environmental microbiology

Environmental Microbiology (2011) **13**(9), 2468–2477 doi:10.1111/j.1462-2920.2011.02519.x

Characterization of the cryptic *Escherichia* lineages: rapid identification and prevalence

Olivier Clermont,[1,2] David M. Gordon,[3] Sylvain Brisse,[4] Seth T. Walk[5] and Erick Denamur[1,2]*

[1]*INSERM, UMR-S 722, Paris, F-75018, France.*
[2]*Univ Paris Diderot, Sorbonne Paris Cité, UMR-S 722, Faculté de Médecine, Site Xavier Bichat, Paris, F-75018, France.*
[3]*Research School of Biology, Australian National University, Canberra, ACT 0200, Australia.*
[4]*Genotyping of Pathogens and Public Health, Institut Pasteur, 75724 Paris, France.*
[5]*Department of Internal Medicine, University of Michigan Health System, 4618 Medical Science Building II, Ann Arbor, MI 48109, USA.*

Summary

Strains phenotypically indistinguishable from *Escherichia coli* and belonging to at least five distinct cryptic lineages, named *Escherichia* clades I to V, that are genetically divergent from *E. coli* yet members of the genus have been recently found using multi-locus sequence typing (MLST). Very few epidemiological data are available on these strains as their detection by MLST is not suitable for large-scale studies. In this work, we developed a rapid PCR method based on *aes* and *chuA* allele-specific amplifications that assigns a strain a cryptic lineage membership. By screening more than 3500 strains with this approach, we show that the cryptic lineages of *Escherichia* are unlikely to be detected in human faecal samples (2–3% frequency) and even less likely to be isolated from extra-intestinal body sites (< 1% frequency). They are more abundant in animal faeces ranging from 3–8% in non-human mammals to 8–28% in birds. Overall, the strains from the clade V are the most abundant and from the clade II very rare. These results suggest that members of the cryptic clades are unlikely to be of significance to human and health but may influence the use of *'E. coli'* as an indicator of water quality.

Received 2 March, 2011; accepted 2 May, 2011. *For correspondence. E-mail erick.denamur@inserm.fr; Tel. (+33) 1 57 27 75 34; Fax (+33) 1 57 27 75 21.

Introduction

Escherichia coli was, until recently, considered to be the best-known bacterial species. First, because the *E. coli* strain K-12 has been used widely as a model organism in genetic and molecular biology studies (Neidhardt *et al.*, 1996). Second, because population geneticists have used *E. coli* to test a variety of hypotheses with each new technological advance (reviewed in Tenaillon *et al.*, 2010). Third, because *E. coli* is a very significant human and animal pathogen it has been the subject of intense scrutiny by the medical community. In nature, *E. coli* encompasses strains living as commensals inhabiting the vertebrate gut (primary habitat) but also intra-and extra-intestinal pathogenic strains (Kaper *et al.*, 2004; Tenaillon *et al.*, 2010). Soil, water and sediments represent the species' secondary habitat, and *E. coli* cells may be as abundant in the secondary habitat as in the primary habitat (Savageau, 1983). The genetic structure of the species is predominantly clonal, meaning that recombination is rare enough to allow the long-term propagation of clonal lineages (Tenaillon *et al.*, 2010). Phylogenetic analyses initially revealed at least five main phylogenetic groups (A, B1, B2, D and E) (Gordon *et al.*, 2008). Recently, an additional group of strains close to B2 phylogroup strains and called F has been delineated (Jaureguy *et al.*, 2008). At the genomic level, *E. coli* is highly diverse with 10 times more genes in the pan-genome (the total number of genes in the species) than in the core genome (the genes present in all the strains) (Rasko *et al.*, 2008; Touchon *et al.*, 2009). Of the other species of *Escherichia*, *E. fergusonii* is the species genetically most similar to *E. coli* (Lawrence *et al.*, 1991; Walk *et al.*, 2009) and, while less similar, *E. albertii* is also clearly a member of the genus (Hyma *et al.*, 2005; Walk *et al.*, 2009). *Escherichia blattae*, *E. hermanii* and *E. vulneris* share minimal genetic similarity with other *Escherichia* and they should not be considered valid members of the genus (Lawrence *et al.*, 1991; Hartl, 1992; Cilia *et al.*, 1996; Paradis *et al.*, 2005; Pham *et al.*, 2007; Priest and Barker, 2010).

Given that *E. coli* has been the subject of intense study for over a century, it was perhaps surprising when strains phenotypically indistinguishable from *E. coli* were found to be genetically divergent from *E. coli* (Walk *et al.*, 2009). Multi-locus sequence analysis based on 22 core genome

genes revealed that these strains fell into five novel 'cryptic' lineages that were clearly members of the genus, but which are distinct from the three named *Escherichia* species. Clade I is very closely related to *E. coli*, whereas clade V strains are the most divergent. Clades III and IV are sister groups, branched between *E. coli* and clade V. Only one strain of clade II has been reported (Walk *et al*., 2009). It has been suggested that the primary niche of these strains could be habitats outside of the host gastrointestinal tract (Walk *et al*., 2009; Ingle *et al*., 2011), which might explain why they have not been found before, as much of our understanding of *E. coli* is based on faecal and clinical isolates.

We know little of the distribution or the ecological characteristics of these novel *Escherichia* lineages. This lack of understanding is largely due to the fact that, at present, the only method by which members of the cryptic clades can be distinguished from *E. coli* is by obtaining nucleotide sequence data for an informative gene. Consequently, we developed a PCR-based method that distinguishes between strains of *E. coli* and members of the novel clades and which allows most cryptic strains to be assigned to one of the novel clades. Using this approach, we then screened various collections of commensal and pathogenic strains in order to determine their ecological distribution.

Results

Multi-locus sequence typing (MLST) and single gene phylogenies

Characterization of 419 strains isolated from humans and other animals living in France and from animals living in various parts of the world, using the Institute Pasteur MLST scheme (Le Gall *et al*., 2007; Jaureguy *et al*., 2008), revealed 37 *Escherichia* strains that could not be considered typical *E. coli*. In order to determine the clade membership of these non-*E. coli* strains as in (Walk *et al*., 2009), a representative subset of strains was characterized using the Achtman MLST scheme (Wirth *et al*., 2006). Trees based on the concatenated data of each of the MLST schemes, including typical *E. coli*, *E. fergusonii* and *E. albertii* strains and rooted on *Salmonella*, were largely congruent (Fig. 1). Both schemes identified clades I–V as distinct from *E. coli*, with clade I between *E. coli* and *E. fergusonii*. However, the phylogeny inferred using the Achtman MLST scheme indicates that the strain ROAR19 is a clade II strain, while the Institute Pasteur scheme depicts ROAR19 as more closely related to *E. coli* than the clade II strain B1147 (Fig. 1). Further, using the Institute Pasteur scheme, *E. albertii* strains appear as the most basal branch whereas it is the clade V strains that are basal using the Achtman MLST scheme. It can be noted that all the nodes of the Institute Pasteur scheme tree are supported by high bootstrap values, which is not the case with the other scheme (Fig. 1). This could be due in part to the fact that the Institute Pasteur scheme is based on a greater number of nucleotides than the Achtman MLST scheme (7032 versus 3423).

Walk and colleagues (2009) reported that only three genes, *lysP*, *rpoS* and *fumC*, appeared capable of unambiguously assigning all *Escherichia* strains to the appropriate cryptic clade or species of *Escherichia*. Recently, it has been shown that *aes*, the gene encoding esterase B, is a powerful phylogenetic marker of *E. coli* (Lescat *et al*., 2009). We sequenced the *aes* gene in a set of 29 representative non-*E. coli* strains as well as *E. fergusonii* and *E. albertii* strains. The phylogenetic tree inferred using these sequences, as well as with sequences from *E. coli* strains belonging to the five main phylogenetic groups, shows that this gene is able to clearly delineate the five cryptic clades from *E. coli*, *E. albertii* and *E. fergusonii*, with two exceptions. There are two identical clade I strains (H442 and TW10509), identified as such based on the MLST data (Fig. 1), having *aes* sequences that are more like *E. fergusonii aes* sequences rather than the *aes* sequences of other clade I strains (Fig. 2). The strain B2H5 is assigned to *E. coli* based on the MLST data; however, its *aes* gene is clearly more like the *aes* genes of clade I strains than those of *E. coli* (Fig. 2).

It can be concluded from these analyses that various lineages of *Escherichia* are phylogenetically robust and can be identified by classical MLST schemes. It has also been shown that, provided the appropriate gene is chosen, nucleotide sequence data for a single gene can reliably assign an *Escherichia* strain to the appropriate lineage. However, a sequencing approach for epidemiological screening is tedious and relatively expensive. Consequently, the goal was to develop a simple PCR-based approach to both identify strains belonging to the cryptic clades and assign these strains to the appropriate clade.

Preliminary identification of cryptic Escherichia *lineages*

The MLST results identify a group of strains that are unambiguously members of one of the five cryptic clades of *Escherichia*. The characteristics of these strains were used to discover which traits might indicate that a strain in a collection of isolates phenotypically resembling *E. coli* could be a member of one the cryptic clades.

It is now common practice to use the triplex PCR method (Clermont *et al*., 2000) to assign *E. coli* strains to the phylo-groups A, B1, B2 or D. This method is based on the amplification of three DNA fragments, named *chuA*, *yjaA* and TSPE4.C2, belonging to *chuA*, *yjaA* and a putative lipase esterase genes respectively (Clermont *et al*., 2000; Gordon *et al*., 2008). The different combinations of

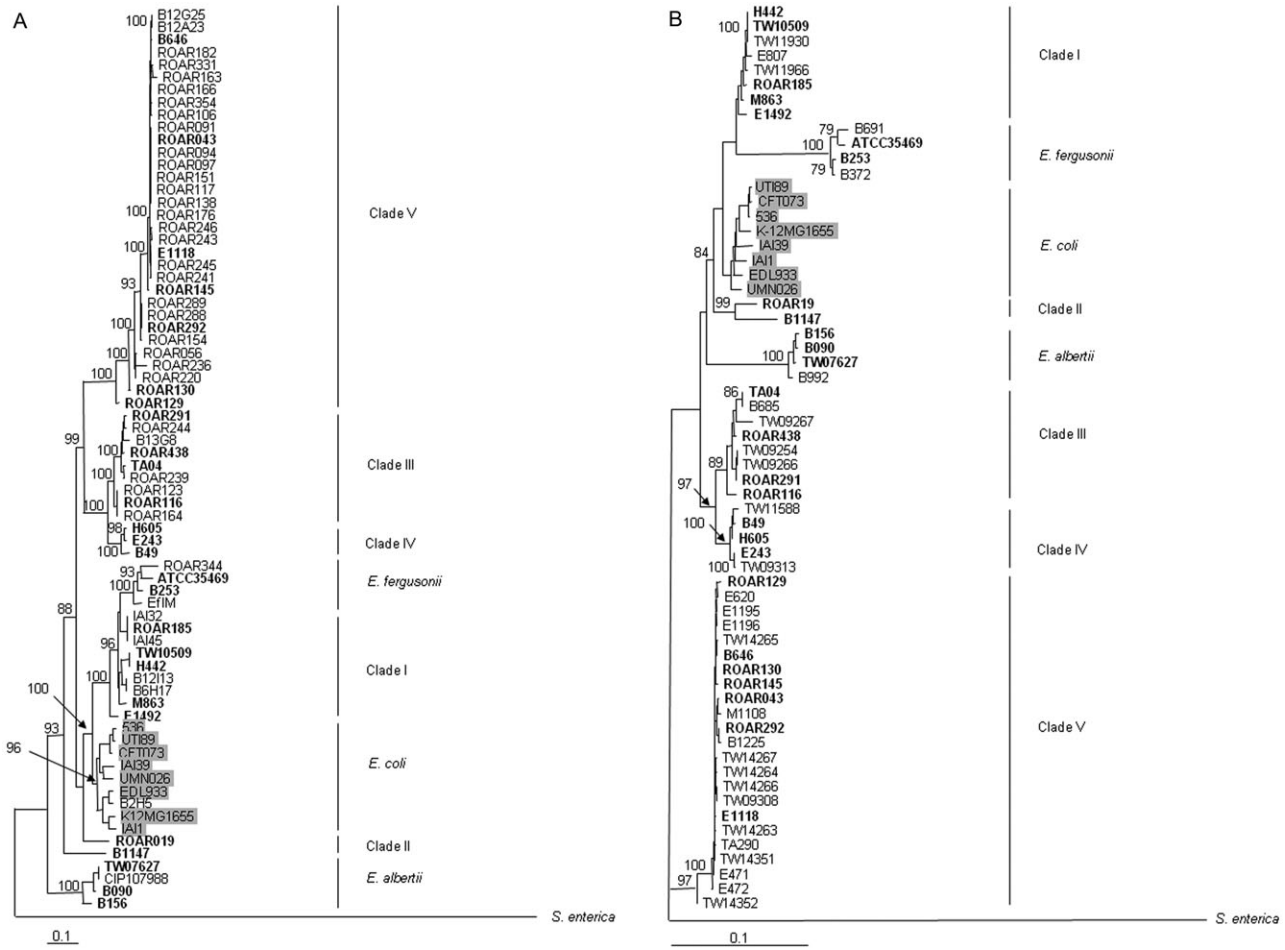

Fig. 1. Phylogenetic trees of Australian and French representative commensal and pathogenic *Escherichia* clade strains as well as representative strains of *E. coli, E. fergusonii* and *E. albertii* reconstructed from (A) 7 complete genes (7032 bp) (Le Gall *et al.*, 2007) and (B) 7 partial genes (3423 bp) (Wirth *et al.*, 2006) using PHYML (Guindon *et al.*, 2005). The trees are rooted on *Salmonella* strains. Bootstrapping was performed on 500 replicates and values above 70 for the major nodes are indicated. Strains common to the *aes* and *chuA* trees (Figs 2 and 3 respectively) are in bold. *E. coli* strains are boxed in grey. The strain B2H5, which has an ambiguous status, is not boxed.

these DNA fragments give rise to 7 genotypes: A_0 and A_1 belonging to the A phylo-group, B1, $B2_2$ and $B2_3$ belonging to the B2 phylo-group and D_1 and D_2 belonging to the D phylo-group (Gordon *et al.*, 2008). The majority (88%) of strains belonging to clades II, III, IV or V failed to yield any of the expected triplex method PCR products (A_0 genotype) (Table 1). Although most clade I strains are positive for *chuA* and *yjaA* ($B2_2$ genotype), clade I strains can have a variety of genotypes (Table 1). Although cryptic clade strains could not be distinguished from *E. coli* strains using biochemical and enzymatic reaction patterns analysed by multidimensional scaling (Walk *et al.*, 2009; Sabarly *et al.*, 2011), we observed specific patterns for lysine and ornithine utilization. The majority of *E. coli* are positive for lysine and ornithine catabolism (Ewing, 1986). Most clade III and IV strains are lysine negative and ornithine positive, while all but one clade V strains are lysine positive and ornithine negative (Table 1). The two clade II strains are also lysine positive and ornithine negative. These traits fail to distinguish clade I strains from *E. coli*.

Some cryptic clade strains yielded a *chuA* product using the triplex PCR method and genome analysis revealed that *chuA* was present in *E. albertii*. These results suggested that *chuA* may be present in all of the cryptic lineages, but divergent from *chuA* of *E. coli*, and consequently the *chuA* PCR was performed at 50°C of annealing [instead of 55°C in the classical Clermont method (Clermont *et al.*, 2000)]. PCR screening using this approach revealed that although some clade I strains (38%) lack *chuA*, all clade III, IV and V strains were *chuA* positive, while the clade II strains are *chuA* negative. A phylogenetic tree inferred using the complete *chuA* sequences of 27 strains from clades I, III, IV and V as well as *E. coli* and rooted on the *E. albertii* sequences indicated that the *chuA* phylogeny is congruent with the phy-

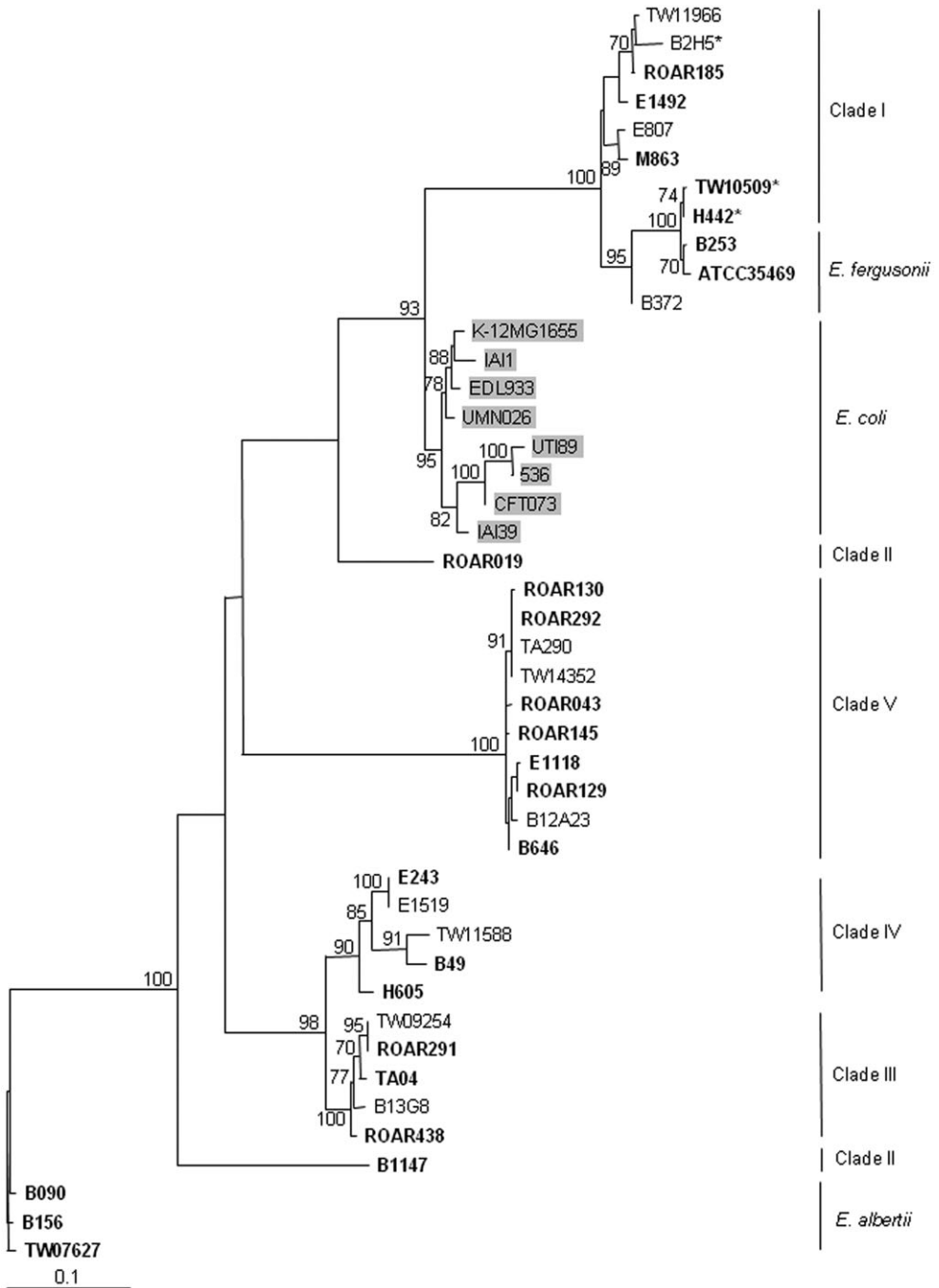

Fig. 2. Phylogenetic tree of *Escherichia* clade, *E. coli* and *E. fergusonii* strains reconstructed from the 894 bp of the *aes* gene using PHYML (Guindon *et al*., 2005). The tree is rooted on the *E. albertii* strains. Bootstrapping was performed on 500 replicates and values above 70 are indicated at the nodes. Strains in bold correspond to the strains present also in Fig. 1 and/or 3. *E. coli* strains are boxed in grey. The strains noted with a star (B2H5, TW10509, H442) correspond to strains having a different position in the MLST trees (Fig. 1).

logenetic history of the strains (Fig. 3), as previously observed for *E. coli* (Gordon *et al*., 2008). The clade I strains except E1492 are monophyletic, however they fall within *E. coli* between the B2/F and D/E phylo-group strains.

A rapid PCR method to assign a strain a cryptic clade membership

The complete *aes* and *chuA* nucleotide sequence data were used to develop an allele-specific PCR of the *aes*

Table 1. PCR triplex genotypes and metabolic properties of strains belonging to the cryptic lineages of *Escherichia* isolated from Australia, France and other parts of the world.

Trait	Clade I ($n = 13$)	Clade II ($n = 2$)	Clade III ($n = 14$)	Clade IV ($n = 6$)	Clade V ($n = 52$)
Triplex PCR[a]					
chuA yjaA TSPE4.C2 genotype					
– – – A_0	0	1	12	4	48
– + – A_1	5	1	0	0	0
– – + B1	0	0	1	1	3
+ + + $B2_3$	1	0	0	0	1
+ + – $B2_2$	7	0	0	0	0
+ – – D_1	0	0	1	1	0
+ – + D_2	0	0	0	0	0
Catabolism					
lysine, ornithine					
+ +	10	0	0	1	1
+ –	1	2	0	0	51
– +	2	0	11	4	0
– –	0	0	3	1	0

a. As in Clermont and colleagues (2000).
+, presence of the PCR product; –, absence of the PCR product.

and *chu*A genes based on clade-specific single nucleotide polymorphisms that allows the amplification of PCR products of different sizes according to the *Escherichia* clade (clades I and II and clades III, IV and V respectively). The primers and the lengths of the PCR products are given Table 2 and an example of the PCR products migrated in an agarose gel is shown in Fig. 4. The robustness of this method, first developed in Erick Denamur's lab (France) was also tested in the David Gordon's lab (Australia) on a panel of strains previously characterized with the MLST data. In total, 14 clade III, 6 clade IV and 52 clade V strains yield *chuA* PCR products of the expected size, while 13 clade I and 2 clade II strains yielded *aes* products of the expected size. No positive signal was obtained using these allele-specific primers neither in a collection of 11 *E. albertii* strains nor in the ECOR collection strains (Ochman and Selander, 1984) that are representative of the *E. coli* genetic diversity. However, the three *E. fergusonii* strains tested yielded a product using the *aes* clade I primers. This outcome was expected given the close similarity of *aes* for clade I and *E. fergusonii* strains (Fig. 2).

According to the classical triplex PCR results of the clade strains (Table 1), we propose a two-step screening strategy based first on the triplex PCR amplifying *chuA*, *yjaA* and TSPE4.C2 (Clermont *et al*., 2000) followed by the allele-specific PCR on A_0 (*chuA*, *yjaA* and TSPE4.C2 negative) and $B2_2$ (*chuA* and *yjaA* positive and TSPE.4C2 negative) genotypes.

Relative abundance of strains belonging to the cryptic clades

The MLST analysis of bacterial isolates identified as '*E. coli*' based on classic phenotypic characteristics should provide an unbiased estimate of the fraction of strains that are not *E. coli*, but rather members of one of the cryptic clades. The MLST data indicate that the relative abundance of members of the cryptic clades in an '*E. coli*' collection will depend on the source of the isolate (Table 3). The frequency of strains belonging to the cryptic clades ranged from 0% among extra-intestinal isolates to 28% for isolates from birds. There also appears to be a locality effect, as the frequency of the cryptic clade strains in the collection of French isolates was 8.8% but only 4.4% in the collection of Australian isolates (contingency analysis, likelihood ratio $\chi^2 = 6.74$, $P > \chi^2 = 0.009$).

The rarity of strains belonging to the cryptic lineages among extra-intestinal isolates from humans was confirmed when strains from two collections of largely clinical isolates were screened using the allele-specific PCR approach: 65 strains isolated in the 1980s from patients living in France (Picard *et al*., 1999) and 1081 strains isolated from septicaemic patients in 2005 during the COLIBAFI study in France (Lefort *et al*., 2011). The 1980's collection yielded five strains with an A_0 profile and 4 $B2_2$ strains and two of these strains were members of clade I. The COLIBAFI collection contained 98 A_0 and 56 $B2_2$ strains, two of these were clade I strains, two were clade V and one a member of clade III. Thus, the frequency of strains belonging to the cryptic lineages in these two collections was 3.1% and 0.5% respectively.

One of the $B2_2$ strain (B2H5) from the COLIBAFI collection, screened using the clade specific *chuA* primers, yielded a PCR product with the size expected for a clade I strain. The nucleotide sequence of the *chuA* gene for B2H5 confirmed that this strain clustered with other clade I strains (Fig. 3). The nucleotide sequence of the *aes* gene for B2H5 also would suggest that it is a clade I strain (Fig. 2). However, the MLST results for this strain has it

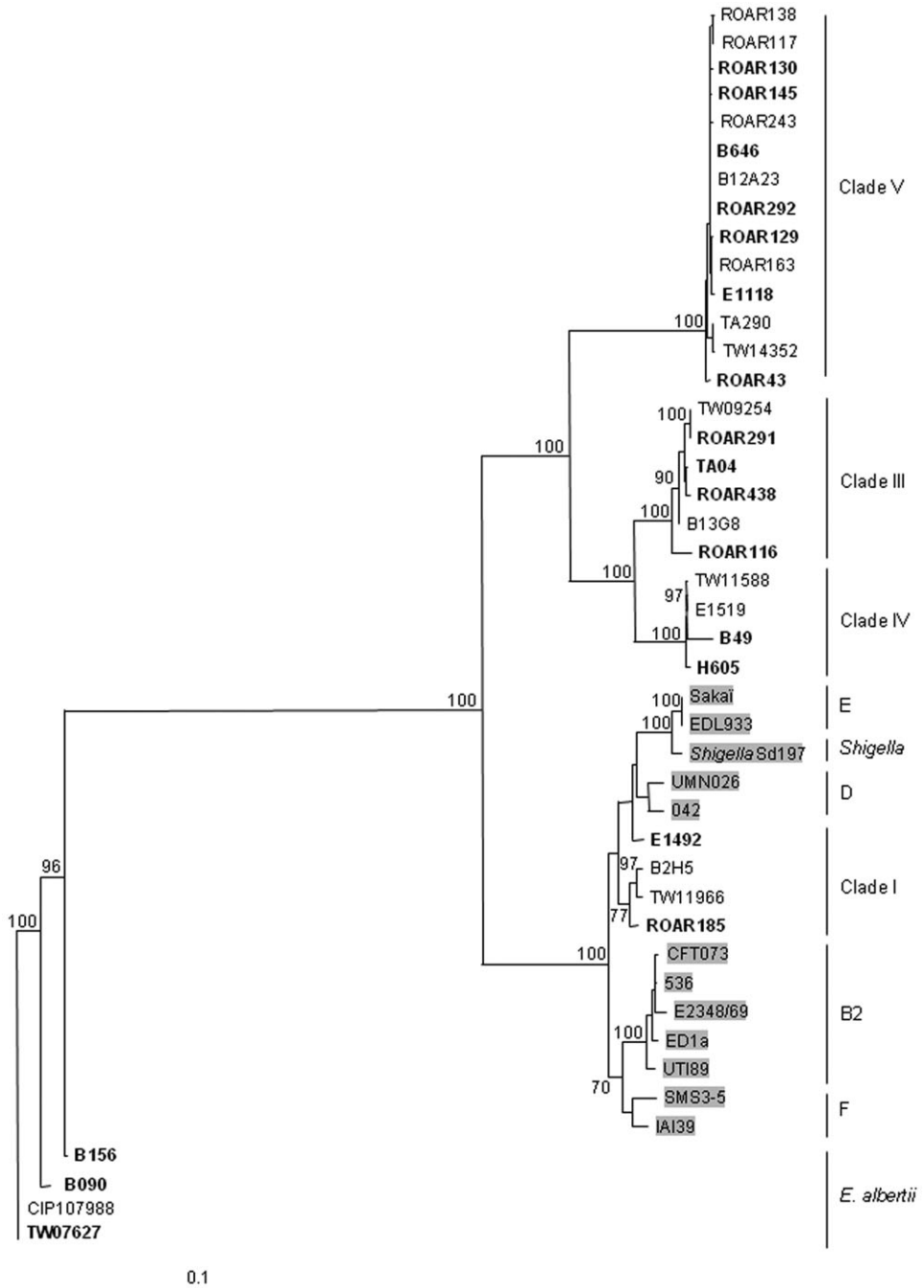

Fig. 3. Phylogenetic tree of *Escherichia* clade, *E. coli* and *E. fergusonii* strains reconstructed from the 1983 bp of the complete *chuA* gene using PHYML (Guindon *et al*., 2005). The tree is rooted on the *E. albertii* strains. Bootstrapping was performed on 500 replicates and values above 70 are indicated at the nodes. Strains in bold correspond to the strains present also in Fig. 1 and/or 2. *E. coli/Shigella* strains are boxed in grey. The strain B2H5, which has an ambiguous status, is not boxed.

clearly clustering with *E. coli* and it is most similar to a phylo-group E strain (Fig. 1A).

Further screening of Australian isolates identified as *E. coli* using the clade specific PCR revealed additional members of the cryptic *Escherichia* lineages. These PCR screening data combined with the MLST results allow the relative abundance of strains of the various cryptic lineages to be assessed in France and Australia. Strains belonging to clade II are rare, while clade V strains are the most common of the cryptic clades (Table 4). Strains belonging to the other clades are of intermediate abundance and their relative frequency may vary with locality,

Table 2. *chuA* and *aes* primers used in the study for the allele-specific PCR amplifications.

Primer designation	Primer sequence	Target	Size of PCR product (bp)
aesI.1	5′-CCTCTACTCACCCAAAAGTC-3′	*aes*	315
aesI.2	5′-ATCACGTAACCACAACGCAC-3′		
aesII.1	5′-CGCCTGTTGTCACTTCCACG-3′	*aes*	125
aesII.2	5′-GTTTATCACGCAGCCACAAG-3′		
chuIII.1	5′-GTGTTGAGATTGTCCGTGGG-3′	*chuA*	183
chuIII.2	5′-CAAAAGCACTGGCGCCCAG-3′		
chuIV.1	5′-CTGGCGAAAGGAACCTGGA-3′	*chuA*	461
chuIV.2	5′-GTTATCTCATCTTGCAGCCAA-3′		
chuV.1	5′-ACTGTATGGCAGTGGCGCAT-3′	*chuA*	600
chuV.2	5′-GCAAAACTATCGGCAAACAGC-3′		

as it appears that clade IV strains are less common in France than clade III strains while the opposite is true for isolates from Australia (contingency analysis, likelihood ratio $\chi^2 = 11.93$, $P < \chi^2 = 0.018$) (Table 4).

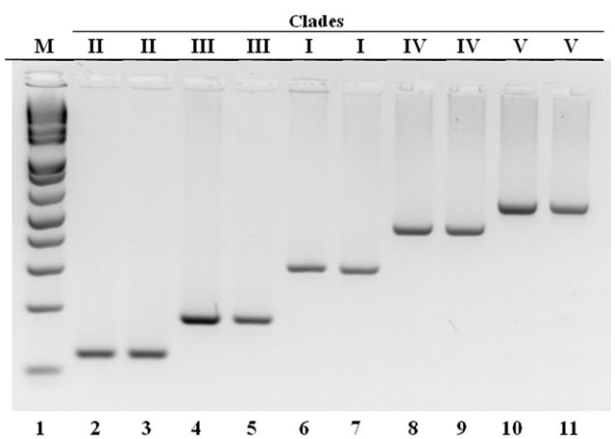

Fig. 4. Allele-specific PCR amplifications of the *Escherichia* clade I–V strains. The PCR products were loaded on a 2% agarose gel. Lane 1 (MW): molecular weight marker (1 kb Plus DNA Ladder, Invitrogen); lanes 2 and 3 (Clade II strains): ROAR19 and B1147 respectively; lanes 4 and 5 (Clade III strains): ROAR116 and ROAR438 respectively; lanes 6 and 7 (Clade I strains): IAI32 and ROAR185 respectively; lanes 8 and 9 (Clade IV strains): B49 and E243 respectively; lanes 10 and 11 (Clade V strains): ROAR43 and ROAR129 respectively.

Discussion

Although *E. fergusonii* and *E. albertii* strains are quite easily distinguished from *E. coli* (Farmer *et al.*, 1985; Oaks *et al.*, 2010), there is no known suite of phenotypic traits capable of distinguishing among all *Escherichia* lineages. Consequently, the recently discovered cryptic lineages of *E. coli* can only be discriminated from each other and from *E. coli* using genetic data. Although, nucleotide sequence data for a single gene, such as *fumC* and *aes*, can determine the lineage membership of an isolate, such an approach is not suitable for large-scale studies. The PCR-based screening method offers a simple method of determining if an isolate with the phenotypic characteristics of *E. coli* is actually a member of one of the cryptic *Escherichia* clades and to what clade it belongs. The method appears to be very accurate for strains belonging to clades II–V.

The method for clade I identification may yield false positives as suggested by the results obtained for the strain B2H5. The MLST data (7032 bp) clearly indicate B2H5 is an *E. coli* strain (Fig. 1), while the nucleotide sequence data for *chuA* and *aes* (2877 bp for both genes) would suggest that it is a clade I strain (Figs 2 and 3). Walk and colleagues (2009) found evidence of allele sharing between clade I strains and *E. fergusonii*, as we observed for the H442 and TW10509 *aes* alleles (Fig. 2),

Table 3. The frequency of strains belonging to the cryptic lineages of *Escherichia* as determined by multi-locus sequence analysis.

	France		Australia	
Source	Number examined	Number of clade members (%)	Number examined	Number of clade members (%)
Soil, water, sediment	0	ND	99	8 (8.1)
Fish/reptiles	0	ND	29	0 (0)
Mammals	279	23 (8.2)	124	4 (3.2)
Birds	39	11 (28.2)	77	6 (7.8)
Human (faecal)	101	2 (2.0)	60	2 (3.3)
Human (clinical)[a]	0	ND	58	0 (0)
Source effect[b]	$P > \chi^2 < 0.001$		$P > \chi^2 = 0.024$	

a. Extra-intestinal isolates.
b. Contingency analysis, likelihood ratio χ^2.

Table 4. Relative abundance of the five cryptic lineages of *Escherichia*.

Clade	France *n* (%)	Australia *n* (%)
Clade I	6 (13.6)	5 (19.2)
Clade II	0	1 (3.8)
Clade III	8 (18.2)	2 (7.7)
Clade IV	0	4 (15.4)
Clade V	30 (68.2)	14 (53.8)

and a recent paper by Luo and colleagues (2011) demonstrates that more recombination events are detected occurring between strains belonging to clade I and *E. coli* than between clade I strains and strains belonging to the other *Escherichia* clades.

Although, it would be ideal if all strains with a phenotype resembling that of *E. coli* were screened using the new PCR method, a more cost-effective approach would be to restrict the screening to those strains yielding the triplex PCR method genotypes of A_0 and $B2_2$. In the absence of any triplex PCR method data, then all strains that are not both lysine and ornithine positive should be screened. The problems associated with correctly identifying clade I strains and the small number of clade II representatives would suggest that all strains identified as clades I and II using the clade specific PCR method should have their identification confirmed using one of the MLST typing schemes.

The available data indicate that strains belonging to any of the cryptic lineages of *Escherichia* are unlikely to be detected in human faecal samples (2–3% frequency) and even less likely to be isolated from extra-intestinal body sites (< 1% frequency), in agreement with the lack of intrinsic extra-intestinal virulence exhibited by these strains (Ingle *et al.*, 2011). However, the samples reported here are from humans living in France and Australia. In humans, the relative abundance of strains belonging to the four main *E. coli* phylo-groups varies substantially with locality (Tenaillon *et al.*, 2010). Further we have shown that when the *Escherichia* isolates are from birds and mammals they are more likely to be members of the cryptic clades if the hosts were sampled in France as compared with Australia. Consequently, members of the cryptic clades may be more prevalent in humans sampled in other countries.

Escherichia coli is the most common member of the *Enterobacteriaceae* to be isolated from mammals (Gordon and FitzGibbon, 1999); however, members of the cryptic clades can approach frequencies of 10% for isolates from non-human mammals. *Escherichia coli* is not prevalent in most species of birds (Gordon and Cowling, 2003), but members of the cryptic clades appear to be relatively common in birds, in both France and Australia. Walk and colleagues (2009) argued that members of the cryptic clades might be more prevalent in water samples than in the mammalian host population, and the results of this study provide some support for this suggestion. In Australia, the cryptic clade strains are more prevalent in water samples than in faecal samples from mammals (Table 3).

These results suggest that members of the cryptic clades are unlikely to be of significance to human and health. However, the cryptic clades appear to inhabit ecological niches that may be distinct from that of most *E. coli* strains. Consequently, the cryptic clades are of significance to studies of population genetics and evolution that concern *E. coli* and related species. The extent to which the existence of these cryptic lineages may influence the use of *'E. coli'* as an indicator of water quality is unknown. Further studies concerning the distribution, ecological and virulence characteristics of members of the cryptic *Escherichia* clades are required. It is hoped that the rapid screening method described here will lead to a better understanding of the significance of these cryptic *Escherichia* lineages.

Experimental procedures

Bacterial strains

MLST data were available for 419 strains characterized using the MLST scheme described at http://www.pasteur.fr/mlst/ (Jaureguy *et al.*, 2008) modified as in Le Gall and colleagues (2007) and for 447 isolates characterized using the MLST scheme described at http://mlst.ucc.ie. (Wirth *et al.*, 2006). Strains in these collections were identified as '*E. coli*' using classical biochemical tests. The collection of 419 strains consisted of 101 isolates taken from the faeces of humans living in France as well as 279 isolates from mammals and 39 isolates from birds living in various parts of the world (Skurnik *et al.*, 2006). The collection of 447 strains were obtained from hosts living in Australia and consisted of 60 faecal and 58 extra-intestinal isolates from humans, together with 124 faecal isolates from mammals, 77 faecal isolates from birds, 29 faecal isolates from reptiles, and 99 isolates from soil sediment or water samples (Gordon *et al.*, 2008). The MLST analysis allowed members of the cryptic *Escherichia* clades to be identified (Walk *et al.*, 2009) and these strains served as the positive controls for the development of the rapid PCR-based screening methods whereas the others served as negative controls. The strains described by Walk and colleagues (2009) were also included in this study.

To gain a better understanding of the distribution and abundance of strains belonging to the cryptic clades of *Escherichia*, several previously published collections of strains identified as *E. coli* by classical biochemical tests were screened using the rapid PCR screening protocols. (i) A collection of 15 faecal and 67 extra-intestinal isolates collected in France during the 1980s (Picard *et al.*, 1999). (ii) A collection of 1081 strains isolated from septicaemic patients in 2005 during the COLIBAFI study in France (Lefort *et al.*,

2011). (iii) The Australian strains used in the MLST analyses were drawn from a much larger collection of strains obtained from hosts living in Australia that consisted of 266 faecal isolates and 353 extra-intestinal isolates from humans (Gordon *et al*., 2005), 12 faecal isolates from fish, 37 faecal isolates from reptiles, 134 faecal isolates from birds and 497 faecal isolates from mammals (Gordon and Cowling, 2003), together with 239 isolates from soil, sediment and water samples (Power *et al*., 2005). The phylo-group membership of all strains in these collections had been determined using the PCR triplex method (Clermont *et al*., 2000).

In addition, the strains of the ECOR collection (Ochman and Selander, 1984) and 6 *E. fergusonii*, 12 *E. albertii* and 2 *Salmonella* strains were used.

chuA *and* aes *gene phylogenies*

The complete *chuA* gene was PCR amplified and sequenced as in Gordon and colleagues (2008). The *aes* gene was PCR amplified and sequenced as described by Lescat and colleagues (2009). Phylogenetic analysis was performed with the maximum likelihood method, as implemented in the PHYML program (Guindon *et al*., 2005), with *E. alberti* as an outgroup.

Allele-specific chuA *and* aes *PCR detection of the* Escherichia *clades*

PCR reaction was carried out in a 20 μl volume containing 2 μl of 10 × buffer (supplied with *Taq* polymerase), 20 pmol of each primer, 2 μM each dNTP, 1 U of *Taq* polymerase (New England Biolabs, Ozyme, St Quentin-en-Yvelines, France), and 3 μl of bacterial lysate or 2 μl of DNA. PCR was performed with an Eppendorf Mastercycler with MicroAm tubes in the following conditions: denaturation 4 min at 94°C, 30 cycles of 5 s at 94°C and 30 s at 63°C, and a final extension step of 5 min at 72°C. All the primers were used in the same reaction (Table 2). PCR products were loaded on 2% agarose gel with SYBR Safe DNA gel stain (Invitrogen, Cergy Pontoise, France). After electrophoresis, gels were photographed under UV light.

Acknowledgements

We are grateful to Bertrand Picard, David Skurnik and Marie Picque for their participation to the gathering and first-line characterization of some of the French collections. This work has been partially funded by the Alliance for the Prudent Use of Antibiotics (APUA) in the frame of the Reservoirs of Antibiotic Resistance (ROAR) projects 2006–2007 obtained by ED and DMG.

References

Cilia, V., Lafay, B., and Christen, R. (1996) Sequence heterogeneities among 16S ribosomal RNA sequences, and their effect on phylogenetic analyses at the species level. *Mol Biol Evol* **13:** 451–461.

Clermont, O., Bonacorsi, S., and Bingen, E. (2000) Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl Environ Microbiol* **66:** 4555–4558.

Ewing, W.H. (1986) *Edwards and Ewing's Identification of Enterobacteriaceae*. New York, NY, USA: Elsevier Science Publishing Co.

Farmer, J.J., 3rd, Fanning, G.R., Davis, B.R., O'Hara, C.M., Riddle, C., Hickman-Brenner, F.W., *et al*. (1985) *Escherichia fergusonii* and *Enterobacter taylorae*, two new species of *Enterobacteriaceae* isolated from clinical specimens. *J Clin Microbiol* **21:** 77–81.

Gordon, D.M., and Cowling, A. (2003) The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* **149:** 3575–3586.

Gordon, D.M., and FitzGibbon, F. (1999) The distribution of enteric bacteria from Australian mammals: host and geographical effects. *Microbiology* **145** (Pt 10)**:** 2663–2671.

Gordon, D.M., Stern, S.E., and Collignon, P.J. (2005) Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology* **151:** 15–23.

Gordon, D.M., Clermont, O., Tolley, H., and Denamur, E. (2008) Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ Microbiol* **10:** 2484–2496.

Guindon, S., Lethiec, F., Duroux, P., and Gascuel, O. (2005) PHYML Online – a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33:** W557–W559.

Hartl, D.L. (1992) Population genetics of microbial organisms. *Curr Opin Genet Dev* **2:** 937–942.

Hyma, K.E., Lacher, D.W., Nelson, A.M., Bumbaugh, A.C., Janda, J.M., Strockbine, N.A., *et al*. (2005) Evolutionary genetics of a new pathogenic *Escherichia* species: *Escherichia albertii* and related *Shigella boydii* strains. *J Bacteriol* **187:** 619–628.

Ingle, D.J., Clermont, O., Skurnik, D., Denamur, E., Walk, S.T., and Gordon, D.M. (2011) Biofilm formation by and thermal niche and virulence characteristics of *Escherichia* spp. *Appl Environ Microbiol* **77:** 2695–2700.

Jaureguy, F., Landreau, L., Passet, V., Diancourt, L., Frapy, E., Guigon, G., *et al*. (2008) Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* **9:** 560.

Kaper, J.B., Nataro, J.P., and Mobley, H.L. (2004) Pathogenic *Escherichia coli*. *Nat Rev Microbiol* **2:** 123–140.

Lawrence, J.G., Ochman, H., and Hartl, D.L. (1991) Molecular and evolutionary relationships among enteric bacteria. *J Gen Microbiol* **137:** 1911–1921.

Le Gall, T., Clermont, O., Gouriou, S., Picard, B., Nassif, X., Denamur, E., and Tenaillon, O. (2007) Extraintestinal virulence is a coincidental by-product of commensalism in B2 phylogenetic group *Escherichia coli* strains. *Mol Biol Evol* **24:** 2373–2384.

Lefort, A., Panhard, X., Clermont, O., Woerther, P.L., Branger, C., Mentre, F., *et al*. (2011) Host factors and portal of entry outweigh bacterial determinants to predict the severity of *Escherichia coli* bacteremia. *J Clin Microbiol* **49:** 777–783.

Lescat, M., Hoede, C., Clermont, O., Garry, L., Darlu, P., Tuffery, P., *et al*. (2009) *aes*, the gene encoding the esterase B in *Escherichia coli*, is a powerful phylogenetic marker of the species. *BMC Microbiology* **9:** 723.

Luo, C., Walk, S.T., Gordon, D.M., Feldgarden, M., Tiedje, J.M., and Konstantinidis, K.T. (2011) 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.

Neidhardt, F.C., Curtiss, R., III, Ingraham, J.L., Lin, E.C.C., Low, K.B., Magasanik, B., *et al*. (1996) *Escherichia Coli and Salmonella: Cellular and Molecular Biology.* Washington, DC, USA: ASM Press.

Oaks, J.L., Besser, T.E., Walk, S.T., Gordon, D.M., Beckmen, K.B., Burek, K.A., *et al*. (2010) *Escherichia albertii* in wild and domestic birds. *Emerg Infect Dis* **16:** 638–646.

Ochman, H., and Selander, R.K. (1984) Standard reference strains of *Escherichia coli* from natural populations. *J Bacteriol* **157:** 690–693.

Paradis, S., Boissinot, M., Paquette, N., Belanger, S.D., Martel, E.A., Boudreau, D.K., *et al*. (2005) Phylogeny of the *Enterobacteriaceae* based on genes encoding elongation factor Tu and F-ATPase beta-subunit. *Int J Syst Evol Microbiol* **55:** 2013–2025.

Pham, H.N., Ohkusu, K., Mishima, N., Noda, M., Monir Shah, M., Sun, X., *et al*. (2007) Phylogeny and species identification of the family *Enterobacteriaceae* based on *dnaJ* sequences. *Diagn Microbiol Infect Dis* **58:** 153–161.

Picard, B., Garcia, J.S., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., *et al*. (1999) The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. *Infect Immun* **67:** 546–553.

Power, M.L., Littlefield-Wyer, J., Gordon, D.M., Veal, D.A., and Slade, M.B. (2005) Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environ Microbiol* **7:** 631–640.

Priest, F.G., and Barker, M. (2010) 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.

Rasko, D.A., Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F., Gajer, P., *et al*. (2008) The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* **190:** 6881–6893.

Sabarly, V., Bouvet, O., Glodt, J., Clermont, O., Skurnik, D., Diancourt, L., *et al*. (2011) The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity. *J Evol Biol* (in press): doi: 10.1111/j.1420-9101.2011.02287.x.

Savageau, M.A. (1983) *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *Am Nat* **122:** 732–744.

Skurnik, D., Ruimy, R., Andremont, A., Amorin, C., Rouquet, P., Picard, B., and Denamur, E. (2006) Effect of human vicinity on antimicrobial resistance and integrons in animal faecal *Escherichia coli*. *J Antimicrob Chemother* **57:** 1215–1219.

Tenaillon, O., Skurnik, D., Picard, B., and Denamur, E. (2010) The population genetics of commensal *Escherichia coli*. *Nat Rev Microbiol* **8:** 207–217.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., *et al*. (2009) Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet* **5:** e1000344.

Walk, S.T., Alm, E.W., Gordon, D.M., Ram, J.L., Toranzos, G.A., Tiedje, J.M., and Whittam, T.S. (2009) Cryptic lineages of the genus *Escherichia*. *Appl Environ Microbiol* **75:** 6534–6544.

Wirth, T., Falush, D., Lan, R., Colles, F., Mensa, P., Wieler, L.H., *et al*. (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60:** 1136–1151.