Environmental Microbiology (2008) **10**(10), 2484–2496 doi:10.1111/j.1462-2920.2008.01669.x

Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method

David M. Gordon,[1] Olivier Clermont,[2] Heather Tolley[1] and Erick Denamur[2]
[1]*School of Botany and Zoology, Australian National University, Canberra ACT 0200, Australia.*
[2]*Institut National de la Santé et de la Recherche Médicale (INSERM) U722 and Faculté de Médecine Xavier Bichat, Université Paris 7 Denis Diderot, 75018 Paris, France.*

Summary

It is well recognized that *Escherichia coli* consists of a number of distinct phylo-groups and that strains of the different phylo-groups vary in their ecological niches, life-history characteristics and propensity to cause disease. Consequently, much can be learnt by assigning a strain of *E. coli* to one of the recognized phylo-groups. A triplex PCR-based method that enables strains of *E. coli* to be assigned to a phylo-group using a dichotomous key approach based on the presence or absence of two genes (*chuA* and *yjaA*) and an anonymous DNA fragment (TSPE4.C2) has been developed. However, the accuracy with which this method assigns strains to their correct phylo-group has not been adequately evaluated. Consequently, 662 strains of *E. coli* were characterized using a multi-locus sequence typing approach. Unsupervised population assignment algorithms were used to assign strains to phylo-groups based on the multi-locus sequence typing data. The analyses revealed that 85–90% of *E. coli* strains can be assigned to a phylo-group and that 80–85% of the phylo-group memberships assigned using the Clermont method are correct. However, the accuracy with which strains are assigned to the correct phylo-group depends on their Clermont genotype. For example, strains yielding a Clermont genotype consistent with phylo-groups B1 and B2 are assigned correctly 95% of the time. Strains failing to yield any PCR products using the Clermont method are seldom members of phylo-group A and strains with such a genotype should not be assigned to a phylo-group.

Received 5 December, 2007; accepted 21 April, 2008. *For correspondence. E-mail David.Gordon@anu.edu.au; Tel. (+61) 26125 3552; Fax (+61) 26125 5573.

Introduction

The existence of distinct phylo-groups or 'subspecies' within *Escherichia coli* has long been acknowledged (Ochman and Selander, 1984; Selander *et al*., 1987; Herzer *et al*., 1990; Desjardins *et al*., 1995; Wirth *et al*., 2006). Currently, there are four well-recognized phylo-groups and these have been designated A, B1, B2 and D. Groups A and B1 are considered to be sister groups and group B2 is considered by some to represent the 'ancestral lineage' of *E. coli* (Lecointre *et al*., 1998). Strains of the four groups differ in their phenotypic characteristics, including their ability to exploit different sugars, their antibiotic-resistance profiles and their growth rate–temperature relationships (Gordon, 2004). Genome size varies among the four phylo-groups with A and B1 strains having smaller genomes than B2 or D strains (Bergthorsson and Ochman, 1998). The distribution (presence/absence) of a variety of genes thought to enable a strain to cause extra-intestinal disease also varies among strains of the four phylo-groups (Johnson *et al*., 2001).

Strains of the four phylo-groups also appear to differ in their ecological niches, life-history characteristics and propensity to cause disease. For example, groups B2 and D strains are less frequently isolated from the environment (Walk *et al*., 2007) or fish, frogs and reptiles than A or B1 strains (Gordon and Cowling, 2003). In mammals, B2 strains are more frequently isolated from hosts possessing hindgut modifications for microbial fermentation than strains of the other phylo-groups (Gordon and Cowling, 2003). B2 strains have been shown to persist for longer periods in infants than other strains of *E. coli* (Nowrouzian *et al*. 2006). Finally, isolates recovered from extra-intestinal body sites are more likely to be B2 or D strains than to be A or B1 strains (Gordon, 2004). Thus, a great deal can be learnt concerning the characteristics of an unknown strain by determining its phylo-group membership.

Clermont and colleagues (2000) developed a multiplex PCR-based method that enables strains of *E. coli* to be

assigned to a phylo-group using a dichotomous key approach based on the presence or absence of two genes (*chuA* and *yjaA*) and an anonymous DNA fragment (TSPE4.C2). To date, the method has been used in over 150 population-level studies of *E. coli*. The utility of the Clermont method was validated as part of the original study. However, the validation process was based on relatively few strains, largely collected from humans or human-associated animals. Since the method appeared in the literature, only one study has commented on the frequency with which strain are correctly assigned using the Clermont phylo-grouping method (Walk *et al*., 2007), and no study has been specifically undertaken to validate the method. Here we use multi-locus sequence typing (MLST) data for 662 isolates of *E. coli* together with a number of population assignment and clustering algorithms to determine the frequency with which *E. coli* strains are appropriately assigned to the well-recognized phylo-groups when using the Clermont method.

Results

In addition to the 72 strains in the *E. coli* reference (ECOR) collection (Ochman and Selander, 1984), strains from two other collections underwent MLST. The first of these was a collection of 153 strains isolated from the faeces of a diversity of host species and a variety of geographical locations, as well as from patients exhibiting different clinical syndromes. These strains were characterized at six loci using a modified version of the MLST scheme described at http://www.pasteur.fr/mlst (French MLST scheme). The second collection consisted of 437 strains isolated from humans and non-human vertebrates living in Australia and from Australian soil, sediment and water samples. Strains in the Australian collection were characterized using the MLST scheme described at http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli (German MLST scheme). The ECOR collection strains and 19 others were typed using both MLST schemes. Two unsupervised population assignment algorithms were used to assign strains to phylo-groups: BAPS (Corander and Marttinen, 2006; Corander and Tang, 2007) and STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003).

The Clermont method has the potential to yield eight distinct phylo-types (Table 1). The phylo-type −++ (*chuA*–, *yjaA*+, TSPE4.C2+) was never observed among the 662 strains characterized in this study and the phylo-group B2 phylo-type ++− (*chuA*+, *yjaA*+, TSPE4.C2–) was rare (Table 1).

The first analysis considered the 72 strains in the ECOR collection and 19 other strains that had been characterized using both the French and German MLST schemes. The nucleotide sequence data consisted of 9446 nt from 13 genes and yielded 762 informative sites.

Table 1. Phylo-types resulting from the application of the Clermont method (Clermont *et al*., 2000), the *E. coli* phylo-group they represent, together with the frequency with which they were observed among the 662 strains characterized for this study.

chuA	*yjaA*	TSPE4.C2	Phylo-group assignment	% frequency
−	−	−	A	7.5
−	+	−	A	14.4
−	−	+	B1	28.9
+	+	−	B2	3.0
+	+	+	B2	27.7
+	−	−	D	13.6
+	−	+	D	4.9

The relationships among the strains were depicted using a neighbour-joining (NJ) tree (Fig. 1).

Both population assignment algorithms found that assuming five populations provided the best fit to the data (Table 2). The populations recognized corresponded to phylo-groups A, B1, B2 and D, as well as previously suggested phylo-group designated as E, and that is represented by the strain ECOR 37 (Selander *et al*., 1987; Escobar-Páramo *et al*., 2004a). Only 11% of the strains could not be unambiguously assigned to phylo-groups.

The Clermont method correctly assigned 85% of the 91 strains to the correct phlyo-group (Table 3). No strains with a Clermont B1 (−−+) or B2 phylo-type (+++ or ++−) were incorrectly assigned. Of the 22 strains with a Clermont A phylo-type of −+−, 9% were mis-classified and all of the strains with the A phylo-type −−− could not be assigned to a phylo-group. Of the 17 strains with a Clermont D phylo-type (+−+ or +−−), 76% were correctly identified as D strains, while the balance were unassigned or assigned to group E.

French MLST scheme and strain collection

STRUCTURE assigned the 225 strains characterized using the French MLST scheme to one of five populations,

Table 2. Results of the population assignment algorithms STRUCTURE and BAPS when applied to the 91 strains characterized using the French and German MLST schemes (762 informative sites from 13 genes).

STRUCTURE	BAPS	Number of strains
A	A	20
B1	B1	23
B2	B2	23
D	D	13
E	E	3
U	A	7
U	B1	1
B1	U	1

U denotes strains that were not assigned to a phylo-group.

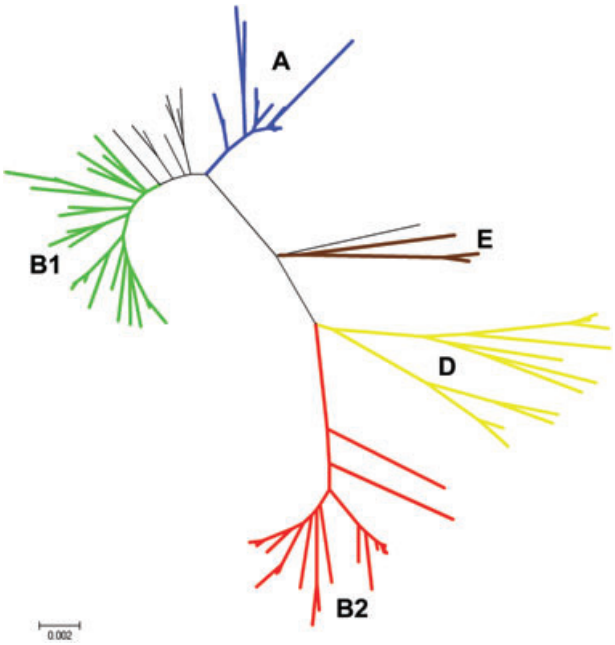


Fig. 1. Among-strain relationships as described by a NJ tree based on 91 strains characterized using the French and German MLST schemes (762 informative sites from 13 genes). B2 strains are in red, D strains are in yellow, B1 strains are in green, E strains are in brown and A strains are in blue. Strains denoted by thin black lines could not be assigned to a phylo-group.

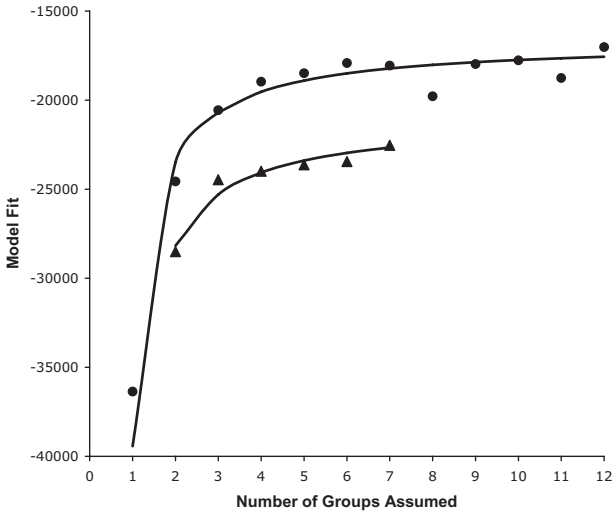


Fig. 2. Fit (ln likelihood) as a function of the number of populations assumed for the unsupervised population assignment algorithms STRUCTURE (circles) and BAPS (triangles) based on the strains characterized using the French MLST scheme (data for 6 genes, 848 informative sites).

although there was only a very slight improvement in the fit of the data when five rather than four genetic groups were assumed (Fig. 2). The five phylo-groups recognized by STRUCTURE were A, B1, B2, D and E. The BAPS analysis found that assuming seven genetic groups provided the best description of the data (Fig. 2). The groups recognized were the same five as recognized by STRUCTURE, but the BAPS analysis subdivided the group D strains (D-1 and D-2) and did the same for group B2 strains (B2-1 and B2-2) (Table 4).

Overall, there was a high degree of correspondence in the assignment of individual strains to genetic groups by the two clustering algorithms (Table 4). A greater number of isolates were unassigned by STRUCTURE compared with BAPS. Overall, 5% of the strains could not be unambiguously assigned to a phylo-group (Table 5). The phylogenetic relationships among the strains analysed using the French MLST scheme are depicted in Fig. 3.

Of the 225 strains examined, 190 (84%) were correctly assigned using the Clermont method (Table 5). However, the extent to which strains were miss-assigned depended on the strain's Clermont phylo-type. Strains with a B2 phylo-type (+++ or ++−) were assigned to the correct group 99% of the time and all strains with a B1 phylo-type (−−+) were considered to be members of the B1 phylo-group. Strains with a D phylo-type (+−+ or +−−) were

Table 3. Comparison of phylo-group assignment of 91 strains using the PCR-based Clermont method and their assignment using the nucleotide sequence data from the combined French and German MLST schemes.

Clermont assignment	MLST assignment	Number of strains
A (−−−)	U	8
A (−+−)	A	20
A (−+−)	B1	2
B1 (−−+)	B1	21
B2 (+++) (++−)	B2	23
D (+−+) (+−−)	D	13
D (+−+) (+−−)	E	3
D (+−+)	U	1

Only strains assigned to the same phylo-group by both STRUCTURE and BAPS were considered a member of that phylo-group. U denotes strains that were not assigned to a phylo-group.

Table 4. Results of the population assignment algorithms STRUCTURE and BAPS when applied to the strains characterized using the French MLST scheme (data for 6 genes, 848 informative sites).

STRUCTURE	BAPS	Number of isolates
A	A	33
B1	B1	63
B2	B2-1	73
B2	B2-2	10
B2	U	1
D	D-2	16
D	D-1	10
E	E	8
U	B1	6
U	U	5

U denotes strains that were not assigned to a phylo-group.

Table 5. Comparison of phylo-group assignment of strains using the PCR-based Clermont method and the assignment of 225 strains using the nucleotide sequence data from the French MLST scheme.

Clermont assignment	MLST assignment	Number of isolates
A (−−−)	B1	4
A (−−−)	U	8
A (−+−)	A	33
A (−+−)	B1	11
A (−+−)	U	2
B1 (−−+)	B1	48
B2 (+++) (++−)	B2	83
B2 (+++)	U	1
D (+−+) (+−−)	D	26
D (+−+) (+−−)	E	8
D (+−+)	U	1

Only strains assigned to the same phylo-group by both STRUCTURE and BAPS were considered a member of that phylo-group. U denotes strains that were not assigned to a phylo-group.

incorrectly described as belonging to group D 26% of the time. Strains with the Clermont phylo-type −+− were correctly classified as group A strains 72% of the time with the remainder being assigned to phylo-group B1. None of the strains that failed to yield any of the Clermont method PCR products (−−−) were correctly classified as group A strains and the majority of these strains clustered with one another in the NJ tree (Fig. 3).

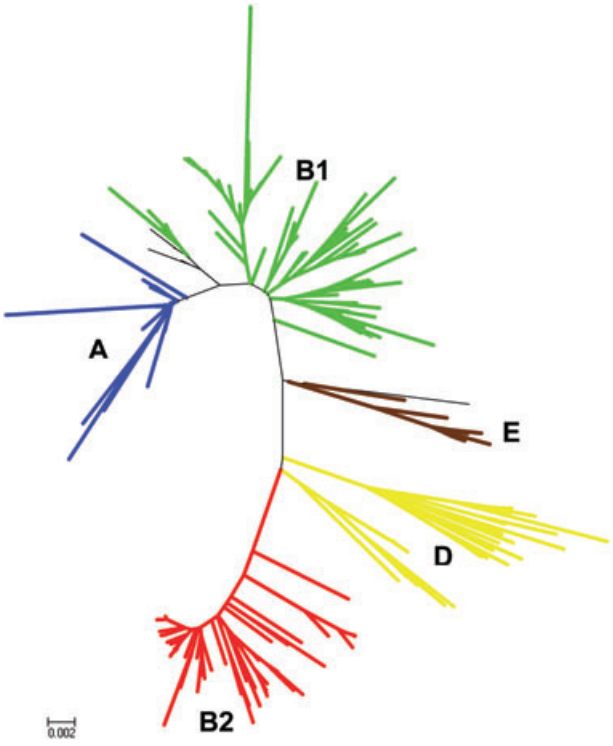


Fig. 3. Among-strain relationships as described by a NJ tree based on the strains characterized using the French MLST scheme (data for 6 genes, 848 informative sites). B2 strains are in red, D strains are in yellow, B1 strains are in green and A strains are in blue. Group E strains are in brown. Strains denoted by thin black lines could not be assigned to a phylo-group.

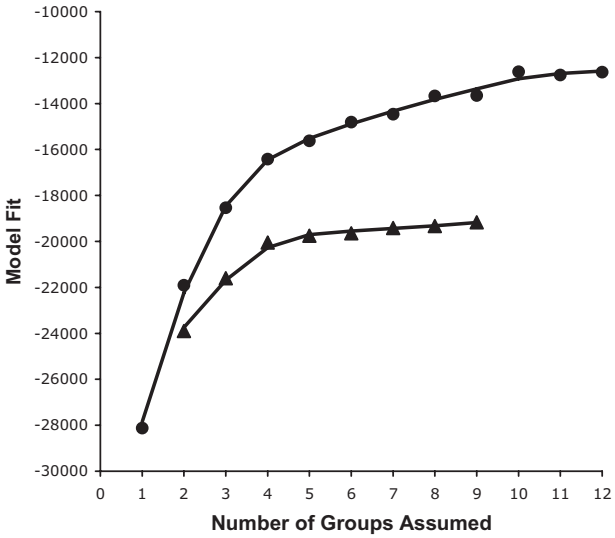


Fig. 4. Fit (ln likelihood) as a function of the number of populations assumed for the unsupervised population assignment algorithms STRUCTURE (circles) and BAPS (triangles) based on the strains characterized using the German MLST scheme (data for 7 genes, 319 informative sites).

German MLST scheme and Australian strain collection

For the 509 strains characterized using the German MLST scheme, the assignment algorithm BAPS found that assuming seven populations provided the best description of the data (Fig. 4). When using STRUCTURE, most of the variation in the data was captured when four populations were assumed, although there was a continued improvement in the fit to the data as the number of assumed populations increased from four to about 10 (Fig. 4). However, when more than seven populations were assumed, strain assignment, especially for the uncommon groups, became increasingly inconsistent. The phylogenetic relationships among the strains analysed using the German MLST scheme are depicted in Fig. 5.

The best concordance between the assignment algorithms was observed for seven populations (Table 6). Both algorithms identified the groups A, B1, B2 and D, as well as phylo-group E. STRUCTURE subdivided group E strains (E-1 and E-2) and also subdivided group B1 strains (B1-1 and B1-2). BAPS subdivided D strains (D-1 and D-2). BAPS also recognized a new clade, one not detected by STRUCTURE.

Of the 509 strains characterized, 400 (79%) were correctly assigned using the Clermont method (Table 7). However, as previously observed, the extent to which strains were miss-assigned depended on the strain's

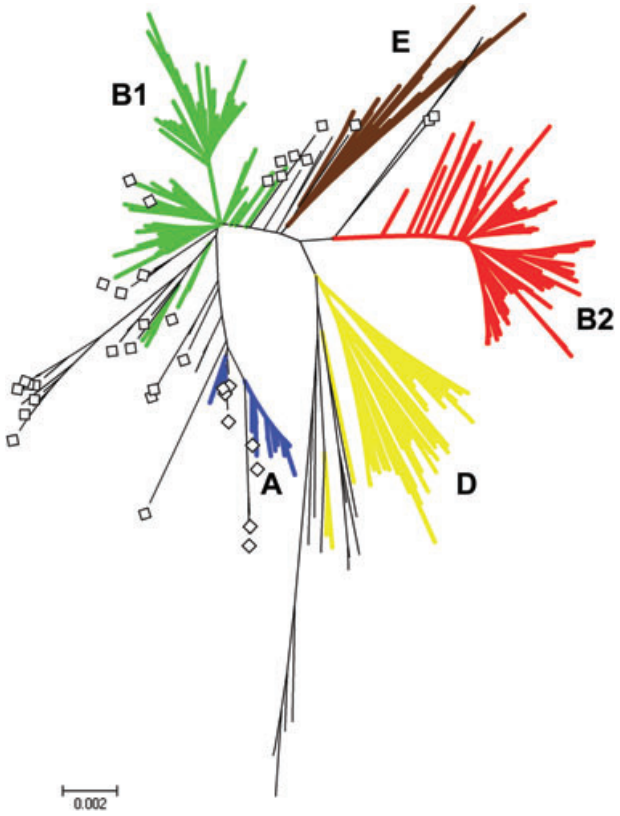


Fig. 5. Among-strain relationships as described by a NJ tree based on the strains characterized using the German MLST scheme (data for 7 genes, 319 informative sites). B2 strains are in red, D strains are in yellow, B1 strains are in green and A strains are in blue. Group E strains are in brown. Strains denoted by thin black lines could not be assigned to a phylo-group. Squares denote strains that failed to yield any PCR products using the method described by Clermont and colleagues (2000).

Table 6. Results of the population assignment algorithms STRUCTURE and BAPS when applied to the 509 strains characterized using the German MLST scheme (data for 7 genes, 319 informative sites).

STRUCTURE	BAPS	Number of strains
A	A	55
B1-1	B1	93
B1-1	U	2
B1-2	B1	78
B2	B2	129
D	U	3
D	D-1	9
D	D-2	62
E-1	E	22
E-2	E	10
U	A	11
U	B1	2
U	B2	2
U	NC	5
U	U	26

U denotes strains that were not assigned to a phylo-group and NC denotes a previously undescribed phylo-group.

Table 7. Comparison of phylo-group assignment of strains using the PCR-based Clermont method and the assignment of 509 strains using the nucleotide sequence data from the German MLST scheme.

Clermont assignment	MLST assignment	Number of strains
A (−−−)	A	9
A (−−−)	B1	8
A (−−−)	E	8
A (−−−)	U	20
A (−+−)	A	46
A (−+−)	B1	9
A (−+−)	D	1
A (−+−)	E	1
A (−+−)	U	10
B1 (−−+)	B1	154
B1 (−−+)	E	7
B2 (+++) (++−)	B2	122
B2 (+++) (++−)	E	5
B2 (+++) (++−)	U	7
D (+−+) (+−−)	B2	7
D (+−+) (+−−)	D	70
D (+−+) (+−−)	E	11
D (+−+) (+−−)	U	14

Only strains assigned to the same phylo-group by both STRUCTURE and BAPS were considered a member of that phylo-group. U denotes strains that were not assigned to a phylo-group.

Clermont genotype. Only 4% of the strains with a B1 phylo-type (−−+) were mis-classified and only 9% of strains with a B2 phylo-type (+++ or ++−) were not members of phylo-group B2. Of the 103 strains yielding a Clermont D phylo-type (+−+ or +−−), 33% were incorrectly assigned to phylo-group D. Strains with a Clermont phylo-type, −+−, were found to be group A strains only 69% of the time, while strains yielding no Clermont method PCR products, −−−, were group A strains only 18% of the time.

In silico *analysis of* chuA*,* yjaA *and TSPE4.C2*

Full genome data were available for the following 18 strains: phylo-group A – HS, K12-MG1655, K12-W3110; phylo-group B1 – 55989, IAI1, E22; phylo-group D – O42, IAI39, UMN026; phlyo-group E – EDL933, Sakai; and phylo-group B2 – 536, APECO1, CFT073, E2348/69, ED1a, S88, UTI89. The genes *chuA* (1983 bp), *yjaA* (384 bp) and the fragment TSPE4.C2 (contained in a putative lipase esterase gene of 909, 921 or 984 bp according to the strain) are in synteny among the different strains. The gene *chuA* is between *yhiF* and *yhiD* and, when it is absent from a strain, the entire 9 kb operon, consisting of *chuS*, *chuA*, *chuT*, *chuW*, *chuX*, *chuY*, *chuU* and *hmuV*, is absent. The gene *yjaA* is flanked by *rrfE* and *yjaB* and, when absent, only this gene is missing. The fragment TSPE4.C2 lies between *yiiD* and *yiiE* and, when absent, only the putative lipase esterase gene is missing. When any of these three genes is absent from a strain, it has not been replaced by another DNA fragment.

chuA, yjaA *and TSPE4.C2 variation in miss-assigned and unassigned strains*

To gain further insights regarding the strains that were not assigned to one of the phlyo-groups A, B1, B2 or D, or were miss-assigned by the Clermont method, nucleotide variation in a 754 bp fragment of *chuA*, a 347 bp fragment of *yjaA* and a 833 bp fragment of the putative lipase esterase gene containing TSPE4.C2, was determined for a subset of the Australian isolates.

Based on Clermont method, the presence of *chuA* denotes a strain belonging to phylo-group B2 or D. Among-strain variation in *chuA* largely reflected the patterns observed for the seven housekeeping loci used in the MLST analysis with the exception of TA326 (Fig. 6). Based on the MLST data, TA326 was unambiguously assigned to phylo-group B2, yet it has a Clermont method D phylo-type (+−−) and is more similar in its *chuA* nucleotide sequence to variants in phylogroup D strains than to *chuA* variants in phylo-group B2 strains. Notably, all of the strains assigned to phylo-group E exhibited similar *chuA* sequence variation, despite these strains having a variety of Clermont method phylo-types.

The gene *yjaA* distinguishes phylo-group B2 from phlyo-group D strains and is present in most of phylo-group A strains. For the most part, the among-strain relationships inferred from *yjaA* nucleotide variation reflect the patterns observed using the MLST data (Fig. 7). Strains assigned to phylo-group B1 using the MLST data, but which yielded a Clermont phylo-type of −+−, exhibited identical *yjaA* alleles and these were distinct from the *yjaA* variants observed in phylo-group A strains. Strains assigned to phylo-group B2 based on the MSLT data exhibited similar *yjaA* alleles. However, phylo-groups A and E strains exhibited either identical or very similar *yjaA* alleles. Strains (e.g. E807, H442) that were identified by the BAPS assignment algorithm as belonging to a novel phylo-group (Table 6) exhibited very different *yjaA* alleles.

The TSPE4.C2 fragment is present in all phylo-group B1 strains, most of the phylo-group B2 strains and few of the phylo-group D strains. Here again, the putative lipase esterase gene containing TSPE4.C2 reflects the MLST phylogeny (Fig. 8).

There were a number of strains that exhibited a Clermont method D phylo-type (+−+ or +−−), but which were assigned to phylo-group B2 based on the MLST data (TA454, B564, E1139, M540, B679 and TA326). Inspection of the virulence profile data for these strains revealed that, with the exception of TA326, all of these strains encoded the gene *ibeA* (invasion of brain epithelium), a trait that is most frequent in phylo-group B2 strains. Inspection of the virulence factor data for other strains isolated in Australia with a Clermont D phylo-type ($n = 272$) revealed 22 strains with a Clermont D phylo-

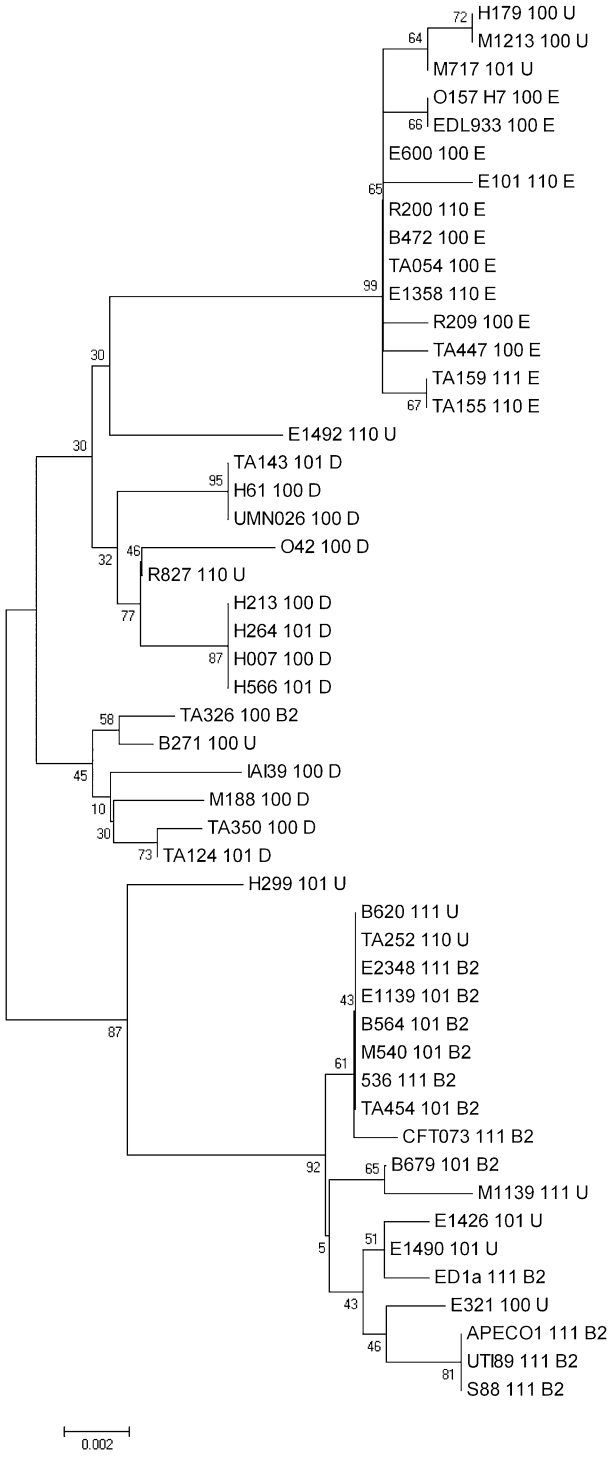


Fig. 6. NJ tree depicting the among-strain relationships based on a 754 bp portion of the *chuA* gene. The label presents the strain name, then the Clermont phylo-type of the strain (*chuA*, *yjaA*, TSPE4.C2; where 1 denotes presence and 0 absence of a PCR product) followed by the strain's phylo-group assignment based on the available MSLT data. Numbers at the nodes represent bootstrap support values.

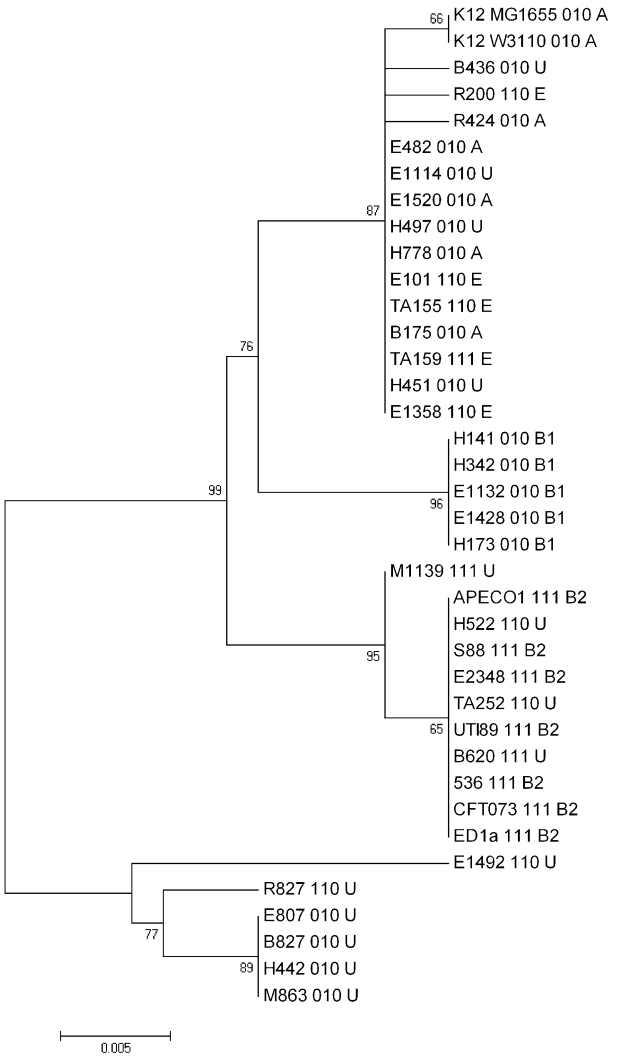


Fig. 7. NJ tree depicting the among-strain relationships based on a 347 bp portion of the *yjaA* gene. The label presents the strain name then the Clermont phylo-type of the strain (*chuA*, *yjaA*, TSPE4.C2; where 1 denotes presence and 0 absence of a PCR product) followed by the strain's phylo-group assignment based on the available MSLT data. Numbers at the nodes represent bootstrap support values.

type and which were *ibeA*-positive. The nucleotide sequences of the *chuA* and, where possible, the putative lipase esterase gene containing TSPE4.C2 fragments, were determined for these strains. In all but a single case, the *ibeA*-positive strains exhibiting a Clermont method D phylo-type clustered, with high bootstrap support, together with strains belonging to phylo-group B2 (Fig. 9). The one exception, B271, exhibits a *chuA* allele that was more similar to phylo-group D strains than to phylo-group B2 strains. However, this strain had a *chuA* allele that was very similar to the allele exhibited by TA326, a strain which the MLST data unambiguously assigned to phylo-group B2.

Discussion

Number of phylo-groups in E. coli

That *E. coli* consists of a number of distinct phylogenetic groups has been demonstrated using a variety of genetic characterisation methods and statistical techniques (Herzer *et al*., 1990; Desjardins *et al*., 1995; Wirth *et al*., 2006). How many phylogenetic groups can be detected depends on the resolving power of both the genotyping method, the analytical approach used to identify strain clusters, as well as the nature of the sample being characterized.

For the ECOR collection of strains and using multi-locus enzyme electrophoresis, 10 appropriately chosen enzyme

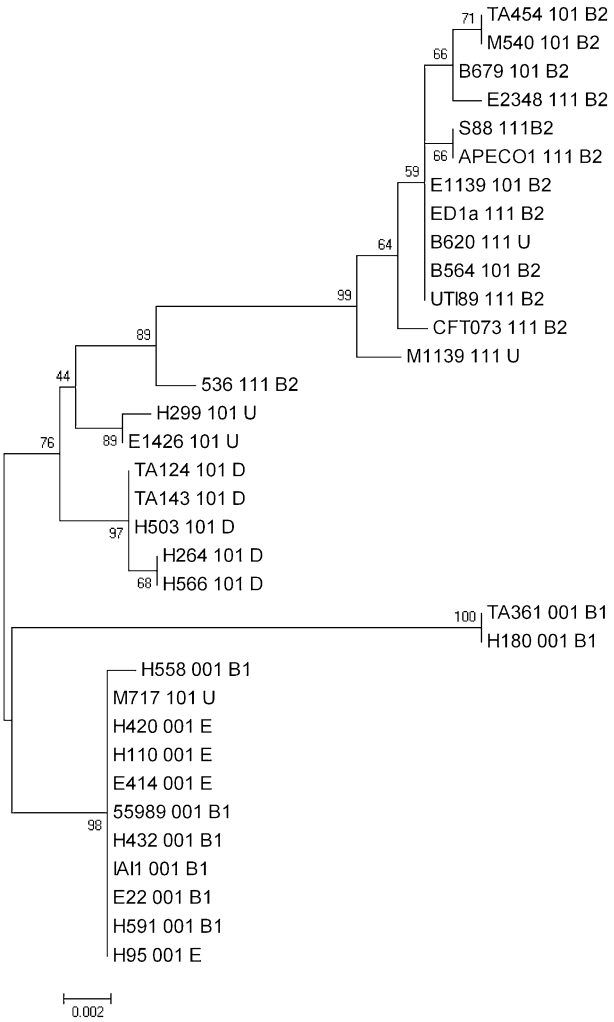


Fig. 8. NJ tree depicting the among-strain relationships based on a 833 bp portion of the TSPE4.C2 DNA fragment. The label presents the strain name then the Clermont phylo-type of the strain (*chuA*, *yjaA*, TSPE4.C2; where 1 denotes presence and 0 absence of a PCR product) followed by the strain's phylo-group assignment based on the available MSLT data. Numbers at the nodes represent bootstrap support values.

loci (Pupo *et al*., 1997) can recover most of the structure found using 38 loci (Herzer *et al*., 1990). In the present study, there is little difference among the three MLST data sets in the fraction of strains unambiguously assigned to phylogenetic groups, despite the fact that there is more than a twofold difference in the number of genes characterized (6, 7 or 13) and in the number of informative sites (319, 762 or 848) among the data sets.

The assignment algorithms STRUCTURE and BAPS produced very similar results. STRUCTURE tended to be more conservative, in that a greater number of strains were not assigned to any population. The results of both assignment algorithms also showed good agreement with the clustering patterns seen using NJ, unweighted pair group method with averages (UPGMA) or minimum evolution tree building algorithms. The variants of STRUCTURE and BAPS that consider linkage tended to be less conservative and assigned a slightly greater fraction of strains to a phylogenetic group. Although the results of these analyses have not been presented, ordination methods also yielded strain groupings similar to those observed using tree building or population assignment approaches.

The analyses of the three data sets indicate that there are five primary groups within *E. coli*: the four well-established groups of A, B1, B2 and D, as well as E. Phylo-group E has been previously recognized (Selander *et al*., 1987; Escobar-Páramo *et al*., 2004a), but strains of this phylo-group appear to be uncommon. Strains assigned by both STRUCTURE and BAPS to phylo-group E exhibited most of the possible Clermont phylo-types: −−−, −+−, −−+, +++, +−+ and +−−. Phylo-group E encompasses the highly virulent enterohemorragic O157 : H7 strains which exhibit the +−− Clermont phylotype (Clermont *et al*., 2000). However, the extent to which group E represents a phlyo-group comparable to, for example, phylo-group B2, as opposed to a clonal complex, is unknown.

Strains traditionally recognized as belonging to phylo-group D were often split into two groups (Figs 1, 3 and 5). The extent to which the 'D' group with the smaller number of strains, of which the ECOR strains 35, 36, 38, 39, 40, 41 and 43 are members, belong with the other D strains is unknown. There is little support for the monophyly of phylo-group D strains and it may well be that the strains in this small group should not be considered as phlyo-group D strains. However, all strains found to be members of this small group of D strains did exhibit the Clermont D phylotype (+−−).

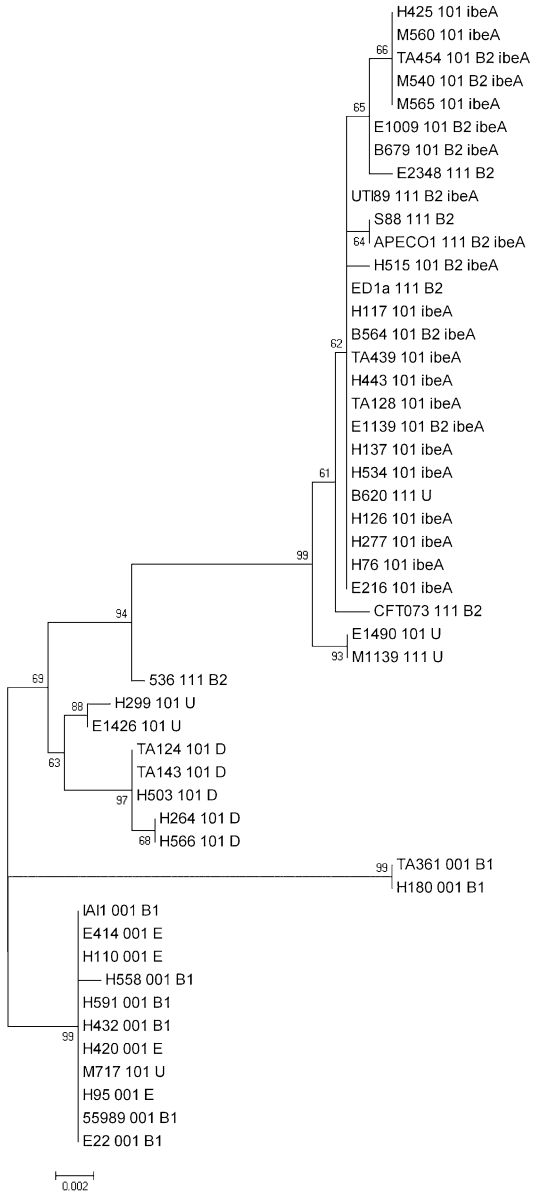
The results of the STRUCTURE analysis of the strains characterized using the German MLST scheme, as well as inspection of the NJ tree (Fig. 5), suggest two distinct groups within phylo-group B1. Recently, a collection of *E. coli* strains collected from fresh-water beaches was characterized using a third MLST scheme (http://www.shigatox.net/cgi-bin/mlst7/index) (Walk *et al*., 2007). In this study, two distinct groups of B1 strains were also observed (Walk *et al*., 2007). It is not known if these two groups of B1 strains have different ecological distributions, virulence factor or biotype profiles.

Groups A and B1 have long been recognized as sister groups and, in the NJ trees depicted, strains assigned to phylo-group A strains tend to appear as a relatively homogenous group among all the strains assigned to phylo-groups A and B1. Indeed, the majority, and by some criteria, 95% of strains assigned to phylo-group A belong to a single very successful clonal complex known as the ST10 complex (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli). The founder of this complex, ST 10, is the single most common ST in the German MLST database as well as in the collection of Australian strains characterized using MLST (D. M. Gordon, unpubl. data). All members of the ST 10 complex characterized using the Clermont method have yielded a −+− genotype. These results suggest that perhaps phylo-group A strains are better thought of as a clonal complex, rather than a phylogenetic group which should consist of many clonal complexes as do the groups B1, B2 and D.

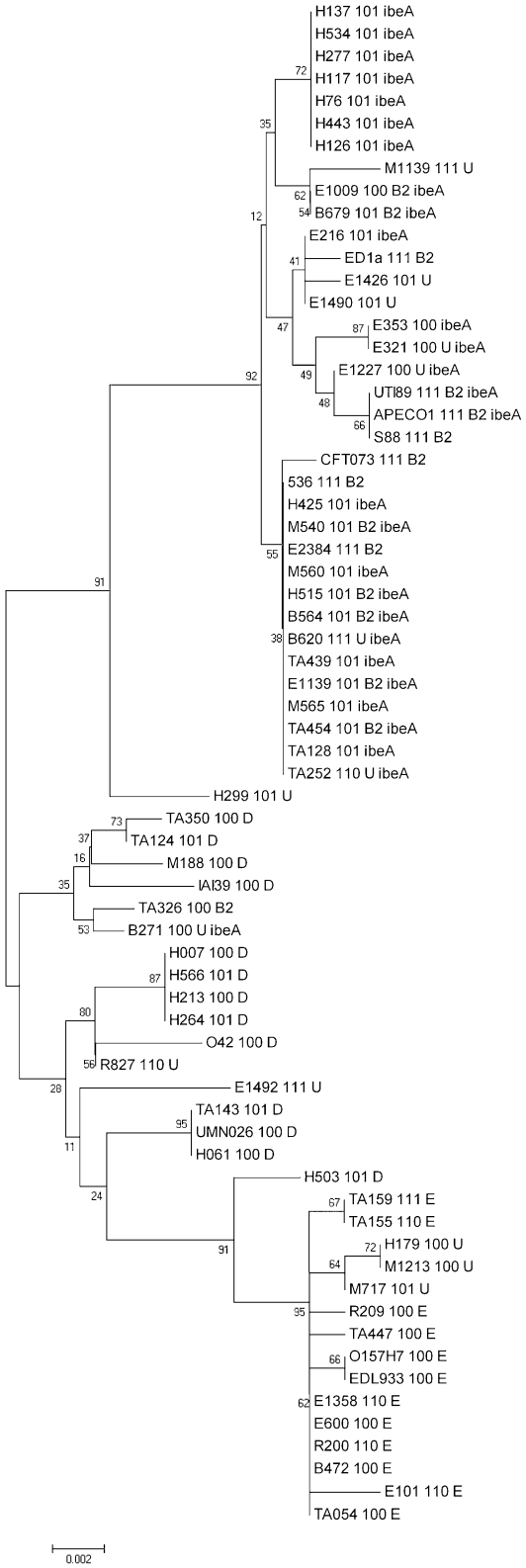
Utility of the Clermont phylo-typing method

The Clermont method was designed to assign *E. coli* strains to the phylo-groups A, B1, B2 and D. The method will clearly fail to correctly assign any strain that does not belong to one of these four phylo-groups. Therefore, it is important to know what fraction of *E. coli* strains cannot be assigned to one of the four main groups of *E. coli* (A, B1, B2 and D). Among the strains not belonging to the ECOR collection and characterized using the French MLST scheme, 90% were assigned to one of these four groups. While among the non-ECOR strains characterized using the German MLST scheme, 84% of the isolates could be assigned to one of these four phylo-groups. The difference between the two results is most likely due to the way the Australian strains were collected and selected. First, the Australian strains were isolated from a wide variety of vertebrate hosts, as well as from soil, sediment and water samples. The isolates characterized using the French MLST scheme were primarily isolated from humans and other mammals. Second, the Australian isolates were selected on the basis of their Clermont genotype and virulence factor profile in an attempt to maximize the diversity of the sample. No such criteria were used in choosing the strains characterized using the French MLST scheme.

Finally, it is important to note that the procedure used to assign strains to a phylo-group was conservative. A strain had to be assigned to the same phylo-group by both



TSPE4.C2



chuA

Fig. 9. Phylo-group membership of strains yielding a Clermont method D phylo-type and which encoded the invasion of brain epithelium gene, *ibeA*. NJ tree depicting the among-strain relationships based on a 833 bp portion of the putative lipase esterase gene encompassing the TSPE4.C2 DNA fragment and a 754 bp fragment of *chuA*. The label presents the strain name then the Clermont phylo-type of the strain (*chuA*, *yjaA*, TSPE4.C2; where 1 denotes presence and 0 absence of a PCR product) followed by the strain's phylo-group assignment based on the MSLT data if available. The *ibeA*-positive strains are labelled as such. Numbers at the nodes represent bootstrap support values.

population assignment algorithms. Further, there was no attempt to 'second-guess' the assignment outcomes by determining the placement of a strain in a tree or in ordination space. However, it is likely that additional data would resolve the phylo-group membership of some of the unassigned strains. For example, the strains TA252, B620 and M1139 exhibited a Clermont method B2 phylo-type, but were not assigned to phylo-group B2 based on the MLST data. Yet, the nucleotide sequence data for *chuA*, *yjaA* and TSPE4.C2 clearly place these strains in phylo-group B2 (Figs 6–8). Therefore, the data indicate that the great majority (> 85%) of *E. coli* isolates can be assigned to the phylo-groups A, B1, B2 or D.

Overall, 80–85% of the phylo-group memberships assigned using the Clermont method are correct. Regardless, the frequency with which the Clermont method yields the correct phylo-group membership depends on the phlyo-type exhibited by the strain. Of the 186 strains exhibiting a B1 Clermont genotype (−−+), 96% were in fact members of phylo-group B1 based on the MLST data. Of the 195 strains yielding a B2 Clermont genotype (+++ or ++−), 94% were members of group B2. Of those strains with genotypes appropriate to group D, 69% are members of phylo-group D. There are two Clermont D genotypes and among the isolates characterized using the German MLST scheme, 71% had a +−− genotype and the balance the +−+ genotype. Of the strains with a +−− genotype, 29% were not assigned to phylo-group D, while 44% of strains with the +−+ genotype were not phylo-group D strains; however, these proportions are not significantly different (contingency table analysis: likelihood ratio $\chi^2 = 1.754$, $P = 0.185$).

All strains that were assigned to phylogroup A exhibit an appropriate Clermont genotype. However, many strains that exhibit a phylo-group A Clermont genotype are not members of group A; 64% of strains exhibiting a −+− genotype and only 17% of strains with a −−− genotype belong to group A. In addition, although strains with a Clermont genotype of −−− are not randomly distributed across *E. coli*, they are not a monophyletic group of strains (Fig. 5).

Thus, the number of strains miss-assigned to a phylo-group using the Clermont method will depend on the composition of the collection of strains being assessed. The phylo-group composition of a sample has been shown to vary with the taxonomic class of the vertebrate host and with host diet in Australian mammals (Gordon and Cowling, 2003), as well as with geographic locality for isolates from humans (Escobar-Páramo *et al*., 2004b). Consequently, for collections consisting of strains largely yielding B1 or B2 phylo-types, there will be little assignment error. While it is likely that more miss-assignments would be observed if the collection consisted of strains yielding those phylo-types associated with phylo-group A.

However, although strains with a −−− genotype are seldom members of phylo-group A, these strains are relatively rare. In Australia, 9% of the isolates from non-human vertebrates, as well as soil, water and sediment samples ($n = 888$) and 5% of isolates ($n = 619$) from humans living in Australia had this genotype. In the collection of strains characterized using the French MLST scheme, 8% of isolates from animals ($n = 75$) and 4% of isolates (*n*-150) from humans failed to yield any Clermont method PCR products. Walk and colleagues (2007) reported that 7% of 191 *E. coli* isolated from beach samples taken from the Great Lakes region of North America yielded a −−− genotype. While, in a set of 460 *E. coli* collected from water samples, animal body tissues and the faeces of humans, domesticated and zoo animals, 8% of the strains exhibited the −−− genotype (Higgins *et al*., 2007).

The foundation of the Clermont method

The genes *chuA*, *yjaA* and the fragment TSPE4.C2 appear to have diverse origins. The GC content of a typical housekeeping gene in *E. coli* is 50.7%, while the GC content of *chuA* is 50.6%, 52.4% for the putative lipase gene, and is 45.3% for *yjaA*. Thus, *yjaA* is likely to have been acquired from a distantly related species. Although a fragment similar to TSPE4.C2 is present in *Salmonella*, *chuA* and *yjaA* are absent from *E. coli*'s sister genus. The three genes are also absent from *Esherichia fergusonii* (data not shown), the closest relative of *E. coli* (Lawrence *et al*., 1991), perhaps indicating their acquisition after the *E. fergusonii–E. coli* divergence. However, as the phylogenies of these genes are congruent with the strain phylogeny inferred from the MLST data, the most parsimonious scenario is that these three genes were acquired early in the *E. coli*'s history and subsequently lost from some lineages. Of the three DNA fragments used in the Clermont method, *chuA* appears to be the most stable, as *chuA* is present in all members of the B2 phylo-group and all but a single member of D phylo-group,

and this concurs with the fact that the gene is part of a large operon. The TSPE4.C2 fragment is the next most stable as it is present in the great majority of B1 and B2 strains, and is absent from the majority of D strains. The gene *yjA* is the least informative of the markers used by the Clermont method, as it can be absent from some B2 strains and a significant fraction of phylo-group A strains and can be the only Clermont marker present in some phylo-group B1 strains.

Concluding remarks

The results of the analyses presented in this paper suggest that improvements to the Clermont method may be difficult to achieve as the gene content of *E. coli* is highly variable. The results of our analysis indicate that all strains exhibiting a phylo-type of +−+ should be screened for the gene *ibeA*, as *ibeA*'s presence indicates that it is highly likely that the strain is a member of phylo-group B2. Among the isolates from Australia, 23% of strains exhibiting a +−+ phylo-type were *ibeA*-positive.

It may be difficult to identify a DNA fragment that is both unique to, as well as common to, all phylo-group E strains, as strains of this group seem to share alleles with many *E. coli* strains. For example, although phylo-group E strains exhibit distinct *chuA* alleles, they share TSPE4.C2 alleles with group B1 strains and *yjaA* alleles with phylo-group A strains. Genome data on additional phylo-group E strains may allow the identification of markers unique to these strains.

The results of this study demonstrate that for population level studies, the Clermont method is an excellent technique for rapidly and inexpensively assigning strains of *E. coli* to phylogenetic groups. For samples taken from the faeces of asymptomatic human hosts, the fraction of strains present in the sample that cannot be assigned to a phylo-group is likely to be low and the fraction of strains that are miss-assigned to a phylo-group is also likely to be low. Strains failing to yield any PCR products using the Clermont method should be scored as unassigned and, as strains with this 'profile' are not monophyletic, they should be excluded from any statistical analyses.

Further studies are needed to determine the validity of group A strains and E strains as phylogenetic groups as compared with phylo-groups B1, B2 and D. Investigations to determine the 'boundaries' of phylo-group D are also required.

Experimental procedures

Strain examined

The ECOR collection of strains (Ochman and Selander, 1984) and two other sets of strains were used in this study. The first of these was a set of 437 strains selected from collections totalling more than 1400 isolates recovered from humans living in Australia (Gordon *et al*., 2005), non-human vertebrates living in Australia (Gordon and Cowling, 2003), and from Australian soil, water and sediment samples (Power *et al*., 2005). All of the Australian strains had been PCR-screened for the presence of 27 virulence genes associated with intestinal and extra-intestinal disease as described in Gordon and colleagues (2005), and their Clermont method genotype had been determined (Clermont *et al*., 2000). The strains selected for further characterization using MLST were chosen on the basis of their phylo-group membership as determined using the Clermont method and their virulence factor profile. The aim was to select, when possible, an approximately equal number of strains from each of phylo-groups A, B1, B2 and D, so that each strain selected had a unique virulence factor profile. When this was not possible, strains were selected on the basis of the host species or geographic locality from which they were isolated. This procedure was repeated for strains from each major collection: humans, animals and the environment. The second set was composed of 153 strains from human and animals, encompassing both commensal and intra- and extra-intestinal pathogenic strains (Escobar-Páramo *et al*., 2004a; Le Gall *et al*., 2007). All strains had been phylo-typed using the Clermont and colleagues (2000) technique. They were chosen for their diversity in term of their phylo-group membership, their host (humans, mammals and birds) and geographic origins (Europe, Africa, America) and the clinical syndrome they produced (urinary tract infection, septicaemia, new-born meningitis, diarrhoea).

Multi-locus sequence typing

The strains isolated from Australia were characterized using the MLST scheme (German) described by Wirth and colleagues (2006). This scheme examines seven loci: *adk* (536 bp), *fumC* (469 bp), *gyrB* (460 bp), *icd* (518 bp), *mdh* (452 bp), *purA* (478 bp) and *recA* (510 bp). The strains of the ECOR collection had already been characterized using this scheme and the MLST profile of these strains is available from (http://web.mpiib-berlin.mpg.de/mlst/dbs/Ecoli).

The balance of the strains examined in this study, as well as that of the strains of the ECOR collection, were characterized using a modified version of the MLST scheme (French) described at the Institut Pasteur's MLST site (http://www.pasteur.fr/mlst). Nucleotide sequence data were obtained for the genes: *icd* (1166 bp), *pabB* (1003 bp), *polB* (1081 bp), *putB* (906 bp), *trpA* (727 bp) and *trpB* (1140 bp).

The nucleotide sequence data for each locus available for a strain were concatenated and the concatenated data sets were used in all analyses.

Phylo-group assignment

Two assignment algorithms were used to assign strains to genetic groups: BAPS (Corander and Marttinen, 2006; Corander and Tang, 2007) and STRUCTURE (Pritchard *et al*., 2000; Falush *et al*., 2003). Both use Bayesian modelling approaches, but STRUCTURE attempts to determine the

number of underlying populations and identify admixture events simultaneously while, in BAPS, the number of ancestral populations is inferred first and then the admixture events are identified. For both STRUCTURE (http://pritch.bsd.uchicago.edu/structure.html) and BAPS (http://web.abo.fi/fak/mnf//mate/jc/software/baps.html), versions are available where the molecular markers are assumed to be either independent or linked.

Three data sets were considered: a set of strains for which MLST data from both MLST schemes were available, the strains characterized using the French MLST scheme and those characterized using the German MLST scheme. When running the simulations, from 2 to 15 populations were assumed to be present and multiple simulations were carried out. In STRUCTURE, the fit of the model can be determined under the assumption that all strains belong to a single population and this was done for each of the data sets. The three data sets were analysed using the linked marker and unlinked marker versions of STRUCTURE and BAPS. Very little difference in the assignment of strains to phylo-groups was observed when the results of the analyses assuming linked or unlinked markers were compared. As what few differences were observed did not affect the conclusions reached, only the results from the analyses assuming unlinked loci are presented. In assigning a strain to a phylo-group, a *Q*-value (probability of membership) of > 0.66 was used. This value was chosen by examining the distribution of *Q*-values in all analyses, as the great majority of *Q*-values were either greater than 0.66 or very much smaller.

The phylo-group membership of strains in the ECOR collection were used to name the populations found using STRUCTURE and BAPS. For each of the data sets, a strain was assigned to a phlyo-group only when both STRUCTURE and BAPS assigned the strain to the same phylo-group.

For each of the concatenated data sets, the number of pair-wise differences among strains was determined and used in Principal Co-ordinates analyses and for constructing NJ, UPGMA and minimum evolution trees.

Phylogenies of chuA*,* yjaA *and TSPE4.C2*

In silico comparative analyses of complete *E. coli* genomes were performed using MaGe interface (Vallenet *et al*., 2006) from the publicly available genomes and from the ColiScope consortium genomes (http://www.genoscope.cns.fr/spip/Escherichia-fergusonii.html). The following primers were used to amplification and nucleotide sequence determination of *chuA* (F-ATCTGGATGGTATTGTGGCCTGGT, R-AGTTTCCGGACGTAAGTTCGGGTT), *yjaA* (F-ATGTCAGTTCTGTATATCCAAATTCGTCG, R-ATTAGTATTCGCCGCTCACG) and TSPE4.C2 (F-GGCACTGGAAAAAGGAATTGC, R-TTACCTTCCCGCTCTCCAGG).

Acknowledgements

We are grateful to Maryvonne Moulin-Schouleur and Eric Oswald for the pathogenic animal strains. This work was partially supported by the 'Fondation pour la Recherche Médicale' and in part by the Alliance for the Prudent Use of Antibiotics through NIH Grant No. U24 AI 50139.

References

Bergthorsson, U., and Ochman, H. (1998) Distribution of Chromosome length variation in natural isolates of *Escherichia coli*. *Mol Biol Evol* **15:** 6–16.

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

Corander, J., and Marttinen, P. (2006) Bayesian identification of admixture events using multi-locus molecular markers. *Mol Ecol* **15:** 2833–2843.

Corander, J., and Tang, J. (2007) Bayesian analysis of population structure based on linked molecular information. *Math Biosci* **205:** 19–31.

Desjardins, P., Picard, B., Kaltenböck, B., Elion, J., and Denamur, E. (1995) 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.

Escobar-Páramo, P., Clermont, O., Blanc-Potard, A.B., Bui, H., Le Bouguénec, C., and Denamur, E. (2004a) A specific genetic background is required for acquisition and expression of virulence factors in *Escherichia coli*. *Mol Biol Evol* **21:** 1085–1094.

Escobar-Páramo, P., Grenet, K., Le Menac'h, A., Rode, L., Salgado, E., Amorin, C., *et al.* (2004b) Large-scale population structure of human commensal *Escherichia coli* isolates. *Appl Environ Microbiol* **70:** 5698–5700.

Falush, D., Stephens, M., and Pritchard, J.K. (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164:** 1567–1587.

Gordon, D.M. (2004) The influence of ecological factors on the distribution and genetic structure of *Escherichia coli*. In *Escherichia Coli and Salmonella Typhimurium: Cellular and Molecular Biology*. Neidhardt, F., *et al.* (eds). Washington, DC, USA: American Society for Microbiology. [Online] http://www.ecosal.org/ecosal/index.jsp

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

Herzer, P.J., Inouye, S., Inouye, M., and Whittam, T.S. (1990) Phylogenetic distribution of branched RNA-linked mulitcopy single-stranded DNA among natural isolates of *Escherichia coli*. *J Bacteriol* **172:** 6175–6181.

Higgins, J., Hohn, C., Hornor, S., Frana, M., Denver, M., and Joerger, R. (2007) Genotyping of *Escherichia coli* from environmental and animal samples. *J Microbiol Methods* **70:** 227–235.

Johnson, J.R., Delavari, P., Kuskowski, M., and Stell, A.L. (2001) Phylogenetic distribution of extraintestinal virulence-associated traits in *Escherichia coli*. *J Infect Dis* **183:** 78–88.

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.

Lecointre, G., Rachdi, L., Darlu, P., and Denamur. E. (1998) *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Mol Biol Evol* **15:** 1685–1695.

Nowrouzian, F.L., Adlerberth, I., and Wold, A.E. (2006) 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.

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

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

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

Pupo, G.M., Karaolis, D.K.R., Lan, R., and Reeves, P.R. (1997) Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* inferred from multilocus enzyme electrophoresis and *mdh* sequence studies. *Infect Immun* **65:** 2685–2692.

Selander, R.K., Caugant, D.A., and Whittam, T.S. (1987) Genetic structure and variation in natural populations of *Escherichia coli*. In *Escherichia Coli and Salmonella Typhimurium, Cellular and Molecular Biology*. Neidhardt, F.C., Ingraham, J.L., Magasanik, B., Low, K.B., Schaechter, M., and Umbarger, H.E. (eds). Washington, DC, USA: American Society of Microbiology, pp. 1625–1648.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., *et al.* (2006) MaGe: a microbial genome annotation system supported by synteny results. *Nucleic Acids Res* **34:** 53–65.

Walk, S.T., Alm, E.W., Calhoun, L.M., Mladonicky, J.M., and Whittam, T.S. (2007) Genetic diversity and population structure of *Escherichia coli* isolated from freshwater beaches. *Environ Microbiol* **9:** 2274–2288.

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.

