragments that do not break up the vertical evolution due to mutation[1]. This clonal structure was first observed using phenotypic serotype determination (somatic (O-serogroup), capsular (K) and flagellar (H) antigens)[53,54] and multilocus enzyme electrophoresis (MLEE)[55]. The development and widespread use of Sanger sequencing technology led to the replacement of these phenotypic tests by the sequencing of ~500-bp segments of seven to eight housekeeping genes, an approach called multilocus sequence typing (MLST) by analogy to MLEE[56]. WGS now enables the in silico typing of strains for various genotyping schemes. It can be used to perform classic typing such as O:H typing[57], MLST[58,59] or *fimH* allele typing based on minor sequence variations[60]. According to the level of relatedness of the isolates, core genome MLST or whole-genome MLST can also be performed[61]. In addition to allele typing, sequencing provides nucleotide sequences, and a phylogenetic tree can be constructed based on SNPs of the core genome that accurately infers the evolutionary history of the isolates.

It is generally agreed that a clone is composed of indistinguishable, or very closely related, isolates that are descended from a common ancestor. Currently, the most commonly used method to define strains is MLST, with the sequence type designation of an isolate considered to represent a clone[62,63]. However, the community of EHEC researchers remains attached to serotype designations[64]. FIGURE 2 provides three examples of well-known clones and/or sequence types (ST^{WU}131, ST^{WU}95 and ST^{WU}117 (sequence types according to the Warwick University scheme)) and shows that the definition of a clone is dependent on the population structure of the sequence type. For the ST^{WU}131 clone, the tree shows a stepwise evolution and the ST^{IP} (sequence types according to the Institute Pasteur scheme), the serotype and the *fimH* allele are congruent and define three major clades: clade A (ST^{IP}506_O16:H5_*fimH*41), clade B (ST^{IP}43_O25:H4_*fimH*22) and clade C (ST^{IP}43_O25:H4_*fimH*30) (FIG. 2a). The ST^{WU}95 clone shows a rapid diversification with short basal branches and the delineation of five major subgroups (A–E), with strains exhibiting six serotypes with various *fimH* alleles widespread within the subgroups (FIG. 2b). Last, the ST^{WU}117 strains show a large O-serogroup diversity (one per strain), but mostly identical ST^{IP} (mainly 48), H type (mainly 4) and *fimH* allele (97) (FIG. 2c). This pattern suggests a strong selective pressure for O-serogroup diversification occurring via recombination at the *rfb* linked to a conserved H type[25].

The level of core genomic divergence, estimated by the nucleotide diversity per site or the mean number of core genome SNPs between strains, is also dependent on the particular sequence type. As an example, ST131 is far more diverse than ST95 or ST117, with clade A and clades B and C each having levels of diversity comparable with ST95 and ST117 (FIG. 2). Strains of a single sequence type can substantially differ in terms of their gene repertoires, with hundreds of genes differing per strain pair[34].

In this Review, we use the widely accepted ST^{WU} and/or serotype designation of the clones, but we should keep in mind that these entities can correspond to lineages with a variable and heterogeneous level of divergence.

The emergence of virulence

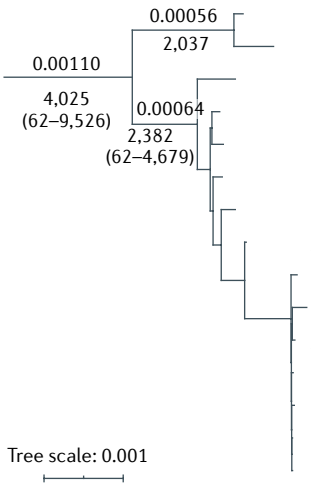
The evolution of virulence is based on three main mechanisms. First, the acquisition of a new gene or genes and/or a new function or functions by horizontal gene transfer mediated by mobile genetic elements, including plasmids, phages, and integrative and conjugating elements. The latter two can integrate chromosomal DNA and thus be replicated by the chromosome[65]. Pathogenicity islands (PAIs), which are large chromosomal genetic elements involved in virulence, are a subset of genomic islands acquired via horizontal gene transfer and frequently associated with tRNA genes that are possibly remnants of mobile genetic elements[66]. All of these acquired elements are characterized by their mosaic and modular structure that can be viewed as molecular building blocks, enabling multiple combinations that lead to multiple phenotypes (FIG. 3a). The second mechanism involved in the evolution of virulence is the inactivation of genes whose expression is incompatible with virulence (antivirulence genes)[67]. In this case, a gene whose expression was advantageous in a non-pathogenic setting is detrimental in the pathogenic setting, a trade-off known as antagonistic pleiotropy[68]. This phenomenon has been shown to especially occur in metabolic pathways[69,70]. The last mechanism involves point mutations that lead to a change of function[71]. Such patho-adaptative mutations have been particularly well described for the adhesive subunit of type I fimbriae, FimH[72]. A few amino-acid variants are responsible for a shift in the binding capacity of *E. coli* strains from digestive to urinary tract epithelial cell binding, which results

Serotype
A group of organisms that have the same association of O-polysaccharide antigen (serogroup), flagellar (H) antigen and capsular (K) antigen. There are currently 53 H types and 67 K antigens. However, as few laboratories had the capability to type the K antigens, serotypes based on O and H antigens became the gold standard.

Serogroup
A group of organisms that have the same surface O-polysaccharide antigen. There are currently ~186 different *Escherichia coli* O serogroups.

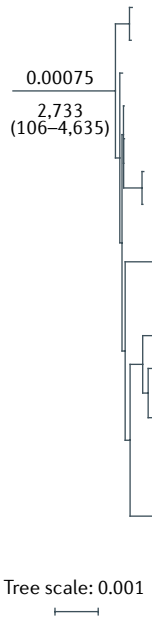
Sequence type
The allelic profile constituted by the alleles at each studied gene locus, usually seven. A group of organisms can be categorized according to the sequence type. Like multilocus enzyme electrophoresis, multilocus sequence typing uses the allele as the unit of comparison, rather than the nucleotide sequence. A sequence type complex (also known as a clonal group) is a simple or double-locus variant of a sequence type.

a



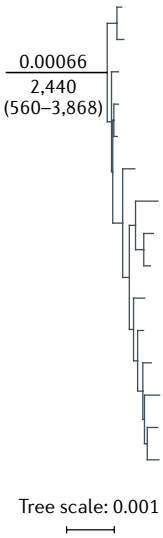
	ST[IP]	O	H	*fimH*	Clade	Fluoroquinolone	CTX-M
SE15	506	O16	H5	*fimH41*	A	S	
NILS53	506	O16	H5	*fimH41*	A	S	
CD347	43	O25	H4	*fimH22*	B3	S	
CD331	43	O25	H4	*fimH22*	B5	S	
CD413	9	O25	H4	*fimH22*	B2	S	
HVM52	9	O25	H4	*fimH22*	B1	S	
ZH063	43	O25	H4	*fimH22*	B4	S	
JJ1897	43	O25	H4	*fimH22*	B0	S	
CD306	43	O25	H4	*fimH30*	C0	S	
KP75	43	O25	H4	*fimH30*	C2	R	15
JJ1886	43	O25	H4	*fimH30*	C2	R	15
BRG151	43	O25	H4	*fimH30*	C1	R	
EcSA01	43	O25	H4	*fimH30*	C1	R	14
KUN5781	43	O25	H4	*fimH30*	C1-M27	R	27
KUN8768	43	O25	H4	*fimH30*	C1-M27	R	27

b



	ST[IP]	O	H	*fimH*	Subgroup
B1J4	1	O18	H7	*fimH15*	B
UTI89	1	O18	H7	*fimH18*	B
APECO1	418	O1	H7	*fimH15*	C
H252	1	O1	H7	*fimH27*	C
BS14	1	O1	H7	*fimH27*	C
004-008	1	O25	H4	*fimH27*	C
BS11	1	O25	H4	*fimH27*	C
H062	1	O2	H4	*fimH27*	E
BS84	1	O2	H4	*fimH27*	E
032-001	1	O2	H4	*fimH27*	E
S88	1	O45	H7	*fimH54*	D
032-003	1	O1	H7	*fimH30*	D
H434	1	O2	H4	*fimH30*	D
025-007	1	O2	H4	*fimH30*	D
H112	1	O1	H7	*fimH41*	A
H397	1	O1	H7	*fimH41*	A
B6H13	1	O2	H7	*fimH41*	A
H263	1	O2	H7	*fimH41*	A

c



	ST[IP]	O	H	*fimH*	CTX-M
HT2012095	48	O183	H18	*fimH97*	
C66	48	O85	H18	*fimH97*	
B1932b	48	O119	H4	*fimH97*	
H561	48	O161	H4	*fimH97*	
ERR435084	48	O159	H4	*fimH97*	
H1-001-0053	48	O33	H4	*fimH97*	
C4 95	48	O45	H18	*fimH97*	
C3 65	539	O11	H4	*fimH97*	
E11701-VNCW	539	O8	H4	*fimH97*	
EFTR 8 7	641	O131	H4	*fimH97*	14
H1-001-0155	48	O114	H4	*fimH97*	
H299	48	O24	H4	*fimH97*	
2H 434 21	48	O111	H4	*fimH97*	
H1-003-0054	48	O143	H4	*fimH97*	1
C2 106	48	O71	H4	*fimH97*	

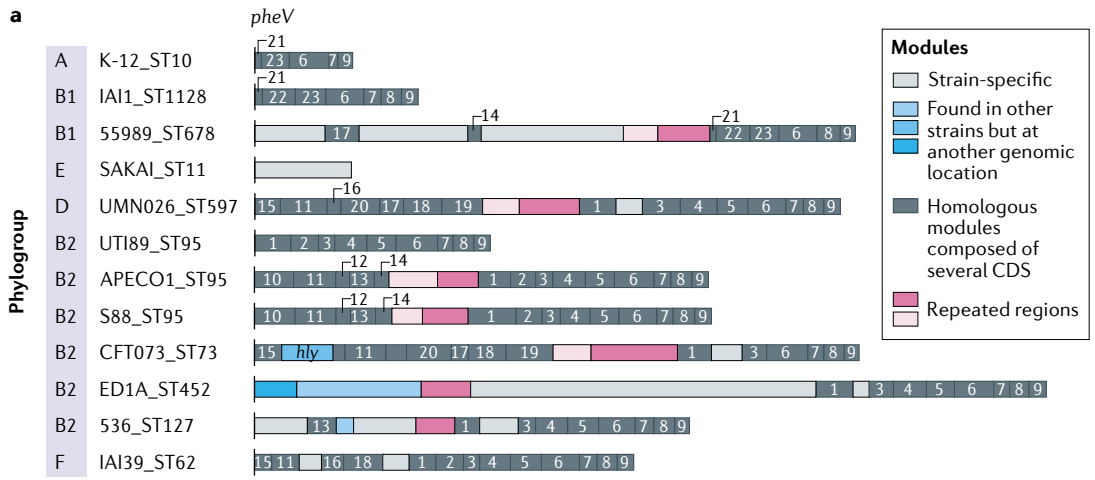
Fig. 2 | **Example phylogenetic history of *Escherichia coli* strains of three main sequence types according to the Warwick University (WU) scheme.** The phylogenetic histories of ST131[WU] (phylogroup B2) (part **a**), ST95[WU] (phylogroup B2) (part **b**) and ST117[WU] (phylogroup G) (part **c**) are shown. The phylogeny was reconstructed from the SNPs of core genome genes using Roary[210] and RAxML[211]. The trees are rooted on *Escherichia coli* ED1a that belongs to phylogroup B2. The core genomes are composed of 3,412 (part **a**), 3,268 (part **b**) and 3,344 (part **c**) genes with 328,470 (part **a**), 467,635 (part **b**) and 205,876 (part **c**) SNPs. The sequence types according to the Institute Pasteur scheme (ST[IP]), serotype (O and H), *fimH* allele, clade or subgroup, fluoroquinolone resistance (R) or susceptibility (S) and the presence and type of CTX-M extended-spectrum β-lactamases are indicated. The nucleotide diversity per site, calculated using the diversity.stats function from the PopGenome R package[213], and the mean (minimum–maximum) numbers of SNPs are indicated above and below the main nodes, respectively.

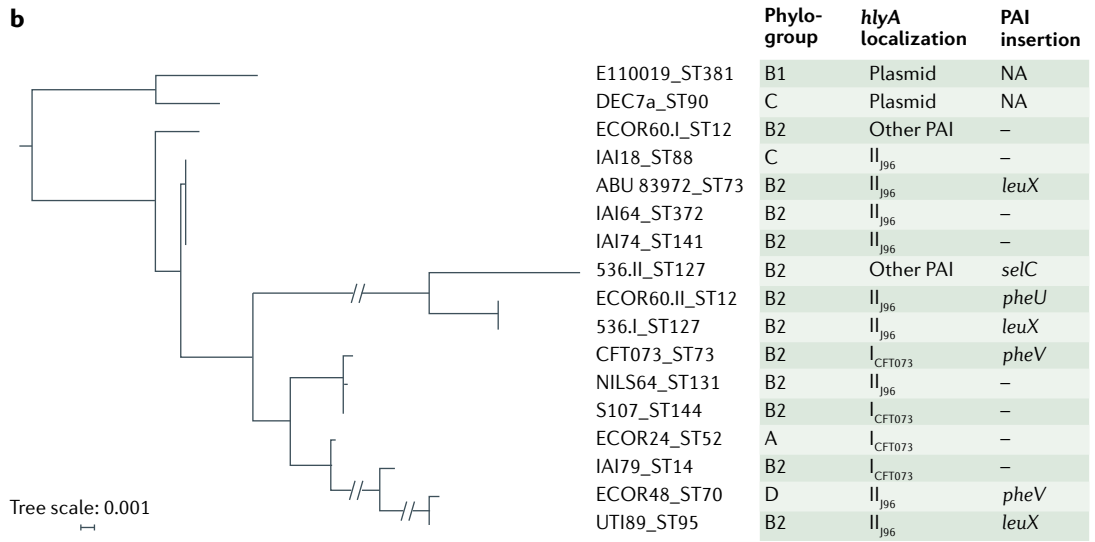
in a functional trade-off as the increase in urovirulence is detrimental to intestinal colonization.

When the evolutionary history of these molecular events is mapped to the phylogenetic history of the strains, two important features emerge. First, each event occurred several times during strain evolution. This convergent evolution is a strong sign of selection[73]. The repeated acquisition of specific virulence genes by plasmids and phages in different lineages was first observed for EHEC and EPEC[74–76] and later extended to the other pathotypes[32,77]. An example of several acquisitions in distinct genomic locations is shown in FIG. 3b for the *hlyA* gene, which encodes a toxin. Similarly, different gene rearrangements (insertion sequences, phages or various deletions) have been observed in the lysine decarboxylase gene (*cadA*) region[78] and the A27V

mutation occurred in several FimH backgrounds[79]. Second, there is a major role of the genetic background in the emergence of the virulence. It has been long recognized that strains from phylogroups B2 and D are frequently isolated from extra-intestinal infections, possess numerous virulence genes[80] and are virulent in a mouse model of sepsis. By contrast, most phylogroup A, B1 and E strains are non-virulent in this model[81–83]. EHEC strains belong mainly to phylogroups E and B1 (REFS[75,84,85]). These specific associations between genetic background, virulence genes and strain phenotype indicate the presence of complex genetic interactions between loci, termed intergenic epistasis[86].

In summary, it can be considered that virulence is the result of the succession of several genetic events[87]. Multiple combinations of such events can lead to virulence but in a specific genetic background[88] and probably in a precise order[89], due to epistatic interactions. Such scenarios are described in more detail in the following sections applied to specific pathogenic clones. Nevertheless, a fundamental question remains: why there are commensals and pathogens within a single species (BOX 1).





	Phylo-group	*hlyA* localization	PAI insertion
E110019_ST381	B1	Plasmid	NA
DEC7a_ST90	C	Plasmid	NA
ECOR60.I_ST12	B2	Other PAI	–
IAI18_ST88	C	II_{J96}	–
ABU 83972_ST73	B2	II_{J96}	*leuX*
IAI64_ST372	B2	II_{J96}	–
IAI74_ST141	B2	II_{J96}	–
536.II_ST127	B2	Other PAI	*selC*
ECOR60.II_ST12	B2	II_{J96}	*pheU*
536.I_ST127	B2	II_{J96}	*leuX*
CFT073_ST73	B2	I_{CFT073}	*pheV*
NILS64_ST131	B2	II_{J96}	–
S107_ST144	B2	I_{CFT073}	–
ECOR24_ST52	A	I_{CFT073}	–
IAI79_ST14	B2	I_{CFT073}	–
ECOR48_ST70	D	II_{J96}	*pheV*
UTI89_ST95	B2	II_{J96}	*leuX*

Fig. 3 | **Modularity and mobility of acquired genomic elements. a** | Schematic representation of the genomic island at the tRNA–PheV insertion hotspot in different *Escherichia coli* strains. The position of the *pheV* gene is indicated by a vertical bar on the left side of the genomic islands. Twenty-three homologous modules composed of several coding sequences (CDS) (mean = 6.7, minimum–maximum = 2–14), often grouped in an operon, have been identified and are represented in different colours. Modules in blue are found in other strains but at another genomic location. One of these modules in CFT073 possesses the *hly* operon and the corresponding genomic island is the pathogenicity island (PAI) I_{CFT073} (see part **b**). **b** | Phylogenetic history of the *hlyA* gene (3,075 nucleotides), which is part of the *hly* operon, in strains of various phylogroups and its localization. The maximum likelihood tree was reconstructed using PHYML[214] and rooted on the two sequences located on a plasmid. The PAIs are indicated by Roman numbers followed by the name of the strain in which they have been described for the first time. Notably, the phylogeny of *hlyA* is not congruent with the strain phylogeny (closely related *hlyA* sequences belong to distinct phylogroup strains) and closely related *hlyA* sequences can be located in distinct PAIs or in the same PAI but at different positions. All of these data indicate multiple gain events of the *hlyA* gene in the strains. The strains are identified as in FIG. 1. NA, not applicable; –, data not available. Part **a** adapted from REF.[32], CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

Box 1 | **Why did *Escherichia coli* pathogens emerge?**

In nature, *Escherichia coli* populations exist in the gut of vertebrates as well as in the environment (water and sediments)[215]. How can virulence be selected in this context? The intestine is the reservoir of extra-intestinal pathogenic *E. coli* (ExPEC) strains, and numerous intestinal pathogenic *E. coli* (InPEC) pathotypes are found as commensals in the gut. It has been proposed that ExPEC strains have evolved for enhanced intestinal colonization and persistence, not to cause extra-intestinal infection, which represents a dead end[91]. Similarly, despite the potential of diarrhoea favouring strain transmission, InPEC strains could also have evolved to enhance persistence in the gut. This 'coincidental evolution' hypothesis — that is, virulence determinants that have evolved for other functions[216] — has recently gained some empirical support.

Phylogroup B2 ExPEC lineages are better at persisting in the gut microbiota of infants[217,218] and piglets[219] compared with strains of the other *E. coli* phylogenetic groups. This is, in part, due to the accumulation of gene-encoding adhesins, protectins, toxins and iron capture systems found in ExPEC pathogenicity islands (PAIs)[218–220]. Additionally, mouse colonization experiments showed that phylogroup B2 ExPEC archetypal strains (F11, 536) efficiently colonize the mouse gut, but not a phylogroup A strain or a 536 strain deleted of its PAIs[221,222]. The molecular mechanisms are complex as F11 *hlyA* and *papG* single mutants have only a transient colonization defect, and deletion of all PAIs of strain 536 is necessary for it to be consistently outcompeted by the wild-type strain[222,223]. Similarly, an O157:H7 *E. coli* strain was more efficient at colonizing the bovine terminal rectal mucosa than a phylogroup A *E. coli* strain or an O157:H7 *eae* mutant[224]. In calves, Shiga-toxin 2a (Stx2a) enhanced O157:H7 animal-to-animal transmission by restricting regeneration and turnover of the colonized gastrointestinal epithelium[225]. Finally, phylogroup B2 ExPEC strains were shown to be resistant to amoeba grazing, with the high-pathogenicity island having a major role in resistance against predation[226]. The role of Stx in protecting O157:H7 from protozoa grazing is still debated[227,228].

The ExPEC clones

The basic characteristics of extra-intestinal pathogenic strains were described by Kauffman in 1947: 75% of ExPEC strains represent just a few O serogroups, whereas faecal isolates exhibited much greater O-serogroup diversity, and these extra-intestinal pathogenic strains exhibit specific virulence phenotypes, such as the haemolytic and the skin necrotizing factors, and a very low lethal dose 50 in mice[90]. MLEE, ribotype or PCR typing[81], MLST[91] and, more recently, WGS[83] studies have confirmed the association of specific clones with numerous extra-intestinal virulence genes, encoding adhesins, toxins, protectins and iron capture systems[3], and virulence in a mouse sepsis model.

Epidemiological studies of thousands of human bloodstream isolates from various continents (the United States, the United Kingdom, France and South Korea) have demonstrated that five sequence types or sequencing type complexes (STcs) represent 48–66% of the isolates[37,92–95]. Among these five STcs, four were always observed (STc131, STc73 and STc95 belonging to phylogroup B2, and STc69 belonging to phylogroup D). The fifth was either STc12 or STc14 (including ST1193), both members of phylogroup B2. The clonal group A (CGA), which corresponds to STc69 that is resistant to trimethoprim–sulfamethoxazole[96] and the extended-spectrum β-lactamase (ESBL)-producing ST131 O25:H4 that is resistant to third-generation cephalosporins[97], was uncommon prior to the turn of the century.

The big four ExPEC clones. The most prevalent ExPEC clone isolated in pathogenic conditions is now ST131 (REF.[98]). This sequence type is in fact composed of three clades[99–103] (FIG. 2a). The most basal is clade A, corresponding to strains with an O16:H5 serotype and the *fimH*41 allele. Clade B then emerged, followed by clade C; all strains in both clades exhibit the O25:H4 serotype but have the *fimH*22 and *fimH*30 alleles, respectively. Within clade C, mutations in *parC* and *gyrA* lead to fluoroquinolone resistance, giving rise to a lineage denoted as C1/*H*30R that frequently encodes CTX-M ESBLs (mostly CTX-M-14), a C1-M-27 lineage that produces CTX-M-27 and the C2/*H*30Rx lineage that encodes CTX-M-15. Molecular clock-based dating estimates the emergence of clade B at around 1950 and the emergence of fluoroquinolone resistance in clade C at around 1987 (REFS[102,103]).

The core genome of the STc131 is around 3,000 genes[100,104,105] with a pan-genome of more than 26,000 genes for 4,000 strains[105], indicative of a highly variable gene pool. A scenario of stepwise evolution has been proposed consisting of the acquisition and/or loss of several genomic islands, prophages and plasmids; recombination events at the *rfb*, *fli* and *fim* loci; and point mutations. Based on single-molecule, real-time sequencing data[106], a scenario of losses and gains of complete plasmids and of specific plasmidic genes has been proposed. Moreover, restricted plasmid–clade associations were evidenced, suggesting strong plasmid–clade adaptations[107].

Numerous epidemiological studies have shown that ST131 mainly colonizes humans or human-associated animals[98]. Indeed, in addition to humans, ST131 strains are now frequently isolated in companion animals such as dogs and cats[108,109], and transmission of a single ST131 strain among human and pets within households has been reported[110]. Poultry meat seems to be a reservoir for ST131 clade B (*fimH*22) strains[111]. ST131 strains have been found in faecal samples of gulls in areas of dense human populations where these birds feed on leftover human food and garbage[112].

The reasons for the success of the ST131 clonal complex are not well understood and several, non-exclusive, hypotheses have been proposed. First, the strains of this sequence type are often found to be resistant to multiple drugs highly prescribed in humans and/or animals (quinolones, third-generation cephalosporins, carbapenems and colistin) and no cost of resistance plasmid carriage has been observed, possibly as a result of epistatic interactions[106]. Second, the ST131 strains have a good in vitro fitness in various media, including human urine[113], can form biofilms with clade-specific kinetics[114], efficiently colonize the mammalian gut and persist long term[113,115], and can be virulent in a mouse model of sepsis[116]. Despite their recent emergence, clade C strains exhibit extensive allelic diversity at loci involved in colonization (protectins, iron capture systems and adhesins) and in anaerobic metabolism. This diversity could reflect selection in situations in which a phenotype is most beneficial to a population when it is rare, such as new antigen or resource-based strategy (negative frequency-dependent selection)[117].

STc95 and STc73 are the most prevalent ExPEC clonal complexes after ST131. One of the characteristics of these two STcs is that they are mostly devoid of antibiotic resistance[118,119], although multiple drug-resistant STc73 strains have been recently reported[120]. Both STcs

exhibit similar phylogenetic history, and STc95 and STc73 are delineated in five (A–E)[121] and four (a–d) subgroups[120], respectively, which have rapidly diverged (FIG. 2b). Each of these subgroups exhibits specific serotype–*fimH* allele combinations[120,121]. The core genome of the STc95 is around 3,000 genes with a pan-genome of 17,000 genes for 200–300 strains, numbers similar to ST131 (REFS[121,122]).

In addition to classic ExPEC pathologies, STc95 represents over half of newborn meningitis *E. coli* isolates[6,123]. It is also the predominant sequence type causing APEC[124,125]. Strains of both STc95 and STc73 are found in companion animals[126], and have been shown to be shared between canine and human members of a household suffering from UTIs[127].

Interestingly, a certain level of specialization is observed at the subgroup level within ST95. Subgroup A encompasses only human strains (O1/O2:H7_*fimH*41), whereas other subgroups encompass both avian and human strains[122]. NBM strains belong to subgroups A and B (O18:H7_*fimH*18/15) and subgroup D (O45:H7_*fimH*54 and O1/O2:H7/H4_*fimH*30) with a local epidemiology, the O18:H7_*fimH*18 and O45:H7_*fimH*54 strains being mainly isolated in the United States and Europe, respectively[128,129]. In addition, O1:H7 subgroup A strains are largely pan-susceptible[121,130]. These strains possess two specific chromosomal regions that encode a restriction-modification system and a DNA-cytosine methyltransferase, which could preclude the acquisition of mobile genetic resistance elements[130].

The fourth ExPEC in term of prevalence is STc69. Also called CGA although it belongs to phylogroup D, ST69 first received attention during the late 1990s as a predominant cause of trimethoprim–sulfamethoxazole-resistant UTIs across the United States[96]. It is now a clonal lineage responsible for extra-intestinal infections in humans worldwide with a stable prevalence[131]. Although the clonal lineage is primarily trimethoprim–sulfamethoxazole-resistant, ESBL producers are increasingly emerging[119]. It can also be found in cattle, pigs and poultry[132] and companion animals[126]. The strains exhibit several O serogroups (O11, O15, O17/73/77 and O117) associated mainly with H18 and *fimH*27 (REFS[37,133]).

Other clonal ExPEC groups. Three other B2 clonal groups are often retrieved as ExPEC: STc12, STc14 and STc127. STc12, mainly O4:H1/H5 serotypes, and STc14, mainly O75:H5 serotype, are increasingly reported to be ESBL producers[119,134]. Among the STc14 strains, at least three clades can be delineated[134], corresponding to the anciently diverged ST550 and ST14_*fimH*27 clades and the recently emerging multidrug-resistant ST1193_K1_*fimH*64 clade[135]. The fluoroquinolone resistance of this clone has been acquired by an unusual 1-step mechanism involving 11 simultaneous homologous recombination events[136]. Using time-scaled phylogenetic analysis, it has been estimated that the current ST1193 clade first emerged 25 years ago[134]. By contrast, STc127 is a very clonal group exhibiting the unique O6:H31 serotype and encompasses mostly antibiotic sensitive strains. Interestingly, STc127 strains are over-represented among human pneumonia isolates[137].

The phylogroup C lineage STc88 is one of the main APEC group of strains[124,138] (TABLE 1) and is represented by the archetypal APEC O78 (ST23_O78:H9_*fimH*35)[139] and 789 (ST88_O78:H19_*fimH*27) strains. It also encompasses NBM strains[123], of which strain S286 (ST23_O78:H4_*fimH*35) is the best-known example[140]. All of these strains host a ColV plasmid with numerous virulence genes that is closely related to a plasmid found in ST95 strains involved in NBM and avian colisepticaemia[141,142]. STc88 includes ST410 strains (mainly O8:H9_*fimH*24) that have recently emerged as a cosmopolitan multiresistant lineage through a process of stepwise evolution similar to that inferred for ST131: around 1987, fluoroquinolone mutational resistance appeared, followed by the addition CTX-M-15 ESBL in 2003, which was followed by the acquisition of a IncX3 plasmid bearing the $bla_{OXA-181}$ carbapenemase gene, and in 2014 a second carbapenemase gene, bla_{NDM-5} on a IncFII plasmid[143].

In conclusion, the virulence of ExPEC clones is multigenic, as it involves numerous genes with weak effects[144], and emerges mainly in B2, D and, to a lesser extent, C and F phylogenetic backgrounds (FIG. 4a). ExPEC has been largely described in human and human-associated animals such as poultry, livestock or pets[138], but rarely in the faeces of wild animals[145–147]. A clear specialization of the strains can be observed at different levels: the host, the organ within a single host and the degree of antibiotic resistance (TABLE 1). The molecular mechanisms of this specialization remain to be determined.

The InPEC clones

STEC and EHEC. STEC is defined by the presence of genes encoding a Shiga toxin (Stx), which are borne by a prophage. Not all STEC strains are pathogenic, and EHEC strains are STEC causing infection in humans, usually characterized by bloody diarrhoea. Based on protein sequence similarity, two types of Stx have been described, Stx1 and Stx2, with Stx2 toxins further divided into subtypes Stx2a–Stx2g (REF.[148]). EHEC strains produce one or more Stx subtypes. Stx1 is closely related to the Stx produced by *Shigella dysenteriae* serotype 1, whereas the origin of Stx2 remains unknown. Stx2 is involved in the vast majority of HUS. Besides the genes encoding Stx, most EHEC strains harbour the locus of enterocyte effacement (LEE), a PAI shared with EPEC (see below). EHEC strains are also characterized by the presence of the plasmid-borne enterohaemolysin gene *ehxA*[2,3]. Therefore, typical EHEC strains illustrate the major modes of virulence gene acquisition via lysogenic phage, plasmid and PAI acquisitions.

The first EHEC haemorrhagic colitis outbreak occurred in the early 1980s in the USA, following the consumption of undercooked beef, and was caused by a strain with the uncommon serotype O157:H7 (REF.[149]). O157:H7 EHEC is distributed worldwide and still represents the major virulent EHEC clonal group. Cattle and sheep are thought to constitute the main reservoir. O157:H7 STEC and EHEC belong to phylogroup E and

STc11 (FIG. 4b). *S. dysenteriae* serotype 1, which produces a Stx encoded by genes belonging to a defective phage, is also a member of phylogroup E. The common phylogenetic origins of these two pathogens clearly indicate that a certain genetic background is necessary or favours the acquisition and/or expression of Stx. Classic MLEE and MLST analyses refined by WGS data revealed the stepwise series of gain and loss events leading to this

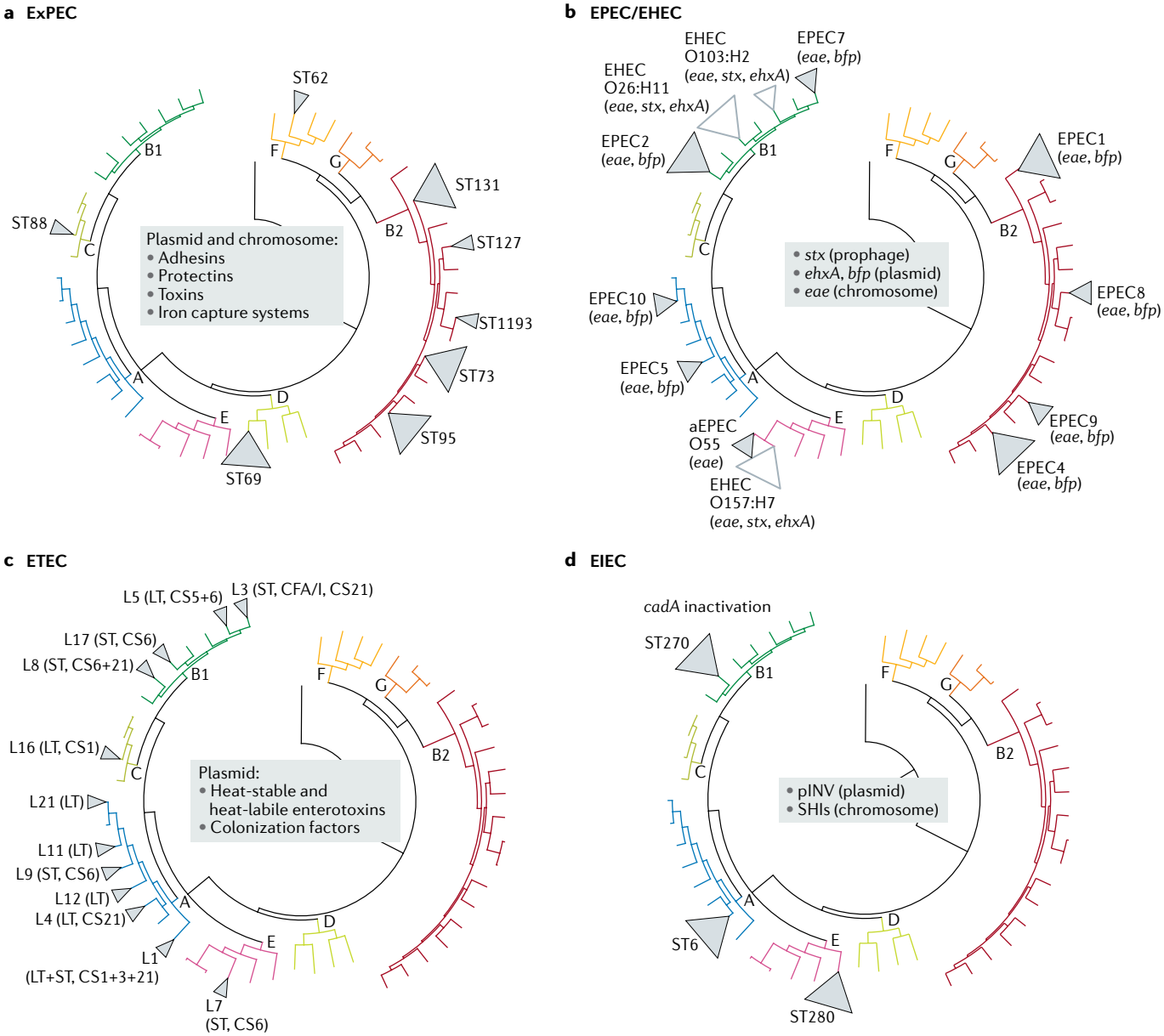


Fig. 4 | **Schematic representation of various evolutionary scenarios involved in the emergence of virulent lineages within *Escherichia coli* species.** The trees are reconstructed from the SNPs of core genome genes using Roary[210] and RAxML[211] from the *Escherichia coli sensu stricto* strains presented in FIG. 1, excluding those of phylogroup H, and rooted on an *Escherichia* clade I strain. Successful lineages are represented within the *E. coli* phylogeny by triangles with a size that is roughly proportional to their prevalence. The virulence determinants are indicated with their genetic locations (chromosomal including pathogenicity island, plasmid or prophage), except for gene inactivation. See TABLE 1 for more details. **a** | Emergence of virulence of extra-intestinal pathogenic *E. coli* (ExPEC) is a multigenic process implicating numerous genes involved mainly in four basic functions. The big four within phylogroups B2 and D and a few others are shown. Various combinations of virulence determinants are present in each lineage. **b** | Enteropathogenic *E. coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC) are represented by the filled and open triangles, respectively. A paucigenic process involving few genes, some with major effects, is at work. EPEC strains, with 10 major lineages[172] (eight represented), are scattered outside phylogroups D, F and G. For EHEC, the big two lineages that belong to phylogroups E and B1 and an example of another lineage are shown. **c** | In enterotoxigenic *E. coli* (ETEC), the virulence is also paucigenic and a high diversity of lineages[176] within phylogroups A, B1, C and E is observed. Only a part of the lineages is presented. Heat-stable (ST) and heat-labile (LT) enterotoxins and diverse colonization factors (CFA and CS) are shown. **d** | Enteroinvasive *E. coli* (EIEC) virulence is paucigenic and has the particularity to involve gene inactivation. The three main lineages[183] within phylogroups A, B1 and E are represented. The virulence plasmid and various combinations of genomic islands are present in all the lineages. aEPEC, atypical enteropathogenic *E. coli*; pINV, virulence plasmid; SHI, genomic island.

lineage becoming pathogenic. It has been shown that O157:H7 STEC is derived from an atypical O55:H7 EPEC ancestor that acquired the *stx2* gene, recombined at the *rfb* locus leading to O-serogroup conversion, and was followed by the loss of β-glucuronidase activity and sorbitol fermentation[150–152]. The pan-genome and core genome of almost 200 O157:H7 EHEC strains are around 14,500 and 4,300 genes, respectively[153]. Phylogeographic analysis using WGS data has been recently performed using 757 O157:H7 isolates from animals and humans from four continents[154]. The common ancestor of this set of isolates originated in the Netherlands around 1890 and then spread to different continents. The analysis of the WGS data showed that O157:H7 strains are organized in seven clades labelled A–G, a result congruent with previous classifications[155–157]. The clades have a non-random geographical distribution. In each country, the population structure of O157:H7 varies between cattle and humans, which indicates that some isolates from cattle are more likely than others to cause infection in humans[158]. Indeed, a machine-learning approach on WGS of 185 O157:H7 EHEC strains from humans and cattle in the United Kingdom found that only a subset of cattle strains (10%) may cause human disease, despite the fact that the phylogenetic analysis shows that cattle and human strains are intermingled in the tree. The major differences between human and bovine *E. coli* O157:H7 isolates were due to the relative abundance of hundreds of predicted prophage proteins[153]. Furthermore, virulence characteristics in patients as well as transmissibility to humans from bovine sources were linked to specific clades and Stx types or subtypes, clade F and the production of the Stx2a variant being associated with severe disease[152,155–157,159]. These epidemiological data are corroborated by in vitro and in vivo experiments using purified toxins that showed a high potency for Stx2a (REF.[160]). The arrival of stx_{2a}-carrying phages occurred several times at various chromosomal integration sites relatively recently (35 years ago), compared with the other subtypes carrying phages[152,154]. Interestingly, a specific phage confers the highest Stx2a production in the clade associated with severe disease[161], showing the complex interplay between the virulence determinant, its vector and the chromosome.

Besides O157:H7 STEC, O26:H11/H$^-$ STEC constitutes the second major public health concern in most industrialized countries. Large-scale genomic analysis of O26 strains revealed that they belong to STc29 of phylogroup B1 (FIG. 4b), with ST21 and ST29 representing most strains[162]. ST21 and ST29 are not exclusively composed of STEC strains as aEPEC strains are also present, suggesting an ongoing microevolutionary scenario in which *stx* is transferred between aEPEC and STEC[163]. Until the 1990s, O26 STEC strains isolated from humans produced only Stx1. In Europe, a genotype shift occurred during the 1990s with the emergence of O26 strains harbouring only *stx2a*, designated as the 'new European clone'. These strains were ST29 and highly associated with HUS[162]. More recent studies have revealed the continuously evolving genome of O26 STEC, and whole-genome SNP analysis splits both sequence types into several lineages[76,164]. O26 isolates display, as do O157 strains, a highly diverse mobilome with a large number of prophages and plasmids, and the genomic heterogeneity of this mobilome is a major contributor to O26:H11 intra-serotype diversity. The high level of diversification of O26:H11/H$^-$ *E. coli* has been very recently illustrated by the emergence of a novel clonal lineage of O26 EHEC within the new European clone in the 2010s (REF.[165]).

Since the first description of O157 and O26 STEC, a diversity of serogroups has been described. The predominant ones are O45, O91, O103, O111, O121 and O145, with each usually associated with a specific H type, and the relative abundance of the serogroups varies with the country of origin[166–169]. All of these STEC lineages belong to phylogroup B1, except O145 (phylogroup E). Half of those lineages belong to unique sequence types, such as ST33 (O91:H14/H$^-$), ST655 (O121:H19/H$^-$) and ST32 (O145:H28/H$^-$)[168]. O45:H2 and O103:H2 isolates belong to ST17, whereas O111:H8/H$^-$ isolates belong to ST16, which is part of the STc29 shared by O26 EHEC[163].

EPEC, ETEC and EIEC. EPEC strains are characterized in part by their ability to induce attaching–effacing lesions in the intestine due to the presence of the LEE genomic island, which encompasses the intimin-encoding *eae* gene. The presence or absence of a plasmid that possesses the *bfp* gene encoding type IV-like bundle-forming pili defines the EPEC as typical (tEPEC) or atypical (aEPEC), respectively[3]. These pathogenic elements are highly diverse. The LEE can be divided into three lineages based on average nucleotide divergence, with each lineage having a preferential tRNA insertion site (*pheU*, *pheU* and/or *pheV* and *selC*, respectively). Among the three lineages, 30 subtypes have been defined, compatible with but providing greater resolution than classic subtyping analyses on individual genes (*eae*, *tir* and *esp*)[170]. A substantial *bfp*-encoding F plasmid diversity is also observed[171]. The genomic background of the strains is highly diverse, with at least 10 lineages (EPEC 1–10) widespread in phylogroups A, B1, E and B2 (FIG. 4b) and exhibiting various serotypes, with each EPEC lineage having mostly a unique H type in combination with several O serogroups (TABLE 1). This serotype pattern is reminiscent of that observed in ST117 (FIG. 2c). tEPEC and aEPEC are found mixed within the lineages[170–172], but additional diversity is observed for lineages encompassing only aEPEC[170]. These EPEC lineages are widely distributed in humans. EPEC, and especially aEPEC, strains are frequently isolated from wild and domesticated mammals, with examples of animal isolates sharing sequence types, serotypes, pulsotypes and *eae* subtypes with human isolates[173].

The global transcriptomes of tEPEC isolates from diverse phylogenomic lineages under virulence-inducing conditions were shown to be highly variable with lineage and isolate-specific differences[174]. Interestingly, crosstalk between the chromosome and the *bfp*-bearing plasmid has been reported, the presence and type of this plasmid influencing the expression of hundreds

of chromosomal genes, including LEE, lipopolysaccharide biosynthesis and iron capture genes[171].

Altogether, the recent phylogenomic analyses, in agreement with earlier MLST data[74,175], show incongruent evolutionary histories of pathogenic elements and core genomes, indicating that stable EPEC lineages emerged from all of the phylogenetic diversity of *E. coli* via repeated acquisitions of LEE variants and/or the *bfp* operon on multiple ancestral plasmids. Various combinations of chromosomal and plasmidic elements influence gene expression.

Enterotoxigenic *E. coli* (ETEC) produce heat-stable enterotoxin including two subtypes (STh and STp) and/or heat-labile enterotoxin (LT), and at least 25 different colonization factors, fimbrial or afimbrial surface structures that enable adherence to intestinal epithelium[3]. These structures are mainly plasmid-encoded. WGS of a large representative collection of 362 human ETEC strains identified 22 robust lineages (L1–L22) belonging to phylogroups A (12 lineages), B1 (8 lineages), C (1 lineage) and E (1 lineage)[176] (FIG. 4c). These lineages exhibit various O serogroups. Interestingly, there is a specific association between the toxin profile, the colonization factors, the plasmid content and the O serogroup within a phylogenetic lineage or sub-lineage[176]. Allelic variants of the heat-labile enterotoxin have also been reported, correlating with specific lineages, with some variants being expressed more than others, suggesting greater virulence potential[177]. These lineages are cosmopolitan, and molecular clock dating of five lineages estimate their emergence as between 50 and 170 years ago[176]. Further phylogenomic studies on more than 200 additional ETEC strains from Chile[178] and Bangladesh[179] confirmed the remarkable sequence type and serotype diversity of the strains within phylogroups A and B1. Similarly, phylogenetic analyses of porcine ETEC strains exhibiting the K88 and F18 adhesins showed that they belong to phylogroups A, B1, C and E[138,180,181]. However, it seems that these lineages are different from the human ones[180].

In sum, the emergence of numerous stable, globally distributed, ETEC lineages among specific phylogenetic groups (A, B1, C and E, but never D, F, G and B2) argues for chromosome and plasmid combinations that optimize fitness and transmissibility, with a certain level of host specificity[176,180].

Enteroinvasive *E. coli* (EIEC) has the ability to invade intestinal epithelial cells owing to attributes closely related to those of *Shigella*; that is, a virulence plasmid (pINV), genomic islands (SHI) and inactivation of genes[3]. WGS-based phylogeny showed that *Shigella* strains, independent of their 'species' designation, belong to five clearly defined monophyletic clades (S1–S5) that are part of phylogroup B1 (S1, S2) or E (S4), or between phylogroups A plus B1 and E (S3, S5)[182], reflecting their multiple independent origins within *E. coli*[13]. Using the same approach, EIEC strains isolated worldwide were shown to belong to three main lineages (phylogroups A, B1 and E) (FIG. 4d) that are distinct from the *Shigella* clades, indicating independent emergence[183,184]. Phylogenies of the pINV and the chromosomal genes are largely congruent in *Shigella* and EIEC, demonstrating that the pINV was acquired early in the emergence of the lineages, and stably co-evolve with the chromosome over time[183,185].

In sum, *Shigella* and EIEC emerged independently multiple times after the emergence of phylogroup E within *E. coli* and stably co-evolved with a virulence plasmid. The acquisition of genomic islands and the inactivation of chromosomal genes are less important in EIEC than in *Shigella*, which could indicate that EIEC lineages are evolving towards *Shigella* phenotypes. As with *Shigella*, EIEC is reported only in humans, indicting a degree of host specificity very rarely observed in the genus *Escherichia*.

EAEC, DAEC and AIEC. EAEC strains are defined by their characteristic aggregative adherence to Hep-2 cells in culture, mainly due to the plasmid-borne *aaf–agg* operons and the *aggR* gene encoding aggregative adherence fimbriae and a transcriptional regulator, respectively, together with chromosomally encoded factors such as a type VI secretion system, *Shigella* enterotoxin 1 and secreted autotransporter toxin Sat[3]. No large-scale phylogenomic studies are available for this pathotype but, based on their phylogroup membership, sequence types and serotypes, it is clear that EAEC clones emerged from at least four phylogroups (A, B1, B2 and D), belong to numerous sequence types and exhibit various serotypes[186–188]. Interestingly, specific combinations of plasmid-encoded and chromosomally encoded virulence genes have been associated with diarrhoea, specifically a gene encoding a SepA autotransporter having a major role[186]. Besides humans, EAEC strains are very frequently isolated from domestic animals with diarrhoea[189].

Diffusely adherent *E. coli* (DAEC) is a heterogeneous group that encompasses both diarrhoea-causing strains (InPEC) and UTI-causing strains (ExPEC), representing an exception to the classic InPEC–ExPEC dichotomy (see below). This DAEC group is characterized by a diffuse adherence pattern on epithelial cells that is mediated by related afimbrial and fimbrial adhesins[3]. DAEC strains are scattered among all of the phylogenetic groups[77,190]. The archetypal C1845 strain (causing diarrhoea in children) and IH11128 strain (causing UTI) belong to the B2_STc14_O75:H5 lineage, whereas the EC7372 strain (causing UTI) and DAEC11 strain (causing diarrhoea) belong to the B2_ST131_O25:H4_*fimH*22 clade[91]. Additional phylogenomic data are needed for a better understanding of this pathovar.

AIEC strains are characterized by their phenotype when interacting with eukaryotic cells; that is, adhesion and invasion of intestinal epithelial cells, and survival and replication within macrophages[191]. They have been linked to Crohn's disease, but their exact role in the disease remains poorly understood[192]. The archetypal strain is LF82, which belongs to phylogroup B2, is a member of a rarely observed sequence type (ST135) and has the O83:H1 serotype. Epidemiological studies on a small number of AIEC strains showed that two-thirds of the strains belong to phylogroup B2, and the remaining strains are distributed in all of the phylogroups[193,194]. Within phylogroup B2, besides ST135, AIEC strains are found within the classic ExPEC lineages such as ST95,

ST73, ST127 and ST131 (REFS[194,195]). To date, no convincing specific molecular property of the AIEC phenotype has been identified[195].

In conclusion, intestinal virulence is paucigenic, as it involves few genes, some with strong effects, with the arrival and persistence of specific virulence genes in most cases in specific chromosomal backgrounds. Strains causing severe diarrhoea (STEC and/or EHEC, ETEC and EIEC) belong mainly to phylogroups A, B1 and E (FIG. 4; TABLE 1). Whereas EIEC and AIEC have been reported only in humans, STEC and/or EHEC, ETEC and EPEC are frequently found in cattle, with some level of host specialization. STEC and/or EHEC, ETEC and EPEC can also be found in wild animals, although the role of such strains as reservoirs of human pathogenic strains is unclear[145,147,196–198].

Hybrid clones: breaking the boundaries

In 2011, a particular EHEC strain caused a devastating outbreak in Germany, infecting nearly 4000 people; 900 of those affected developed HUS, which was fatal for 54 individuals[199]. This EHEC–EAEC hybrid strain of serotype O104:H4 and ST678 belongs to phylogroup B1. The exceptional rate of HUS and mortality was, in part, due to a combination of properties characteristic of EHEC, Stx2a production (although not harbouring the LEE of EPEC) and EAEC-type adherence. Besides these intestinal virulence factors, at least two virulence factors, mainly encountered in ExPEC, were also observed — the iron capture systems yersiniabactin and aerobactin. This combination of virulence factors illustrates the ability of *E. coli* to break the boundaries between pathotypes, enhancing their clinical virulence and leading to substantial morbidity and mortality. Fortunately, this outbreak was short-lived, and since 2012 this clone has been rarely observed. However, as the reservoir has not yet been established, a resurgence of HUS due to this hybrid pathotype remains a threat (for a review, see REF.[10]). Other intestinal hybrid pathotypes, such as STEC–ETEC[200] and EPEC–ETEC[201], have been reported in clones belonging to phylogroups A and B1 and cause diarrhoea in humans.

Although harbouring some extra-intestinal virulence factors, none of the patients infected with the O104:H4 hybrid pathotype developed an invasive infection during the outbreak in Germany. More recently, an emerging hybrid clone has been described, able to cause severe HUS but also invasive infections such as bacteraemia[202,203]. Initially described in France, it is now present in several European countries[204,205]. This clone of serotype O80:H2 (STc165, ST301) harbours all of the typical virulence factors of EHEC (Stx, intimin and enterohaemolysin) and a large plasmid (>100 kb). Similar plasmids are known to be associated with extra-intestinal invasive infections in humans, notably neonatal meningitis (plasmid pS88), and to be present in APEC. As for pS88, the large plasmid of the O80:H2 clone encodes virulence factors such as aerobactin, salmochelin, an iron uptake protein encoded by *sitABCD*, a serum resistance protein, a putative secretion system, an omptin and a haemolysin, and also contains two bacteriocins, *cia* and *cva*. Of note, this plasmid, unlike the pS88 plasmid, has incorporated a resistance cassette conferring resistance to penicillins, cotrimoxazole, tetracyclines, streptomycin and heavy metals, such as mercury[202] (FIG. 5).

The O80:H2 clone also has an atypical phylogenetic affiliation, as it belongs to phylogroup A unlike all of

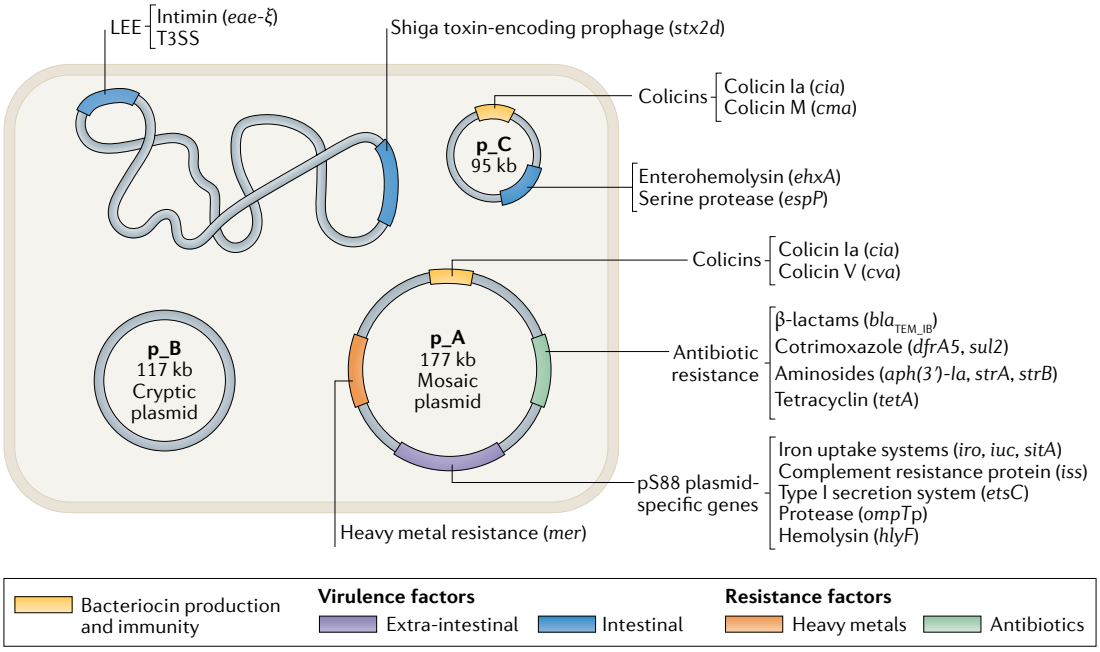


Fig. 5 | **Virulence and resistance factors of the hybrid InPEC–ExPEC pathotype O80:H2 *Escherichia coli* clone.** Schematic representation of the genes that encode virulence and resistance factors and their genetic localization. The chromosome and the three plasmids (p_A, p_B, p_C) are represented. The virulence, bacteriocin and resistance determinants are presented in different colours. ExPEC, extra-intestinal pathogenic *Escherichia coli*; InPEC, intestinal pathogenic *E. coli*; LEE, locus of enterocyte effacement; T3SS, type III secretion system.

the other major EHEC lineages. Therefore, the O80:H2 clone seems to have crossed two boundaries by the successful combination of intestinal and extra-intestinal virulence factors with clinical expression, in an atypical genetic background. Considering that some isolates of this clone have been recently shown to produce additional antibiotic-resistance mechanisms, such as ESBL[206], the O80:H2 clone and, more broadly, STc165 strains, owing to their capacity to integrate genetic elements conferring virulence and resistance, represent a serious health threat. The O80:H2 clone is unlikely to be an exception. Such 'heteropathogenicity' has been also observed in some typical B2 ExPEC lineages exhibiting EPEC and STEC genetic determinants. O2:H6 uropathogenic *E. coli* isolates that belong to ST141 possess the arsenal, although incomplete, of the EHEC pathotype[207,208] and an O4:H1 ST12 strain containing the LEE and *bfp* genes of tEPEC caused severe diarrhoea followed by bacteraemia in an immunocompromised patient[209].

Conclusions

Due to the population genetic structure and the ecology of *E. coli*, we will be constantly faced with the emergence of new pathogenic clones. The many combinations of genes linked to virulence found in specific genomic backgrounds reflect complex intergenic epistasis, and thus the emergence of novel clones is unpredictable. The classification of pathogenic clones into pathotypes has been of invaluable help in comprehending *E. coli* pathogenesis and epidemiology. However, WGS in the context of population genetics, describing the phylogenetic history of the strains based on the core genome in association with the variable genome, will enable a more meaningful and constantly updated classification of these pathogenic strains. Future research should use the power of WGS and statistics to decipher the crosstalk between virulence determinants and the remaining genome.

Published online: 21 August 2020

1. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal *Escherichia coli*. *Nat. Rev. Microbiol.* **8**, 207–217 (2010).
2. Kaper, J. B., Nataro, J. P. & Mobley, H. L. Pathogenic *Escherichia coli*. *Nat. Rev. Microbiol.* **2**, 123–140 (2004).
This paper presents a precise, exhaustive and concise review of pathogenic *E. coli* that remains at the forefront.
3. Croxen, M. A. & Finlay, B. B. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat. Rev. Microbiol.* **8**, 26–38 (2010).
4. Russo, T. A. & Johnson, J. R. Medical and economic impact of extraintestinal infections due to *Escherichia coli*: focus on an increasingly important endemic problem. *Microbes Infect.* **5**, 449–456 (2003).
5. Bruyand, M. et al. Paediatric haemolytic uraemic syndrome related to Shiga toxin-producing *Escherichia coli*, an overview of 10 years of surveillance in France, 2007 to 2016. *Euro Surveill.* **24**, 1800068 (2019).
6. Basmaci, R. et al. *Escherichia coli* meningitis features in 325 children from 2001 to 2013 in France. *Clin. Infect. Dis.* **61**, 779–786 (2015).
7. Lefort, A. et al. Host factors and portal of entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. *J. Clin. Microbiol.* **49**, 777–783 (2011).
8. Abernethy, J. K. et al. Thirty day all-cause mortality in patients with *Escherichia coli* bacteraemia in England. *Clin. Microbiol. Infect.* **21**, 251 e251–251.e8 (2015).
9. Vihta, K. D. et al. Trends over time in *Escherichia coli* bloodstream infections, urinary tract infections, and antibiotic susceptibilities in Oxfordshire, UK, 1998–2016: a study of electronic health records. *Lancet Infect. Dis.* **18**, 1138–1149 (2018).
10. Karch, H. et al. The enemy within us: lessons from the 2011 European *Escherichia coli* O104:H4 outbreak. *EMBO Mol. Med.* **4**, 841–848 (2012).
11. Cassini, A. et al. Attributable deaths and disability-adjusted life-years caused by infections with antibiotic-resistant bacteria in the EU and the European economic area in 2015: a population-level modelling analysis. *Lancet Infect. Dis.* **19**, 56–66 (2019).
12. Walk, S. T. The 'cryptic' *Escherichia*. *EcoSal Plus* https://doi.org/10.1128/ecosalplus.ESP-0002-2015 (2015).
13. Pupo, G. M., Lan, R. & Reeves, P. R. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl Acad. Sci. USA* **97**, 10567–10572 (2000).
14. The, H. C., Thanh, D. P., Holt, K. E., Thomson, N. R. & Baker, S. The genomic signatures of *Shigella* evolution, adaptation and geographical spread. *Nat. Rev. Microbiol.* **14**, 235–250 (2016).
15. Walk, S. T. et al. Cryptic lineages of the genus *Escherichia*. *Appl. Environ. Microbiol.* **75**, 6534–6544 (2009).
This paper describes the *Escherichia* clades and thus demonstrates that an important discovery on *E. coli* diversity can be made 125 years after its first isolation.
16. Priest, F. G. & Barker, M. Gram-negative bacteria associated with brewery yeasts: reclassification of *Obesumbacterium proteus* biogroup 2 as *Shimwellia pseudoproteus* gen. nov., sp. nov., and transfer of *Escherichia blattae* to *Shimwellia blattae* comb. nov. *Int. J. Syst. Evol. Microbiol.* **60**, 828–833 (2010).
17. Hata, H. et al. Phylogenetics of family Enterobacteriaceae and proposal to reclassify *Escherichia hermannii* and *Salmonella subterranea* as *Atlantibacter hermannii* and *Atlantibacter subterranea* gen. nov., comb. nov. *Microbiol. Immunol.* **60**, 303–311 (2016).
18. Alnajar, S. & Gupta, R. S. Phylogenomics and comparative genomic studies delineate six main clades within the family Enterobacteriaceae and support the reclassification of several polyphyletic members of the family. *Infect. Genet. Evol.* **54**, 108–127 (2017).
19. Jain, C., Rodriguez, R. L., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
20. Liu, S. et al. *Escherichia marmotae* sp. nov., isolated from faeces of *Marmota himalayana*. *Int. J. Syst. Evol. Microbiol.* **65**, 2130–2134 (2015).
21. Clermont, O., Christenson, J. K., Denamur, E. & Gordon, D. M. The Clermont *Escherichia coli* phylotyping method revisited: improvement of specificity and detection of new phylo-groups. *Environ. Microbiol. Rep.* **8**, 58–65 (2013).
22. Luo, C. et al. Genome sequencing of environmental *Escherichia coli* expands understanding of the ecology and speciation of the model bacterial species. *Proc. Natl Acad. Sci. USA* **108**, 7200–7205 (2011).
23. Lu, S. et al. Insights into the evolution of pathogenicity of *Escherichia coli* from genomic analysis of intestinal *E. coli* of *Marmota himalayana* in Qinghai–Tibet plateau of China. *Emerg. Microbes Infect.* **5**, e122 (2016).
24. Clermont, O., Bonacorsi, S. & Bingen, E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl. Environ. Microbiol.* **66**, 4555–4558 (2000).
25. Clermont, O. et al. Characterization and rapid identification of phylogroup G in *Escherichia coli*, a lineage with high virulence and antibiotic resistance potential. *Environ. Microbiol.* **21**, 3107–3117 (2019).
26. Clermont, O., Gordon, D. M., Brisse, S., Walk, S. T. & Denamur, E. Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence. *Environ. Microbiol.* **13**, 2468–2477 (2011).
27. Smati, M. et al. Quantitative analysis of commensal *Escherichia coli* populations reveals host-specific enterotypes at the intra-species level. *Microbiologyopen* **4**, 604–615 (2015).
28. Ooka, T. et al. Defining the genome features of *Escherichia albertii*, an emerging enteropathogen closely related to *Escherichia coli*. *Genome Biol. Evol.* **7**, 3170–3179 (2015).
29. Lindsey, R. L., Garcia-Toledo, L., Fasulo, D., Gladney, L. M. & Strockbine, N. Multiplex polymerase chain reaction for identification of *Escherichia coli*, *Escherichia albertii* and *Escherichia fergusonii*. *J. Microbiol. Methods* **140**, 1–4 (2017).
30. Beghain, J., Bridier-Nahmias, A., Le Nagard, H., Denamur, E. & Clermont, O. ClermonTyping: an easy-to-use and accurate in silico method for *Escherichia* genus strain phylotyping. *Microb. Genom.* **4**, e000192 (2018).
31. Zhou, Z. et al. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. *Genome Res.* **30**, 138–152 (2020).
This paper presents a unique database of *E. coli* (and other) genomes with an integrated software environment representing a great tool for the scientific community.
32. Touchon, M. et al. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 https://doi.org/10.1371/journal.pgen.1000344 (2009).
This paper presents a comprehensive comparative genomic analysis of *E. coli* performed on a small number of high-quality sequences that lays the basic concepts of *E. coli* genomic structure.
33. Lukjancenko, O., Wassenaar, T. M. & Ussery, D. W. Comparison of 61 sequenced *Escherichia coli* genomes. *Microb. Ecol.* **60**, 708–720 (2010).
34. Touchon, M. et al. Phylogenetic background and habitat drive the genetic diversification of *Escherichia coli*. *PLoS Genet.* **16**, e1008866 (2020).
35. Bergthorsson, U. & Ochman, H. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* **15**, 6–16 (1998).
36. Rasko, D. A. et al. The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J. Bacteriol.* **190**, 6881–6893 (2008).
37. Kallonen, T. et al. Systematic longitudinal survey of invasive *Escherichia coli* in England demonstrates a stable population structure only transiently disturbed by the emergence of ST131. *Genome Res.* **27**, 1437–1449 (2017).
This paper is the first study using WGS to type a large number of bacteraemia *E. coli* strains over a 10-year period.
38. Milkman, R. & Bridges, M. M. Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. *Genet.* **133**, 455–468 (1993).
39. Mau, B., Glasner, J. D., Darling, A. E. & Perna, N. T. Genome-wide detection and analysis of homologous

recombination among sequenced strains of *Escherichia coli*. *Genome Biol.* **7**, R44 (2006).
40. Didelot, X., Meric, G., Falush, D. & Darling, A. E. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* **13**, 256 (2012).
41. Dixit, P. D., Pang, T. Y., Studier, F. W. & Maslov, S. Recombinant transfer in the basic genome of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **112**, 9070–9075 (2015).
42. Milkman, R., Jaeger, E. & McBride, R. D. Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. *Genetics* **163**, 475–483 (2003).
43. Schubert, S. et al. Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. *PLoS Pathog.* **5**, e1000257 (2009).
44. Williams, L. E., Wireman, J., Hilliard, V. C. & Summers, A. O. Large plasmids of *Escherichia coli* and *Salmonella* encode highly diverse arrays of accessory genes on common replicon families. *Plasmid* **69**, 36–48 (2013).
45. Brolund, A., Franzen, O., Melefors, O., Tegmark-Wisell, K. & Sandegren, L. Plasmidome-analysis of ESBL-producing *Escherichia coli* using conventional typing and high-throughput sequencing. *PLoS ONE* **8**, e65793 (2013).
46. de Toro, M., Garcillaon-Barcia, M. P. & De La Cruz, F. Plasmid diversity and adaptation analyzed by massive sequencing of *Escherichia coli* plasmids. *Microbiol. Spectr.* **2**, https://doi.org/10.1128/microbiolspec.PLAS-0031-2014 (2014).
47. Boyd, E. F., Hill, C. W., Rich, S. M. & Hartl, D. L. Mosaic structure of plasmids from natural populations of *Escherichia coli*. *Genetics* **143**, 1091–1100 (1996).
This important paper before the genomic era shows that plasmids are distributed within the *E. coli* species according to the strain phylogeny.
48. Smillie, C., Garcillan-Barcia, M. P., Francia, M. V., Rocha, E. P. & de la Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74**, 434–452 (2010).
49. Branger, C. et al. Extended-spectrum β-lactamase-encoding genes are spreading on a wide range of *Escherichia coli* plasmids existing prior to the use of third-generation cephalosporins. *Microb. Genom.* **4**, e000203 (2018).
50. Branger, C. et al. Specialization of small non-conjugative plasmids in *Escherichia coli* according to their family types. *Microb. Genom.* **5**, e000281 (2019).
51. Maynard Smith, J., Smith, N. H., O'Rourke, M. & Spratt, B. G. How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4384–4388 (1993).
52. Desjardins, P., Picard, B., Kaltenbock, B., Elion, J. & Denamur, E. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. *J. Mol. Evol.* **41**, 440–448 (1995).
53. Orskov, F. et al. Special *Escherichia coli* serotypes among enterotoxigenic strains from diarrhoea in adults and children. *Med. Microbiol. Immunol.* **162**, 73–80 (1976).
54. Fratamico, P. M. et al. Advances in molecular serotyping and subtyping of *Escherichia coli*. *Front. Microbiol.* **7**, 644 (2016).
55. Selander, R. K. & Levin, B. R. Genetic diversity and structure in *Escherichia coli* populations. *Science* **210**, 545–547 (1980).
This seminal paper demonstrates the clonal structure of the *E. coli* species using MLEE.
56. Maiden, M. C. et al. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145 (1998).
57. Ingle, D. J. et al. *In silico* serotyping of *E. coli* from short read data identifies limited novel O-loci but extensive diversity of O:H serotype combinations within and between pathogenic lineages. *Microb. Genom.* **2**, e000064 (2016).
58. Wirth, T. et al. Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.* **60**, 1136–1151 (2006).
59. Jaureguy, F. et al. Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* **9**, 560 (2008).
60. Roer, L. et al. Development of a web tool for *Escherichia coli* subtyping based on *fimH* alleles. *J. Clin. Microbiol.* **55**, 2538–2543 (2017).
61. Maiden, M. C. et al. MLST revisited: the gene-by-gene approach to bacterial genomics. *Nat. Rev. Microbiol.* **11**, 728–736 (2013).
62. Riley, L. W. Pandemic lineages of extraintestinal pathogenic *Escherichia coli*. *Clin. Microbiol. Infect.* **20**, 380–390 (2014).
63. Pitout, J. D. & DeVinney, R. *Escherichia coli* ST131: a multidrug-resistant clone primed for global domination. *F1000Res* https://doi.org/10.12688/f1000research.10609.1 (2017).
64. Karmali, M. A. et al. Association of genomic O island 122 of *Escherichia coli* EDL 933 with verocytotoxin-producing *Escherichia coli* seropathotypes that are linked to epidemic and/or serious disease. *J. Clin. Microbiol.* **41**, 4930–4940 (2003).
65. Johnson, C. M. & Grossman, A. D. Integrative and conjugative elements (ICEs): what they do and how they work. *Annu. Rev. Genet.* **49**, 577–601 (2015).
66. Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. Genomic islands in pathogenic and environmental microorganisms. *Nat. Rev. Microbiol.* **2**, 414–424 (2004).
67. Bliven, K. A. & Maurelli, A. T. Antivirulence genes: insights into pathogen evolution through gene loss. *Infect. Immun.* **80**, 4061–4070 (2012).
68. Williams, G. C. Pleiotropy, natural selection, and the evolution of senescence. *Evolution* **11**, 398–411 (1957).
69. Maurelli, A. T., Fernandez, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of *Shigella spp*. and enteroinvasive *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **95**, 3943–3948 (1998).
70. Cooper, V. S. & Lenski, R. E. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature* **407**, 736–739 (2000).
71. Sokurenko, E. V., Hasty, D. L. & Dykhuizen, D. E. Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol.* **7**, 191–195 (1999).
72. Sokurenko, E. V. et al. Pathogenic adaptation of *Escherichia coli* by natural variation of the FimH adhesin. *Proc. Natl Acad. Sci. USA* **95**, 8922–8926 (1998).
73. Tenaillon, O. et al. The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
74. Reid, S. D., Herbelin, C. J., Bumbaugh, A. C., Selander, R. K. & Whittam, T. S. Parallel evolution of virulence in pathogenic *Escherichia coli*. *Nature* **406**, 64–67 (2000).
This is the first paper showing nicely convergent evolution of intestinal virulence in *E. coli*, a strong sign of selection.
75. Ogura, Y. et al. Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **106**, 17939–17944 (2009).
76. Ogura, Y. et al. Population structure of *Escherichia coli* O26: H11 with recent and repeated *stx2* acquisition in multiple lineages. *Microb. Genom.* **3**, e000141 (2017).
77. Escobar-Paramo, P. et al. A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol. Biol. Evol.* **21**, 1085–1094 (2004).
This paper indicates the important role of the genetic background in the emergence of virulent *E. coli* clones at the species level.
78. Day, W. A. Jr., Fernández, R. E. & Maurelli, A. T. Pathoadaptive mutations that enhance virulence: genetic organization of the *cadA* regions of *Shigella* spp. *Infect. Immun.* **69**, 7471–7480 (2001).
79. Hommais, F. et al. The FimH A27V mutation is pathoadaptive for urovirulence in *Escherichia coli* B2 phylogenetic group isolates. *Infect. Immun.* **71**, 3619–3622 (2003).
80. Sannes, M. R., Kuskowski, M. A., Owens, K., Gajewski, A. & Johnson, J. R. Virulence factor profiles and phylogenetic background of *Escherichia coli* isolates from veterans with bacteremia and uninfected control subjects. *J. Infect. Dis.* **190**, 2121–2128 (2004).
81. Picard, B. et al. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect. Immun.* **67**, 546–553 (1999).
82. Johnson, J. R. et al. Experimental mouse lethality of *Escherichia coli* isolates, in relation to accessory traits, phylogenetic group, and ecological source. *J. Infect. Dis.* **194**, 1141–1150 (2006).
83. Johnson, J. R. et al. Accessory traits and phylogenetic background predict *Escherichia coli* extraintestinal virulence better than does ecological source. *J. Infect. Dis.* **219**, 121–132 (2019).
84. Girardeau, J. P. et al. Association of virulence genotype with phylogenetic background in comparison to different seropathotypes of Shiga toxin-producing *Escherichia coli* isolates. *J. Clin. Microbiol.* **43**, 6098–6107 (2005).
85. Mellmann, A. et al. Analysis of collection of hemolytic uremic syndrome-associated enterohemorrhagic *Escherichia coli*. *Emerg. Infect. Dis.* **14**, 1287–1290 (2008).
86. Domingo, J., Baeza-Centurion, P. & Lehner, B. The causes and consequences of genetic interactions (epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).
87. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
88. Smati, M. et al. Strain-specific impact of the high-pathogenicity island on virulence in extra-intestinal pathogenic *Escherichia coli*. *Int. J. Med. Microbiol.* **307**, 44–56 (2017).
89. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
90. Kauffmann, F. The serology of the *coli* group. *J. Immunol.* **57**, 71–100 (1947).
This impressive review published just after the Second World War synthesizes knowledge of the *E. coli* extra-intestinal clone population structure based on the serotypes that sounds remarkably accurate today.
91. Le Gall, T. et al. Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. *Mol. Biol. Evol.* **24**, 2373–2384 (2007).
92. Adams-Sapper, S., Diep, B. A., Perdreau-Remington, F. & Riley, L. W. Clonal composition and community clustering of drug-susceptible and -resistant *Escherichia coli* isolates from bloodstream infections. *Antimicrob. Agents Chemother.* **57**, 490–497 (2013).
93. Day, M. J. et al. Population structure of *Escherichia coli* causing bacteraemia in the UK and Ireland between 2001 and 2010. *J. Antimicrob. Chemother.* **71**, 2139–2142 (2016).
94. Clermont, O. et al. Two levels of specialization in bacteraemic *Escherichia coli* strains revealed by their comparison with commensal strains. *Epidemiol. Infect.* **145**, 872–882 (2017).
95. Yoon, E. J. et al. Impact of host–pathogen–treatment tripartite components on early mortality of patients with *Escherichia coli* bloodstream infection: prospective observational study. *EBioMedicine* **35**, 76–86 (2018).
96. Manges, A. R. et al. Widespread distribution of urinary tract infections caused by a multidrug-resistant *Escherichia coli* clonal group. *N. Engl. J. Med.* **345**, 1007–1013 (2001).
97. Nicolas-Chanoine, M. H. et al. Intercontinental emergence of *Escherichia coli* clone O25:H4-ST131 producing CTX-M-15. *J. Antimicrob. Chemother.* **61**, 273–281 (2008).
98. Nicolas-Chanoine, M. H., Bertrand, X. & Madec, J. Y. *Escherichia coli* ST131, an intriguing clonal group. *Clin. Microbiol. Rev.* **27**, 543–574 (2014).
99. Price, L. B. et al. The epidemic of extended-spectrum-β-lactamase-producing *Escherichia coli* ST131 is driven by a single highly pathogenic subclone, H30-Rx. *MBio* **4**, e00377–e00413 (2013).
100. Petty, N. K. et al. Global dissemination of a multidrug resistant *Escherichia coli* clone. *Proc. Natl Acad. Sci. USA* **111**, 5694–5699 (2014).
101. Matsumura, Y. et al. Global *Escherichia coli* sequence type 131 clade with *blaCTX-M-27* gene. *Emerg. Infect. Dis.* **22**, 1900–1907 (2016).
102. Stoesser, N. et al. Evolutionary history of the global emergence of the *Escherichia coli* epidemic clone ST131. *MBio* **7**, e02162 (2016).
103. Ben Zakour, N. L. et al. Sequential acquisition of virulence and fluoroquinolone resistance has shaped the evolution of *Escherichia coli* ST131. *MBio* **7**, e00347–e00416 (2016).
This paper is a clear demonstration of the stepwise evolution of the pandemic ExPEC ST131 *E. coli* clonal group.
104. McNally, A. et al. Combined analysis of variation in core, accessory and regulatory genome regions provides a super-resolution view into the evolution

of bacterial populations. *PLoS Genet.* **12**, e1006280 (2016).
105. Decano, A. G. & Downing, T. An *Escherichia coli* ST131 pangenome atlas reveals population structure and evolution across 4,071 isolates. *Sci. Rep.* **9**, 17394 (2019).
106. Johnson, T. J. et al. Separate F-type plasmids have shaped the evolution of the *H*30 subclone of *Escherichia coli* sequence type 131. *mSphere* **1**, e00121–e00216 (2016).
107. Kondratyeva, K., Salmon-Divon, M. & Navon-Venezia, S. Meta-analysis of pandemic *Escherichia coli* ST131 plasmidome proves restricted plasmid–clade associations. *Sci. Rep.* **10**, 36 (2020).
108. Ewers, C. et al. Emergence of human pandemic O25:H4-ST131 CTX-M-15 extended-spectrum-β-lactamase-producing *Escherichia coli* among companion animals. *J. Antimicrob. Chemother.* **65**, 651–660 (2010).
109. Zogg, A. L., Zurfluh, K., Schmitt, S., Nuesch-Inderbinen, M. & Stephan, R. Antimicrobial resistance, multilocus sequence types and virulence profiles of ESBL producing and non-ESBL producing uropathogenic *Escherichia coli* isolated from cats and dogs in Switzerland. *Vet. Microbiol.* **216**, 79–84 (2018).
110. Johnson, J. R. et al. Household clustering of *Escherichia coli* sequence type 131 clinical and fecal isolates according to whole genome sequence analysis. *Open Forum Infect. Dis.* **3**, ofw129 https://doi.org/10.1093/ofid/ofw129 (2016).
111. Liu, C. M. et al. *Escherichia coli* ST131-*H*22 as a foodborne uropathogen. *MBio* **9,** e00470 https://doi.org/10.1128/mBio.00470-18 (2018).
112. Mukerji, S. et al. Resistance to critically important antimicrobials in Australian silver gulls (*Chroicocephalus novaehollandiae*) and evidence of anthropogenic origins. *J. Antimicrob. Chemother.* **74**, 2566–2574 (2019).
113. Vimont, S. et al. The CTX-M-15-producing *Escherichia coli* clone O25b:H4-ST131 has high intestine colonization and urinary tract infection abilities. *PLoS ONE* **7**, e46547 (2012).
114. Flament-Simon, S. C. et al. Association between kinetics of early biofilm formation and clonal lineage in *Escherichia coli*. *Front. Microbiol.* **10**, 1183 (2019).
115. Sarkar, S. et al. Intestinal colonization traits of pandemic multidrug-resistant *Escherichia coli* ST131. *J. Infect. Dis.* **218**, 979–990 (2018).
116. Johnson, J. R., Porter, S. B., Zhanel, G., Kuskowski, M. A. & Denamur, E. Virulence of *Escherichia coli* clinical isolates in a murine sepsis model in relation to sequence type ST131 status, fluoroquinolone resistance, and virulence genotype. *Infect. Immun.* **80**, 1554–1562 (2012).
117. McNally, A. et al. Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *MBio* **10**, e00644 https://doi.org/10.1128/mBio.00644-19 (2019).
118. Bengtsson, S., Naseer, U., Sundsfjord, A., Kahlmeter, G. & Sundqvist, M. Sequence types and plasmid carriage of uropathogenic *Escherichia coli* devoid of phenotypically detectable resistance. *J. Antimicrob. Chemother.* **67**, 69–73 (2012).
119. Roer, L. et al. WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. *J. Antimicrob. Chemother.* **72**, 1922–1929 (2017).
120. Bogema, D. R. et al. Whole-genome analysis of extraintestinal *Escherichia coli* sequence type 73 from a single hospital over a 2 year period identified different circulating clonal groups. *Microb. Genom.* **6**, e000255 (2020).
121. Gordon, D. M. et al. Fine-scale structure analysis shows epidemic patterns of clonal complex 95, a cosmopolitan *Escherichia coli* lineage responsible for extraintestinal infection. *mSphere* **2**, e00168–e00217 (2017).
122. Jorgensen, S. L. et al. Diversity and population overlap between avian and human *Escherichia coli* belonging to sequence type 95. *mSphere* **4**, 00333–00418 (2019).
123. Bidet, P. et al. Combined multilocus sequence typing and O serogrouping distinguishes *Escherichia coli* subtypes associated with infant urosepsis and/or meningitis. *J. Infect. Dis.* **196**, 297–303 (2007).
124. Moulin-Schouleur, M. et al. Extraintestinal pathogenic *Escherichia coli* strains of avian and human origin: link between phylogenetic relationships and common virulence patterns. *J. Clin. Microbiol.* **45**, 3366–3376 (2007).
125. Danzeisen, J. L., Wannemuehler, Y., Nolan, L. K. & Johnson, T. J. Comparison of multilocus sequence analysis and virulence genotyping of *Escherichia coli* from live birds, retail poultry meat, and human extraintestinal infection. *Avian Dis.* **57**, 104–108 (2013).
126. Bourne, J. A., Chong, W. L. & Gordon, D. M. Genetic structure, antimicrobial resistance and frequency of human associated *Escherichia coli* sequence types among faecal isolates from healthy dogs and cats living in Canberra, Australia. *PLoS ONE* **14**, e0212867 (2019).
127. Johnson, J. R., Clabots, C. & Kuskowski, M. A. Multiple-host sharing, long-term persistence, and virulence of *Escherichia coli* clones from human and animal household members. *J. Clin. Microbiol.* **46**, 4078–4082 (2008).
128. Bonacorsi, S. et al. Molecular analysis and experimental virulence of French and North American *Escherichia coli* neonatal meningitis isolates: identification of a new virulent clone. *J. Infect. Dis.* **187**, 1895–1906 (2003).
129. Geslain, G. et al. Genome sequencing of strains of the most prevalent clonal group of O1:K1:H7 *Escherichia coli* that causes neonatal meningitis in France. *BMC Microbiol.* **19**, 17 (2019).
130. Stephens, C. M., Adams-Sapper, S., Sekhon, M., Johnson, J. R. & Riley, L. W. Genomic analysis of factors associated with low prevalence of antibiotic resistance in extraintestinal pathogenic *Escherichia coli* sequence type 95 strains. *mSphere* **2**, 00390–00416 (2017).
131. Johnson, J. R. et al. Global distribution and epidemiologic associations of *Escherichia coli* clonal group A, 1998–2007. *Emerg. Infect. Dis.* **17**, 2001–2009 (2011).
132. Ramchandani, M. et al. Possible animal origin of human-associated, multidrug-resistant, uropathogenic *Escherichia coli*. *Clin. Infect. Dis.* **40**, 251–257 (2005).
133. Tartof, S. Y., Solberg, O. D., Manges, A. R. & Riley, L. W. Analysis of a uropathogenic *Escherichia coli* clonal group by multilocus sequence typing. *J. Clin. Microbiol.* **43**, 5860–5864 (2005).
134. Johnson, T. J. et al. Phylogenomic analysis of extraintestinal pathogenic *Escherichia coli* sequence type 1193, an emerging multidrug-resistant clonal group. *Antimicrob. Agents Chemother.* **63,** e01913-18 https://doi.org/10.1128/AAC.01913-18 (2019).
135. Tchesnokova, V. L. et al. Rapid and extensive expansion in the United States of a new multidrug-resistant *Escherichia coli* clonal group, sequence type 1193. *Clin. Infect. Dis.* **68**, 334–337 (2019).
136. Tchesnokova, V. et al. Pandemic fluoroquinolone resistant *Escherichia coli* clone ST1193 emerged via simultaneous homologous recombinations in 11 gene loci. *Proc. Natl Acad. Sci. USA* **116**, 14740–14748 (2019).
137. La Combe, B. et al. Pneumonia-specific *Escherichia coli* with distinct phylogenetic and virulence profiles, France, 2012–2014. *Emerg. Infect. Dis.* **25**, 710–718 (2019).
138. Clermont, O. et al. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infect. Genet. Evol.* **11**, 654–662 (2011).
139. Mangiamele, P. et al. Complete genome sequence of the avian pathogenic *Escherichia coli* strain APEC O78. *Genome Announc.* **1**, e0002613 (2013).
140. Lemaitre, C. et al. A conserved virulence plasmidic region contributes to the virulence of the multiresistant *Escherichia coli* meningitis strain S286 belonging to phylogenetic group C. *PLoS ONE* **8**, e74423 (2013).
141. Peigne, C. et al. The plasmid of *Escherichia coli* strain S88 (O45:K1:H7) that causes neonatal meningitis is closely related to avian pathogenic *E. coli* plasmids and is associated with high-level bacteremia in a neonatal rat meningitis model. *Infect. Immun.* **77**, 2272–2284 (2009).
142. Huja, S. et al. Genomic avenue to avian colisepticemia. *MBio* **6,** e01681 https://doi.org/10.1128/mBio.01681-14 (2015).
143. Roer, L. et al. *Escherichia coli* sequence type 410 is causing new international high-risk clones. *mSphere* **3**, 00337–00418 (2018).
144. Tourret, J., Diard, M., Garry, L., Matic, I. & Denamur, E. Effects of single and multiple pathogenicity island deletions on uropathogenic *Escherichia coli* strain 536 intrinsic extra-intestinal virulence. *Int. J. Med. Microbiol.* **300**, 435–439 (2010).
145. Lescat, M. et al. Commensal *Escherichia coli* strains in Guiana reveal a high genetic diversity with host-dependant population structure. *Environ. Microbiol. Rep.* **5**, 49–57 (2013).
146. Skurnik, D. et al. Emergence of antimicrobial-resistant *Escherichia coli* of animal origin spreading in humans. *Mol. Biol. Evol.* **33**, 898–914 (2016).
147. Mora, A. et al. Impact of human-associated *Escherichia coli* clonal groups in Antarctic pinnipeds: presence of ST73, ST95, ST141 and ST131. *Sci. Rep.* **8**, 4678 (2018).
148. Melton-Celsa, A. R. Shiga toxin (Stx) classification, structure, and function. *Microbiol. Spectr.* **2,** https://doi.org/10.1128/microbiolspec.EHEC-0024-2013 (2014).
149. Riley, L. W. et al. Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *N. Engl. J. Med.* **308**, 681–685 (1983).
150. Feng, P., Lampel, K. A., Karch, H. & Whittam, T. S. Genotypic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. *J. Infect. Dis.* **177**, 1750–1753 (1998).
This MLEE-based paper proposes an evolutionary scenario, later confirmed, for the emergence of O157:H7 EHEC.
151. Feng, P. C. et al. Genetic diversity among clonal lineages within *Escherichia coli* O157:H7 stepwise evolutionary model. *Emerg. Infect. Dis.* **13**, 1701–1706 (2007).
152. Dallman, T. J. et al. Applying phylogenomics to understand the emergence of Shiga-toxin-producing *Escherichia coli* O157:H7 strains causing severe human disease in the UK. *Microb. Genom.* **1**, e000029 (2015).
153. Lupolova, N., Dallman, T. J., Matthews, L., Bono, J. L. & Gally, D. L. Support vector machine applied to predict the zoonotic potential of *E. coli* O157 cattle isolates. *Proc. Natl Acad. Sci. USA* **113**, 11312–11317 (2016).
This paper presents an original support vector machine analysis showing that a minor subset of bovine O157:H7 *E. coli* has the potential to cause human disease.
154. Franz, E. et al. Phylogeographic analysis reveals multiple international transmission events have driven the global emergence of *Escherichia coli* O157:H7. *Clin. Infect. Dis.* **69**, 428–437 (2019).
155. Kim, J., Nietfeldt, J. & Benson, A. K. Octamer-based genome scanning distinguishes a unique subpopulation of *Escherichia coli* O157:H7 strains in cattle. *Proc. Natl Acad. Sci. USA* **96**, 13288–13293 (1999).
156. Zhang, Y. et al. Genome evolution in major *Escherichia coli* O157:H7 lineages. *BMC Genomics* **8**, 121 (2007).
157. Manning, S. D. et al. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proc. Natl Acad. Sci. USA* **105**, 4868–4873 (2008).
This paper shows that within a clonal group, such as O157:H7 STc11, clades can be associated with different levels of virulence.
158. Strachan, N. J. et al. Whole genome sequencing demonstrates that geographic variation of *Escherichia coli* O157 genotypes dominates host association. *Sci. Rep.* **5**, 14145 (2015).
159. Iyoda, S. et al. Phylogenetic clades 6 and 8 of enterohemorrhagic *Escherichia coli* O157:H7 with particular *stx* subtypes are more frequently found in isolates from hemolytic uremic syndrome patients than from asymptomatic carriers. *Open Forum Infect. Dis.* **1**, ofu061 (2014).
160. Fuller, C. A., Pellino, C. A., Flagler, M. J., Strasser, J. E. & Weiss, A. A. Shiga toxin subtypes display dramatic differences in potency. *Infect. Immun.* **79**, 1329–1337 (2011).
161. Ogura, Y. et al. The Shiga toxin 2 production level in enterohemorrhagic *Escherichia coli* O157:H7 is correlated with the subtypes of toxin-encoding phage. *Sci. Rep.* **5**, 16663 (2015).
162. Bielaszewska, M. et al. Enterohemorrhagic *Escherichia coli* O26:H11/H^{-}: a new virulent clone emerges in Europe. *Clin. Infect. Dis.* **56**, 1373–1381 (2013).
163. Eichhorn, I. et al. Highly virulent non-O157 enterohemorrhagic *Escherichia coli* (EHEC) serotypes reflect similar phylogenetic lineages, providing new insights into the evolution of EHEC. *Appl. Environ. Microbiol.* **81**, 7041–7047 (2015).
164. Delannoy, S., Mariani-Kurkdjian, P., Webb, H. E., Bonacorsi, S. & Fach, P. The mobilome; a major contributor to *Escherichia coli stx2*-positive O26:H11 strains intra-serotype diversity. *Front. Microbiol.* **8**, 1625 (2017).

165. Karnisova, L. et al. Attack of the clones: whole genome-based characterization of two closely related enterohemorrhagic *Escherichia coli* O26 epidemic lineages. *BMC Genomics* **19**, 647 (2018).
166. Gould, L. H. et al. Increased recognition of non-O157 Shiga toxin-producing *Escherichia coli* infections in the United States during 2000–2010: epidemiologic features and comparison with *E. coli* O157 infections. *Foodborne Pathog. Dis.* **10**, 453–460 (2013).
167. Valilis, E., Ramsey, A., Sidiq, S. & DuPont, H. L. Non-O157 Shiga toxin-producing *Escherichia coli* — a poorly appreciated enteric pathogen: systematic review. *Int. J. Infect. Dis.* **76**, 82–87 (2018).
168. Gonzalez-Escalona, N. & Kase, J. A. Virulence gene profiles and phylogeny of Shiga toxin-positive *Escherichia coli* strains isolated from FDA regulated foods during 2010–2017. *PLoS ONE* **14**, e0214620 (2019).
169. Mellmann, A. et al. Phylogeny and disease association of Shiga toxin-producing *Escherichia coli* O91. *Emerg. Infect. Dis.* **15**, 1474–1477 (2009).
170. Ingle, D. J. et al. Evolution of atypical enteropathogenic *E. coli* by repeated acquisition of LEE pathogenicity island variants. *Nat. Microbiol.* **1**, 15010 (2016).
This detailed genomic study based on almost 200 EPEC strains refines the concept of convergent evolution of intestinal virulence.
171. Hazen, T. H., Kaper, J. B., Nataro, J. P. & Rasko, D. A. Comparative genomics provides insight into the diversity of the attaching and effacing *Escherichia coli* virulence plasmids. *Infect. Immun.* **83**, 4103–4117 (2015).
This paper presents an original study of the global transcriptome of an EPEC strain and its virulence plasmid mutants, showing molecular crosstalk between the plasmid and the chromosome.
172. Hazen, T. H. et al. Genomic diversity of EPEC associated with clinical presentations of differing severity. *Nat. Microbiol.* **1**, 15014 (2016).
173. Moura, R. A. et al. Clonal relationship among atypical enteropathogenic *Escherichia coli* strains isolated from different animal species and humans. *Appl. Environ. Microbiol.* **75**, 7399–7408 (2009).
174. Hazen, T. H., Daugherty, S. C., Shetty, A. C., Nataro, J. P. & Rasko, D. A. Transcriptional variation of diverse enteropathogenic *Escherichia coli* isolates under virulence-inducing conditions. *mSystems* **2**, e00024 https://doi.org/10.1128/mSystems.00024-17 (2017).
175. Lacher, D. W., Steinsland, H., Blank, T. E., Donnenberg, M. S. & Whittam, T. S. Molecular evolution of typical enteropathogenic *Escherichia coli*: clonal analysis by multilocus sequence typing and virulence gene allelic profiling. *J. Bacteriol.* **189**, 342–350 (2007).
176. von Mentzer, A. et al. Identification of enterotoxigenic *Escherichia coli* (ETEC) clades with long-term global distribution. *Nat. Genet.* **46**, 1321–1326 (2014).
177. Joffre, E. et al. Allele variants of enterotoxigenic *Escherichia coli* heat-labile toxin are globally transmitted and associated with colonization factors. *J. Bacteriol.* **197**, 392–403 (2015).
178. Rasko, D. A. et al. Comparative genomic analysis and molecular examination of the diversity of enterotoxigenic *Escherichia coli* isolates from Chile. *PLoS Negl. Trop. Dis.* **13**, e0007828 (2019).
179. Sahl, J. W. et al. Insights into enterotoxigenic *Escherichia coli* diversity in Bangladesh utilizing genomic epidemiology. *Sci. Rep.* **7**, 3402 (2017).
180. Shepard, S. M. et al. Genome sequences and phylogenetic analysis of K88- and F18-positive porcine enterotoxigenic. *Escherichia coli. J. Bacteriol.* **194**, 395–405 (2012).
181. Wyrsch, E. et al. Comparative genomic analysis of a multiple antimicrobial resistant enterotoxigenic *E. coli* O157 lineage from Australian pigs. *BMC Genomics* **16**, 165 (2015).
182. Sahl, J. W. et al. Defining the phylogenomics of *Shigella* species: a pathway to diagnostics. *J. Clin. Microbiol.* **53**, 951–960 (2015).
183. Hazen, T. H. et al. Investigating the relatedness of enteroinvasive *Escherichia coli* to other *E. coli* and *Shigella* isolates by using comparative genomics. *Infect. Immun.* **84**, 2362–2371 (2016).
184. Pettengill, E. A., Pettengill, J. B. & Binet, R. Phylogenetic analyses of *Shigella* and enteroinvasive *Escherichia coli* for the identification of molecular epidemiological markers: whole-genome comparative analysis does not support distinct genera designation. *Front. Microbiol.* **6**, 1573 (2015).
185. Lan, R., Lumb, B., Ryan, D. & Reeves, P. R. Molecular evolution of large virulence plasmid in *Shigella* clones and enteroinvasive *Escherichia coli*. *Infect. Immun.* **69**, 6303–6309 (2001).
186. Boisen, N. et al. Genomic characterization of enteroaggregative *Escherichia coli* from children in Mali. *J. Infect. Dis.* **205**, 431–444 (2012).
187. Imuta, N. et al. Phylogenetic analysis of enteroaggregative *Escherichia coli* (EAEC) isolates from Japan reveals emergence of CTX-M-14-producing EAEC O25:H4 clones related to sequence type 131. *J. Clin. Microbiol.* **54**, 2128–2134 (2016).
188. Zhang, R. et al. Comparative genetic characterization of enteroaggregative *Escherichia coli* strains recovered from clinical and non-clinical settings. *Sci. Rep.* **6**, 24321 (2016).
189. Tang, F. et al. Comparative genomic analysis of 127 *Escherichia coli* strains isolated from domestic animals with diarrhea in China. *BMC Genomics* **20**, 212 (2019).
190. Czeczulin, J. R., Whittam, T. S., Henderson, I. R., Navarro-Garcia, F. & Nataro, J. P. Phylogenetic analysis of enteroaggregative and diffusely adherent *Escherichia coli*. *Infect. Immun.* **67**, 2692–2699 (1999).
191. Boudeau, J., Glasser, A. L., Masseret, E., Joly, B. & Darfeuille-Michaud, A. Invasive ability of an *Escherichia coli* strain isolated from the ileal mucosa of a patient with Crohn's disease. *Infect. Immun.* **67**, 4499–4509 (1999).
192. Mirsepasi-Lauridsen, H. C., Vallance, B. A., Krogfelt, K. A. & Petersen, A. M. *Escherichia coli* pathobionts associated with inflammatory bowel disease. *Clin. Microbiol. Rev.* **32,** e00060 https://doi.org/10.1128/CMR.00060-18 (2019).
193. Martinez-Medina, M. et al. Molecular diversity of *Escherichia coli* in the human gut: new ecological evidence supporting the role of adherent-invasive *E. coli* (AIEC) in Crohn's disease. *Inflamm. Bowel Dis.* **15**, 872–882 (2009).
194. Desilets, M. et al. Genome-based definition of an inflammatory bowel disease-associated adherent-invasive *Escherichia coli* pathovar. *Inflamm. Bowel Dis.* **22**, 1–12 (2016).
195. O'Brien, C. L. et al. Comparative genomics of Crohn's disease-associated adherent-invasive *Escherichia coli*. *Gut* **66**, 1382–1389 (2017).
196. Mora, A. et al. Seropathotypes, phylogroups, *stx* subtypes, and intimin types of wildlife-carried, Shiga toxin-producing *Escherichia coli* strains with the same characteristics as human-pathogenic isolates. *Appl. Environ. Microbiol.* **78**, 2578–2585 (2012).
197. Alonso, C. A. et al. Occurrence and characterization of *stx* and/or *eae*-positive *Escherichia coli* isolated from wildlife, including a typical EPEC strain from a wild boar. *Vet. Microbiol.* **207**, 69–73 (2017).
198. Espinosa, L., Gray, A., Duffy, G., Fanning, S. & McMahon, B. J. A scoping review on the prevalence of Shiga-toxigenic *Escherichia coli* in wild animal species. *Zoonoses Public Health* **65**, 911–920 (2018).
199. Frank, C. et al. Epidemic profile of Shiga-toxin-producing *Escherichia coli* O104:H4 outbreak in Germany. *N. Engl. J. Med.* **365**, 1771–1780 (2011).
200. Bai, X. et al. Molecular characterization and comparative genomics of clinical hybrid Shiga toxin-producing and enterotoxigenic *Escherichia coli* (STEC/ETEC) strains in Sweden. *Sci. Rep.* **9**, 5619 (2019).
201. Hazen, T. H. et al. Comparative genomics and transcriptomics of *Escherichia coli* isolates carrying virulence factors of both enteropathogenic and enterotoxigenic *E. coli*. *Sci. Rep.* **7**, 3513 (2017).
202. Soysal, N. et al. Enterohemorrhagic *Escherichia coli* hybrid pathotype O80:H2 as a new therapeutic challenge. *Emerg. Infect. Dis.* **22**, 1604–1612 (2016).
This paper presents a thorough analysis of a hybrid intestinal and extra-intestinal pathogenic emerging *E. coli* clone rendering obsolete the classical ExPEC–InPEC boundaries.
203. Mariani-Kurkdjian, P. et al. Haemolytic–uraemic syndrome with bacteraemia caused by a new hybrid *Escherichia coli* pathotype. *New Microbes New Infect.* **2**, 127–131 (2014).
204. De Rauw, K. et al. Characteristics of Shiga toxin producing- and enteropathogenic *Escherichia coli* of the emerging serotype O80:H2 isolated from humans and diarrhoeic calves in Belgium. *Clin. Microbiol. Infect.* **25**, 111 e115–111 e118 (2019).
205. Nuesch-Inderbinen, M., Cernela, N., Wuthrich, D., Egli, A. & Stephan, R. Genetic characterization of Shiga toxin producing *Escherichia coli* belonging to the emerging hybrid pathotype O80:H2 isolated from humans 2010–2017 in Switzerland. *Int. J. Med. Microbiol.* **308**, 534–538 (2018).
206. Cointe, A. et al. *Escherichia coli* O80 hybrid pathotype strains producing Shiga toxin and ESBL: molecular characterization and potential therapeutic options. *J. Antimicrob. Chemother.* **75**, 537–542 (2020).
207. Gati, N. S., Middendorf-Bauchart, B., Bletz, S., Dobrindt, U. & Mellmann, A. Origin and evolution of hybrid Shiga toxin-producing and uropathogenic *Escherichia coli* strains of sequence type 141. *J. Clin. Microbiol.* **58**, e01309 https://doi.org/10.1128/JCM.01309-19 (2020).
208. Bielaszewska, M. et al. Heteropathogenic virulence and phylogeny reveal phased pathogenic metamorphosis in *Escherichia coli* O2:H6. *EMBO Mol. Med.* **6**, 347–357 (2014).
209. Kessler, R. et al. Diarrhea, bacteremia and multiorgan dysfunction due to an extraintestinal pathogenic *Escherichia coli* strain with enteropathogenic *E. coli* genes. *Pathog. Dis.* **73**, ftv076 (2015).
210. Page, A. J. et al. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
211. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
212. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
213. Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S. E. & Lercher, M. J. PopGenome: an efficient Swiss army knife for population genomic analyses in R. *Mol. Biol. Evol.* **31**, 1929–1936 (2014).
214. Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online — a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res.* **33**, W557–W559 (2005).
215. Savageau, M. A. *Escherichia coli* habitats, cell types, and moleculat mechanisms of gene control. *Am. Nat.* **122**, 732–744 (1983).
216. Levin, B. R. The evolution and maintenance of virulence in microparasites. *Emerg. Infect. Dis.* **2**, 93–102 (1996).
217. Nowrouzian, F. L., Wold, A. E. & Adlerberth, I. *Escherichia coli* strains belonging to phylogenetic group B2 have superior capacity to persist in the intestinal microflora of infants. *J. Infect. Dis.* **191**, 1078–1083 (2005).
218. Nowrouzian, F. L. et al. *Escherichia coli* B2 phylogenetic subgroups in the infant gut microbiota: predominance of uropathogenic lineages in Swedish infants and enteropathogenic lineages in Pakistani infants. *Appl. Environ. Microbiol.* **85,** e01681 https://doi.org/10.1128/AEM.01681-19 (2019).
219. Schierack, P. et al. ExPEC-typical virulence-associated genes correlate with successful colonization by intestinal *E. coli* in a small piglet group. *Environ. Microbiol.* **10**, 1742–1751 (2008).
220. Nowrouzian, F. L., Adlerberth, I. & Wold, A. E. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes Infect.* **8**, 834–840 (2006).
221. Tourret, J. et al. Small intestine early innate immunity response during intestinal colonization by *Escherichia coli* depends on its extra-intestinal virulence status. *PLoS ONE* **11**, e0153034 (2016).
222. Russell, C. W. et al. Context-dependent requirements for FimH and other canonical virulence factors in gut colonization by extraintestinal pathogenic *Escherichia coli*. *Infect. Immun.* **86,** e00746 https://doi.org/10.1128/IAI.00746-17 (2018).
223. Diard, M. et al. Pathogenicity-associated islands in extraintestinal pathogenic *Escherichia coli* are fitness elements involved in intestinal colonization. *J. Bacteriol.* **192**, 4885–4893 (2010).
224. Sheng, H., Lim, J. Y., Knecht, H. J., Li, J. & Hovde, C. J. Role of *Escherichia coli* O157:H7 virulence factors in colonization at the bovine terminal rectal mucosa. *Infect. Immun.* **74**, 4685–4693 (2006).
225. Fitzgerald, S. F. et al. Shiga toxin sub-type 2a increases the efficiency of *Escherichia coli* O157 transmission between animals and restricts epithelial regeneration in bovine enteroids. *PLoS Pathog.* **15**, e1008003 (2019).
This paper is a very mechanistic demonstration of the concept of 'coincidental evolution'.
226. Adiba, S., Nizak, C., van Baalen, M., Denamur, E. & Depaulis, F. From grazing resistance to pathogenesis: the coincidental evolution of virulence factors. *PLoS ONE* **5**, e11882 (2010).

227. Steinberg, K. M. & Levin, B. R. Grazing protozoa and the evolution of the *Escherichia coli* O157:H7 Shiga toxin-encoding prophage. *Proc. Biol. Sci.* **274**, 1921–1929 (2007).
228. Schmidt, C. E., Shringi, S. & Besser, T. E. Protozoan predation of *Escherichia coli* O157:H7 is unaffected by the carriage of Shiga toxin-encoding bacteriophages. *PLoS ONE* **11**, e0147270 (2016).

Acknowledgements
The authors are grateful to B. Condamine for help with the figure drafts. O.C., S.B. and E.D. are partially supported by the Fondation pour la Recherche Médicale (Equipe FRM 2016, grant number DEQ20161136698).

Author Contributions
E.D. and D.G. conceived and wrote the main part of the Review. O.C. provided numerous reflections and epidemiologic and genomic data for the main text and the figures. S.B. wrote part of the InPEC and hybrid clones sections.

Competing interests
The authors declare no competing interests.

Peer review information
Nature Reviews Microbiology thanks J. Nataro and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Supplementary information
Supplementary information is available for this paper at https://doi.org/10.1038/s41579-020-0416-x.

RELATED LINKS
EnteroBase: https://enterobase.warwick.ac.uk
Harvest: https://www.cbcb.umd.edu/software/harvest