The population genetics of commensal *Escherichia coli*

Olivier Tenaillon, David Skurnik‡, Bertrand Picard*§ and Erick Denamur**

Abstract | The primary habitat of *Escherichia coli* is the vertebrate gut, where it is the predominant aerobic organism, living in symbiosis with its host. Despite the occurrence of recombination events, the population structure is predominantly clonal, allowing the delineation of major phylogenetic groups. The genetic structure of commensal *E. coli* is shaped by multiple host and environmental factors, and the determinants involved in the virulence of the bacteria may in fact reflect adaptation to commensal habitats. A better characterization of the commensal niche is necessary to understand how a useful commensal can become a harmful pathogen. In this Review we describe the population structure of commensal *E. coli*, the factors involved in the spread of different strains, how the bacteria can adapt to different niches and how a commensal lifestyle can evolve into a pathogenic one.

**INSERM U722 and Université Paris 7 Denis Diderot, Paris 75018, France.*
‡Harvard Medical School, Department of Medicine, Channing Laboratory, 181 Longwood Ave, Boston, Massachusetts 02115, USA.
§Hôpital Avicenne and Université Paris 13 Nord, Bobigny 93009, France.
Correspondence to E.D.
e-mail: erick.denamur@inserm.fr
doi:10.1038/nrmicro2298

Escherichia coli is one of the best characterized model organisms. The reference strain *Escherichia coli* K-12 and its derivatives have been key in the advancement of genetics, molecular biology, physiology and biochemistry. However, *E. coli* is not a single clone growing in laboratories[1]. In the wild its total population size has been estimated to be 10^{20} (REF. 2), and it has the interesting characteristic of being both a widespread gut commensal of vertebrates and a versatile pathogen, thought to kill more than 2 million humans per year through both intraintestinal and extraintestinal diseases[3,4]. As such, it is the perfect candidate to study the transition between commensalism and pathogenicity[5] or, more broadly, how the close link between a bacterium and its host can fluctuate between mutualism, commensalism and opportunistic pathogenesis or even specialized pathogenesis. Although pathogenic strains have been extensively investigated, few studies have focused on commensal strains, resulting in a bias towards pathogenic strains in the data sets. However, it is necessary to decipher the ecological and evolutionary forces that shape the population structure of the commensal strains to wholly understand the virulence and antibiotic resistance of pathogenic strains. Indeed, the selective pressures in the habitats of commensal strains may coincidentally promote the emergence of virulence factors and antibiotic resistance, rendering commensal *E. coli* strains reservoirs of virulent and resistant strains.

Commensal niches of *Escherichia coli*

E. coli, a Gram-negative, non-sporulating facultative anaerobe, is an inhabitant of the intestines and faeces of warm-blooded animals and reptiles[6,7]. *E. coli* is found in the gut microbiota, which consists of more than 500 species of bacteria that total 10^{10}–10^{11} cells per gram of large-intestinal content. Although the anaerobic bacteria in the bowel outnumber *E. coli* by 100/1 to 10,000/1 (REF. 6), *E. coli* is the predominant aerobic organism in the gastrointestinal tract. As *E. coli* can transit in water and sediment, it is often used as an indicator of faecal pollution of water; using intuitive calculations, it has been estimated that half of the *E. coli* population resides in these secondary habitats[8]. Some recent studies have revisited the role of these environments and have shown that they can support the growth of some specific strains (those that are capable of saprophytism), depending on nutrient availability and temperature[9,10].

The various *E. coli* hosts have distinct body sizes, gut morphologies, diets, digesta retention times and microbiota. These characteristics can have a substantial influence on the prevalence and density of *E. coli*, which vary from 0% to 100% and over 6 orders of magnitude, respectively, among host species. In humans, the prevalence is more than 90%[11,12], but it is only 56% in wild mammals, 23% in birds and 10% in reptiles[7]. The concentration per gram of faeces varies in humans from 10^7 to 10^9 colony-forming units (cfu)[11–13] and is much lower in domestic animals, averaging between 10^4 and 10^6 cfu[13] (TABLE 1). In the digestive tract, commensal *E. coli* strains are located in the large intestine, especially in the caecum and the colon. They reside in the mucus layer that covers the epithelial cells throughout the tract and are shed into the intestinal lumen with the degraded mucus

Table 1 | **Characteristics of the *Escherichia coli* in the microbiota**

Characteristic	Humans	Domestic animals	Wild animals
Prevalence (%)	>90	No data	10–55
Quantity (cfu per g of faeces)	10^7–10^9	10^4–10^6	No data
Main phylogroup	A (and B2)	B1	B1
Extraintestinal 'virulence genes'	Common	Rare	Rare
Antibiotic resistance score*	18	11	0–7
Integron prevalence (%)	16	7	0

cfu, colony-forming units. *The resistance score is defined as $(R/nA) \times 100$, where n is the number of strains tested, A is the number of antibiotics tested and R is the summation of the number of strains resistant to each antibiotic[101].

component and excreted in the faeces[14]. The mucus defines a nutritional ecological niche to which *E. coli* metabolism has adapted[15]. Strains that are isolated from that part of the intestine grow on nutrients acquired from mucus, including at least seven mucus-derived sugars, of which gluconate seems to have a predominant role[16]. Although the concentrations of these sugars in the intestine are low[17], *E. coli* maximizes its growth by using micro-aerobic and anaerobic respiration in the intestine[18]. This results in a 30 minute generation time *in vitro* on intestinal mucus[19] compared with 40–80 minutes in the intestines of streptomycin-treated mice, in which the cells in the luminal content are static[14], and 120 minutes when the mice are 'conventionalized' by removing the streptomycin and feeding them with mouse caecal content[20]. This change in growth rate in the presence of other species illustrates that *E. coli* competes with those other species. However, these interactions are complex and sometimes mutualistic, as *E. coli* may benefit from anaerobe-mediated degradation of mucosal polysaccharides and dietary fibres and may help these anaerobes by limiting the oxygen content of the intestine[18].

E. coli is among the first bacterial species to colonize the intestine during infancy, reaching very high density (higher than 10^9 cfu per gram of faeces)[11,12] before the expansion of anaerobes[21]. After 2 years, the density stabilizes and remains at around 10^8 cfu per gram of faeces until it gradually decreases in the elderly[11]. The initial strains may originate from the maternal faecal microbiota and also from the maternity nursing staff[22]. In fact, increased hygiene in hospitals and in families living in industrial countries has reduced early colonization by *E. coli*[23,24].

The relationship between *E. coli* and the host should be defined as commensalism, in which one of the two organisms benefits from the interaction between them, whereas the other is neither notably harmed nor helped. *E. coli* strains derive from their host a steady supply of nutrients[25], a stable environment and protection against some stresses, as well as transport and dissemination. However, the normal *E. coli* microbiota provides some benefits to its host by preventing colonization by pathogens (that is, by inducing colonization resistance in the host), which it does through the production of bacteriocins and through other mechanisms[26–29].

Panmictic
Pertaining to a population in which all individuals are potential recombination partners.

Overdominance
Occurs when natural selection favours the heterozygote over the homozygote in a diploid organism. Selectionists proposed overdominance as the driving force underlying the high level of polymorphism that is observed in natural populations.

A classical model in population genetics

To characterize how *E. coli* adapts to different commensal niches, it is necessary to unravel how the species is genetically structured on a global scale. A population structure is largely defined by the balance between recombination and mutation, shifting from a clonal structure when recombination is low to a panmictic structure when recombination is high[30]. As *E. coli* is both a pathogen and a commensal and is easy to isolate and grow in the laboratory, it has been used as a model in population genetics studies for decades, and researchers have benefited from each conceptual or technical advance. Although this has led to somewhat conflicting visions of the importance of recombination, genomic data have helped to reconcile these differences.

The pre-sequencing era: the basis of polymorphism. The clonal structure of the population was first supported by serotyping analysis[31]. As the relative frequency of many of the antigens varied with the isolate source but the O (somatic), K (capsular) and H (flagellar) antigens were non-randomly associated, and as some serotypes were distributed worldwide, it was postulated that the species *E. coli* consists of an array of stable lineages (called clones), among which little recombination of chromosomal genes occurs. Concomitantly, multilocus enzyme electrophoresis (MLEE) analyses revealed that there are only a few distinctive genotypes, despite the huge global genetic diversity, and showed furthermore that clones isolated from geographically and temporally distinct hosts were identical[32,33] (see BOX 1 for a description of typing methods). These MLEE experiments were initially performed to test whether the genetic variability in a haploid species would be as high as that in a diploid species, despite the absence of an overdominance contribution to the maintenance of polymorphism[32] in haploid populations. When framed in the clonal concept of a species, the MLEE experiments supported neutral theory, which stipulates that the vast majority of the molecular variability observed is not affected by natural selection but is neutral in terms of organism fitness[33]. Further MLEE studies, complemented by other techniques (such as biotyping[34], serotyping, outer-membrane protein electrophoretic analysis[35,36], random amplified polymorphic DNA and restriction fragment length polymorphism of ribosomal RNA gene regions[37]) revealed that the genetic markers were mutually corroborative, therefore reinforcing the clonal concept.

The sequence era: population structure and recombination. With the arrival of DNA sequence data for individual genes in the 1980s, studies could demonstrate recombination at the molecular level. The sequenced regions consisted mainly of the *trp* operon (which controls the biosynthesis of tryptophan in the cell) and the genes coding for the enzymes previously studied using MLEE. As early as 1983, Milkman and Crawford identified clustered base substitutions in the translated regions of the *trp* operon, which they interpreted as

Box 1 | **Tools for studying *Escherichia coli* population genetics**

Four main techniques have been used to study the genetic entities (that is, units of population structure) of *Escherichia coli*.

- Serotyping: this was developed in the 1940s by Kauffman[131], and the work was continued by Orskov[132]. Based on the combinations of 173 O antigens, 80 K antigens and 56 H antigens, an extremely high number of serotypes have been described[133]. Molecular alternatives based on PCR have now been developed, especially for the typing of O antigens[134].
- Multilocus enzyme electrophoresis: the 1980s saw the development of multilocus enzyme electrophoresis (MLEE) methods for studying bacteria[63]. Isolates are characterized by the relative electrophoretic mobility of several water-soluble housekeeping cellular enzymes. Mobility variants of an enzyme can be directly equated with alleles at the corresponding locus[63]. The alleles at each locus define an electrophoretic type, and the relatedness of isolates can be visualized on a dendrogram produced from a matrix of pairwise differences between the electrophoretic types.
- Multilocus sequence typing: in the late 1990s, multilocus sequence typing (MLST) emerged as a powerful tool for bacterial population genetics[135]. The nucleotide sequence of several housekeeping genes is determined for each isolate. The data can then be studied in two ways. They can be analysed similarly to MLEE data, such that the alleles at each locus are assigned on the basis of their sequence. The alleles at the different loci provide an allelic profile, which defines the sequence type (ST). No weighting is given to take into account the number of nucleotide differences between the alleles. The relatedness of isolates is displayed in a manner analogous to that in MLEE analysis. Alternatively, phylogenetic reconstructions can be performed from the nucleotide sequences, with or without corrections for recombination events[68,73]. Currently, three MLST schemas[67–69] are available for *E. coli*, using each a different combination of genes. The results obtained with these three schemas are highly correlated[71,136], arguing for the robustness of the clonal structure of the species.
- Phylogrouping triplex PCR: this approach allows strains to be assigned to one of the four main phylogenetic groups (A, B1, B2 and D)[137]. Since its introduction in 2000, it has become widely used owing to its simplicity and rapidity. The method, based on triplex PCR, uses the combination of two genes (*chuA*, the outer-membrane hemin receptor gene, and *yjaA*, which encodes an uncharacterized protein) and a DNA fragment that has been recently identified as part of a putative lipase esterase gene[71]. The accuracy with which this method assigns strains to their correct MLST-based phylogroup is good (80–85%)[71].

With the arrival of next-generation sequencing technology[138], it will soon be possible to study hundreds of strains to help understand at the whole-genome level the evolutionary processes acting in populations[139–141], opening the era of 'population genomics'.

possible recombination events[38]. The establishment of the *E. coli* reference collection (ECOR) in 1984 (REF. 39) provided researchers with a valuable tool for deciphering the population structure of commensal *E. coli* (BOX 2). Several studies then showed that the phylogenetic trees constructed from individual genes were incongruent with each other (that is, the groupings of the strains were not the same in trees constructed from different genes) and different from the species phylogeny taken from a consensus tree of the enzyme-coding genes or from the MLEE tree[40–43], and these differences are hallmarks of recombination at the gene level. Furthermore, the spatial distribution of the inconsistent sites was not random; it was, instead, highly clustered in a discontinuous series, suggesting that intragenic recombination had occurred. As proposed by Milkman, these recombined segments correspond to the 'clonal segments' (REF. 44), whereas the DNA that is inherited purely vertically from the clonal ancestor can be regarded as the 'clonal frame' (REF. 45) and can be used to reconstruct the species phylogeny. The sizes of the recombination events that were deduced from these sequence comparisons were small, at around 1 kb[43]. *In vitro* experiments of transduction and conjugation in *E. coli* found that large fragments of incoming DNA could be reduced into 1 kb fragments by restriction systems and exonucleases[46,47]. Moreover, the superimposition of large fragments corresponding to successive overlapping incorporations resulted in a mosaic of small segments[43].

These studies highlighted the importance of recombination and led to the suggestion that it could be much more frequent than mutation[48]. It was also acknowledged that recombination varied among the genes studied: some genes had no trace of recombination (for example, the glyceraldehyde-3-phosphate dehydrogenase A gene (*gapA*)[49], and three phosphotransferase system enzyme III genes encoding proteins that are specific for β-glucoside sugars (*chbA*; also known as *celC*), glucose (*crr*) and glucitol (*gutB*; also known as *srlB*)[50]), whereas others were highly recombined, such as the gene encoding gluconate-6-phosphate dehydrogenase (*gnd*)[41]. This high level of recombination of *gnd* is, in fact, related to its close proximity to the *rfb* operon that encodes the O antigen, which is under diversifying selection. It has been shown that a similar O antigen type could be shared by very different genotypes[35]. The *rfb* operon has been described as a "bastion of polymorphism" (REF. 51), as has the host specificity for DNA (*hsd*) operon coding for the type I restriction and modification systems that enable bacteria to distinguish 'foreign' DNA from their own[52].

The genomic era brings reconciliation: organized disorder. How can such a level of recombination be compatible with a clonal population structure? Thanks to the accumulation of whole-genome sequences (31 to date), it has become possible to carry out a large-scale analysis of homologous recombination in *E. coli*.

However, before many genomes were fully sequenced, the genomic era shifted the debate to another form of recombination: the acquisition and loss of genes, or horizontal gene transfer. Indeed, the most striking difference among strains at the genomic level is

Box 2 | The *Eschericha coli* reference collection (ECOR)

In 1984, Ochman and Selander established a set of 72 *Escherichia coli* strains isolated from the faeces of healthy human and zoo animal hosts or during human urinary tract infections, using hosts from a variety of geographical locations (mainly the United States and Sweden). They chose the strains from their collection of 2,600 *E. coli* natural isolates on the basis of the allelic diversity at 11 enzyme loci, which they took to represent the range of genotypic variation in the species as a whole[39]. This collection was widely distributed and rapidly became a common substrate for various phenotypic and genetic analyses all over the world. Now, 26 years and numerous studies later, we can conclude that they did indeed capture most of the variation in *E. coli*.

the variation in gene content[53,54]. This is illustrated by an analysis of genomes of 20 different *E. coli* strains. This analysis found that the average *E. coli* genome contains 4,721 genes, but only 2,000 genes with high homology are conserved among all strains; these conserved genes form the core genome[55]. Hence, each strain or group of strains contains some genes sampled from a pool that currently contains up to 8,000 genes (excluding insertion sequences, phage and phage-like elements, as well as core genes) but that keeps increasing with the number of strains studied[56,57] (FIG. 1). This provides a high level of plasticity in the genome, which results in a large diversity of adaptive paths. These observations changed the picture of the genome into that of a highly dynamic structure in which there is a constant flux of insertions and deletions[58]. Despite its dynamic nature, the genome keeps a strong structure as, apart from in *Shigella* spp. (and potential assembly errors), very few genome rearrangements have been observed. Indeed, most of the horizontal gene transfers seem to occur in particular hotspots of integration. Some of these hotspots correspond to tRNA or phage integration hotspots, as described previously, but most have no specific molecular signature to date[57].

Recombination in the core genome was analysed with quantitative estimates based on a coalescent framework and approximate Bayesian computations[59], and the observed pattern of linkage disequilibrium was compatible with a rate of genetic exchange that was twice as high as the mutation rate for short fragments of 50 bp[57]. In other words, a base has a 100-fold greater chance of being involved in genetic transfer than being mutated. This supports the strong role of recombination in genome evolution. However, because recombination events involve short fragments, simulations suggest that they will not alter the global topology of the phylogeny, provided that the sequence analysed to build the phylogeny is long enough. The conclusion is that, despite a greater rate of recombination than mutation, the mode of recombination (which requires a double crossing over, similar to gene conversion) is compatible with an apparent clonal population structure and a clear phylogenetic signal[57] (FIG. 2a) that appropriately reflects the relationship between strains.

In agreement with the previous gene analyses, genome-wide analysis revealed that variations in the frequency of recombination occurred along the genome. A large-scale pattern showed that recombination was lower around the terminus macrodomain, which is centred on the terminus of replication[57]. This presumably results from the low copy number of the terminus macrodomain in the cell and from its aggregation, which is mediated by the macrodomain Ter protein (MatP)–macrodomain Ter sequence (*matS*) association[60]. On a smaller scale, the two main bastions of polymorphism that had previously been reported[61] were identified using phylogenetic analysis; these are the *rfb* operon and the region containing both the *hsd* and mannose-sensitive type 1 pilus (*fim*) operons[57]. Finally, a link was found between some hotspots of integration and some hotspots of recombination. It seems that once a new cluster of genes is incorporated into the genome of a strain, the locus can spread through the species if it provides any advantage, owing to homologous recombination using the conserved flanking portions[57,62] (FIG. 2b,c).

Phylogenetic history and the genetic structure of the species. Apart from a few genes in regions in which the combination of selection and recombination strongly alters the phylogenetic signal, most combinations of genes can provide a phylogenetic signal that can be used to cluster strains in a relevant way. This is illustrated by the similarities in the results obtained from MLEE and different multilocus sequence typing (MLST) schemes used for this species to date.

An MLEE-based phenogram using 38 enzymes[63,64] identified four main groups (A, B1, B2 and D) and two accessory groups (C and E) in the species[65,66]. The concatenation of individual gene sequences obtained by MLST, with or without the removal of recombination events, also identified all these groups except group C[67–71]. These five groups were recovered using the 1,878 genes of the *Escherichia* spp. core genome and the 2.6 million nucleotides of the *E. coli* chromosomal backbone[57]. The use of *Escherichia fergusonii*, the closest relative of *E. coli*[72], instead of *Salmonella enterica* as the outgroup[57,73] limited the long-branch attraction artefact[74] and strongly supported the polyphyletic origin of group D.

This allowed a robust phylogeny to be built, in which the first split in the *E. coli* phylogenetic history leads to one branch containing the strains of group B2 and a subgroup within D that we called group F[75] and another branch containing the rest of the species[57]. The remaining strains of group D then emerge from this second branch, followed by group E. Finally, the A and B1 groups appear as sister groups[57,75] (FIG. 3). The B2 group exhibits the highest diversity at both the nucleotide and the gene content level[57], supporting its early emergence in the species lineage and suggesting that it has subspecies status[76]. This is further reinforced by the clear genetic structure observed in this group, which has at least nine phylogenetic subgroups that are well correlated with a flexible gene pool and, to a lesser extent, with the O antigen type[77].

The epidemiology of commensalism

Using the tools and technologies described above, numerous studies have furthered our understanding of the ecological structure of the *E. coli* population and

Coalescent framework
Coalescent theory is a retrospective model of population genetics. It builds the genealogy of gene copies isolated from a sample of individuals from a population back to a single ancestral copy (known as the most recent common ancestor).

Approximate Bayesian computation
A family of computational likelihood-free inference techniques that operate on summary data (such as population mean or variance) to make broad inferences. They are especially useful in situations in which evaluation of the likelihood is computationally prohibitive or whenever suitable likelihoods are not available.

Linkage disequilibrium
The non-random association of alleles at two or more loci. It describes a situation in which some combinations of alleles or genetic markers occur more or less frequently in a population than would be expected if there were a random association of alleles on the basis of their frequencies.

MLEE-based phenogram
A dendrogram resulting from hierarchical clustering that is computed from multilocus enzyme electrophoresis (MLEE) data.

Long-branch attraction artefact
The erroneous grouping of two or more long branches as sister groups due to methodological artefacts of phylogenetic reconstruction. In the case discussed in this Review, a distant out-group (*Salmonella enterica*) works as an attractor of long-branched in-group taxa.

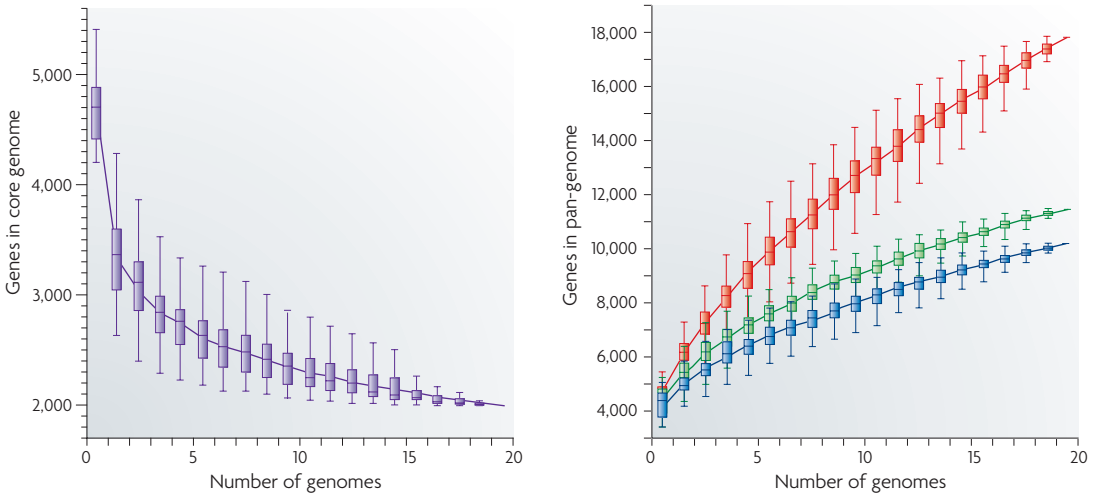

Figure 1 | ***Escherichia coli* core genome and pan-genome evolution according to the number of sequenced genomes.** For each data point, the upper and lower edges of the box indicate the first quartile and third quartile, respectively, of 1,000 random input orders of the genomes. The horizontal line indicates the sample median, and the vertical lines show the full range of the data, to a distance of at most 1.5 interquartile ranges (that is, the distance between the first and third quartile values). With 20 sequenced genomes, the core genome (left) comprises 1,976 genes and the pan-genome (right) comprises 17,838 genes in total (red), 11,432 of which show <80% similarity in sequence (green). If insertion sequence-like elements (found in 3,834 genes) and prophage-like elements (found in 3,873 genes) are discounted, then the pan-genome comprises 10,131 genes (blue). The studied strains corresponded mainly to pathogenic isolates[57]. Figure is reproduced from REF. 57.

started to answer simple questions such as whether individual strains are randomly spread among hosts or specific to their host.

Intra-host diversity. By studying numerous colonies (up to 300) per stool from some human individuals, it was established that at any one time each person commonly carries a predominant strain that constitutes more than half the colonies isolated, with the other strains present at various levels, and also that over time they carry a resident strain that is present for months or years as well as transient strains that are found for only a few days or weeks[78–82]. The predominant strain and the resident strain tend to be identical for a given individual[78,82]. More recently, it has been shown that the level of intra-host diversity is variable among human populations, with the highest diversity observed in populations living in tropical regions[83,84]. Domesticated animals were found to exhibit lower strain diversity than their wild counterparts[85].

Several epidemiological arguments have attributed this intra-host diversity to the recurrent migration of strains. Studies in human household members and pets have shown that strain sharing is more frequent within households than across households and that sexual partners share strains more commonly than other adults[86,87]. In tropical human populations or wild animals the lower hygienic quality of food might increase the rate of incoming *E. coli*. Experimental attempts to colonize adult humans with new orally administrated strains of *E. coli* for an extended period of time have been disappointing[78,80,88], whereas colonization is easier to achieve in neonates[89]. Moreover, an *E. coli* strain that fails to colonize the intestines of mice harbouring a full intestinal microbiota will colonize quite well when introduced into streptomycin-treated mice[90] or as a monocontaminant into germ-free animals[15]. Interestingly, in the case of germ-free animals, the bacterial strain will persist even after implantation of an indigenous microbiota[15]. This suggests an association between the resident strain and its host that is stronger than expected and in which the rest of the flora is involved.

Between-host diversity. Although pioneering studies based on MLEE analyses have shown that many clones have broad geographical and host distributions[39], the observed genetic diversity of *E. coli* exhibits both host taxonomic and environmental components[91,92]. This can be illustrated by the prevalence of the four main phylogenetic groups in various human and animal populations. In humans, strains of group A are predominant (40.5%), followed by B2 strains (25.5%), whereas B1 and D strains (17% each) are less common (these data were compiled from 1,117 subjects[83,93–100]). In animals, a predominance of B1 strains (41%), followed by A (22%), B2 (21%), and, to a lesser extent, D (16%) strains is observed (these data were compiled from 1,154 animals[7,39,85,101,102]) (TABLE 1).

This variation in the prevalence of phylogenetic groups among different hosts is not attributable to the existence of host-specific strains. Indeed, only a few strains seem to be host specific: some haemolysin-producing B1 strains, exhibiting distinct O antigen types and MLST profiles, have so far been found exclusively in animals[85], whereas an avirulent B2 clone of the O81 type has been reported only in humans[103]. Apart from these recent findings, however, MLST groups have no clear association patterns among the different hosts but do have variable prevalence.

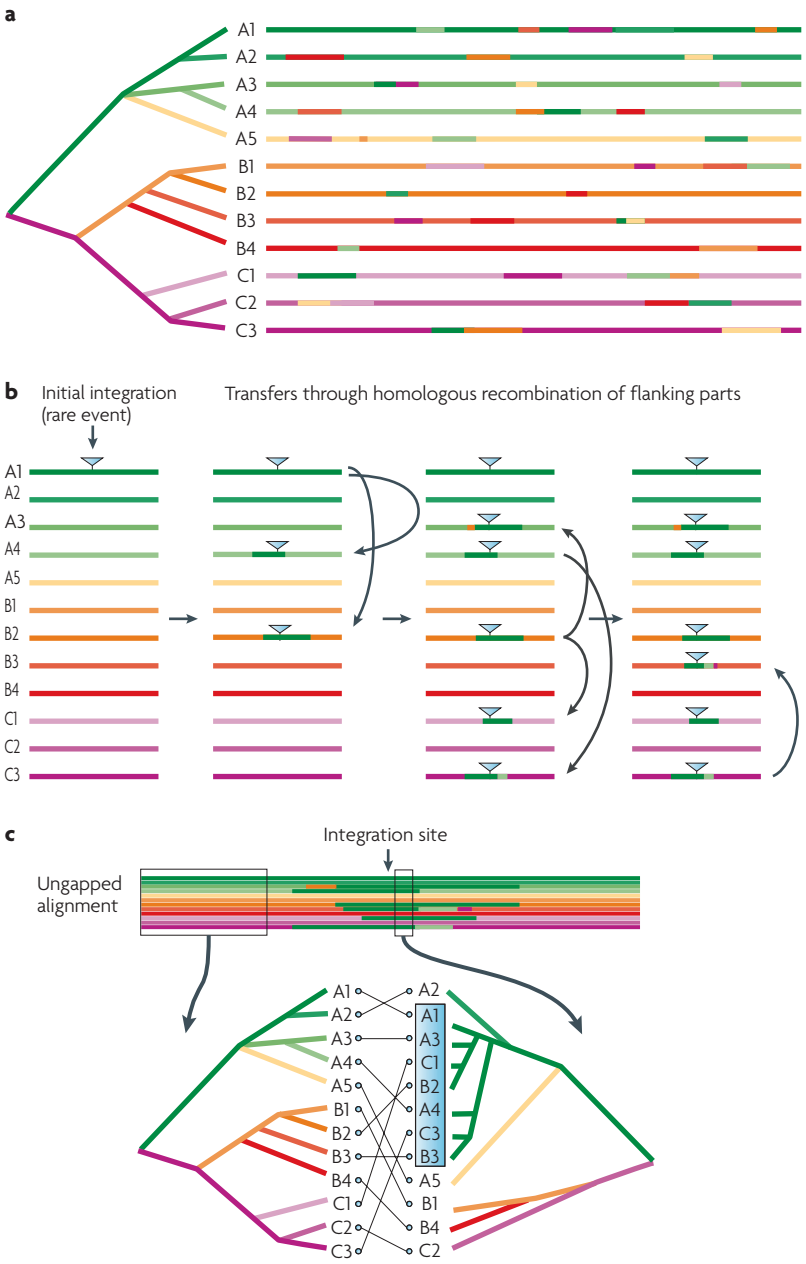

Figure 2 | **Phylogenetic reconstruction and recombination. a** | The short size of the fragments involved in recombination is not sufficient to blur the phylogenetic signal emerging from the non-recombining parts. Hence, provided that enough loci are studied, the topology of the phylogeny is recovered despite frequent recombination. **b** | The integration of a DNA fragment through horizontal gene transfer is a rare event. Once integrated, it can propagate through the species if it provides a selective advantage, thanks to homologous recombination of the flanking conserved regions. **c** | This results in phylogenetic incongruence (as illustrated by the non-congruent trees) between the global phylogeny (left tree) and the phylogeny derived from sequences around the integration site (right tree)[62].

Factors involved in the distribution of strains. Some recent studies have begun to unravel the determinants of the associations between strain type and host. However, this task is difficult, as even in a given species the variation is high. Animal hosts are even more complex owing to the huge diversity of animal populations (including both wild and domesticated mammals, birds and reptiles), the difficulty of sampling and the absence of health status data for wild animals. Furthermore, comparisons between human and animal strains are difficult, because ideally strains should be sampled in the same area and at the same time, a case that is rarely observed owing to the heterogeneity of the collections and the small number of studied strains. In spite of these difficulties, the factors shaping the observed genetic structure of the *E. coli* population can be schematically divided into host characteristics and environmental factors.

As illustrated previously, notable differences in the prevalence of a given strain are found between humans and animals. Indeed, host characteristics such as diet, gut morphology and body mass seem to be important predictors of the distribution of the phylogenetic groups[7,85]. For example, the physical complexity of the hindgut in the herbivores and omnivores seems to favour B2 strains[7]. Similarly, although much less documented, different niches might exist in the gut, as group A strains are more likely to be isolated from the upper gastrointestinal tract, and B1 strains from the faeces[104].

However, greater variability arises from the environment in which a given animal or human population lives. In animals, the main environmental force shaping the genetic structure of the *E. coli* gut population is the domestication status of the host[85]. Domesticated animals have a decreased proportion of B2 strains than their wild counterparts (from 30% in wild animals to 14% and 11% in farm and zoo animals, respectively) and an increased proportion of A strains (from 14% in wild animals to 27% and 26% in farm and zoo animals, respectively) (these data were compiled from 1,154 animals[7,39,85,102]).

Similarly, large changes in the prevalence of *E. coli* groups are found among different human populations. According to their *E. coli* group prevalence, human populations can be roughly split into two groups. Commensal strains isolated from Europe (France and Croatia) in the 1980s and from Africa (Mali and Benin), Asia (Pakistan), and South America (French Guiana, Colombia and Bolivia) belong mainly to the A (55%) and B1 (21%) phylogenetic groups, whereas strains from the D (14%) and B2 (10%) groups are uncommon (these data were compiled from 550 subjects[83,93,99,100]). Conversely, strains isolated from Europe (France and Sweden) in the 2000s and from North America (USA), Japan and Australia belong mainly to the B2 group (43%), followed by the A (24%), D (21%), and B1 (12%) groups (these data were compiled from 567 subjects[94–98]) (TABLE 2). The importance of the environment has been confirmed by comparing the *E. coli* microbiota of subjects who expatriated to French Guiana from metropolitan France with the microbiota of either metropolitan French residents or the natives of French Guiana. The *E. coli* microbiota of the expatriates was intermediate between those of the two other populations[84], with the same prevalence of group A as the French residents and the same prevalence of group B2 as the native residents. Socioeconomic factors, such as dietary habits and the level of hygiene, are presumably the main factors accounting for this phylogenetic group distribution, rather than geographical,

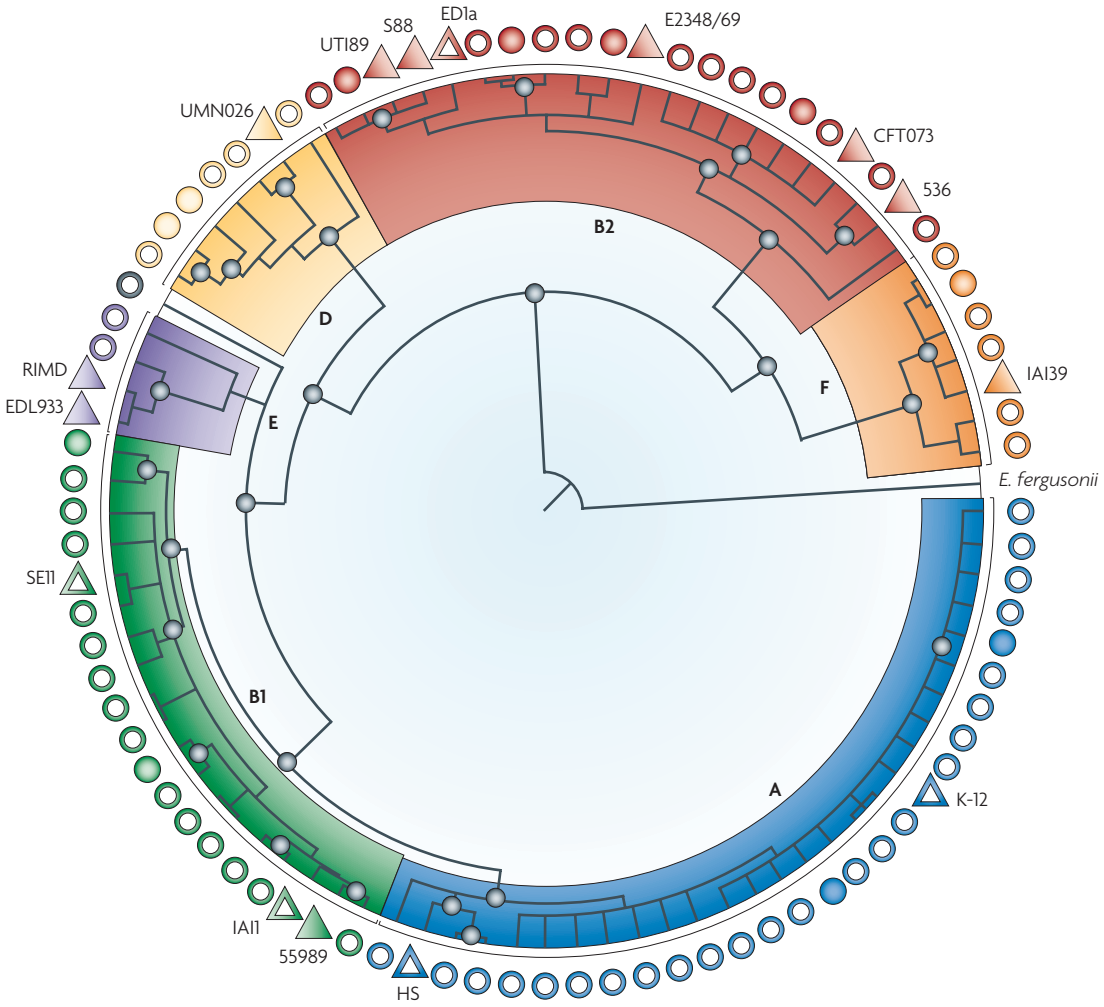

Figure 3 | **Phylogenetic history of *Escherichia coli*.** ClonalFrame[142] analysis based on the sequences of 8 housekeeping genes (4,095 nucleotides in total)[75], 72 strains from the *Escherichia coli* reference collection (ECOR)[39] (outer circles) and 15 genome reference strains[57,143,144] (outer triangles), rooted on *Escherichia fergusonii*. The commensal strains (61 ECOR strains and 5 sequenced strains (that is, strains K-12, HS, IAI1, SE11 and ED1a) are indicated by an open symbol, whereas the pathogenic strains are represented by a full symbol. Colours indicate the 6 main phylogenetic groups. Blue dots on nodes indicate that the clade defined by the node is monophyletic, with more than 80% support.

climatic or host genetic conditions, as indicated by the dramatic shifts in the proportions of B2 strains (from 10% to 30%) and A strains (from 60% to 30%) in France during the past 20 years (TABLE 2). Furthermore, the morphological, physiological and dietary differences that occur among human individuals of different sexes or ages influence the distribution of the *E. coli* genotypes[98].

Support and expression of diversity

Large-scale epidemiological studies provide insight into the diversity and complexity of *E. coli* niches. Here, we discuss some known molecular factors that sustain this diversity.

The coincidental hypothesis for 'virulence factors'. As *E. coli* must face variable environments, many different adaptive strategies can simultaneously exist in the species. The selective pressure in each host is intense, and the genome size of *E. coli* is finite. Therefore, the plasticity of the genome may illustrate the diversity of adaptive paths present in the species: some clusters of genes or genomic islands (including pathogenic islands[105]) should be found only in a subset of strains and favoured in some specific environments. In addition, several alternative combinations of genes could promote similar adaptations to a given environment.

Epidemiological data and experimental studies in animal models have identified and extensively studied genes that are associated with virulence: the so-called 'virulence genes'. There is now growing evidence that these virulence genes evolved and are maintained by selection for other roles that they have in the ecology of the bacteria, especially in commensalism[77,106]. Extraintestinal virulence genes coding for adhesins, iron capture systems, toxins and protectins have been correlated with successful gut colonization in humans[107–109], dogs[110] and piglets[111], whereas the adhesin intimin,

Table 2 | **Prevalence* of the main *Escherichia coli* groups in humans**

Population	Phylogenetic group‡				Ref.
	A	B1	D	B2	
France (1980)	61	12.5	16	10.5	93
Croatia	35	32	14	19	93
Mali	24	58	16	2	93
Benin	50	32.5	0	17.5	83
Pakistan	47	18	23	12	100
French Guiana (native populations)	63.5	20.5	13	3	83
Bolivia (native populations)	77	10	8	5	99
Colombia	57	3.5	14.5	25	83
France (2000)	25.5	21	24	29.5	83
Sweden	29	11	14	46	97
USA	20.5	12.5	19	48	94
Japan	28	0	28	44	95
Australia	19.5	12.5	23	45	98

*Prevalence is given as a percentage of the four phylogenetic groups of *Escherichia coli* in human faecal samples from the indicated populations. ‡Determined by triplex-PCR[137].

which is involved in intraintestinal pathogenicity, has been shown to be essential for the colonization of the bovine rectal mucosa[112]. Likewise, lipopolysaccharide (LPS)[113], Shiga toxins[114] and extraintestinal virulence genes[115] enhance survival by providing protection against predation by protozoa (such as amoebae[113] and *Tetrahymena* spp.[114]) or nematodes[115]. Finally, antigenic variability that is usually attributed to immune system selective pressure may also be driven by these predators and bacteriophages. Protozoan grazers such as *Entamoeba* spp. are common in commensal niches[116] and have been shown to attack strains differentially according to their O antigen type[117]. Similarly, bacteriophages can attack cells expressing LPS or many antigenic receptors, thereby promoting diversification and the maintenance of the resulting diversity[118].

The prevalence of these virulence factor genes is variable among commensal populations. On a global scale, the human microbiota is characterized by a higher prevalence of virulence genes than the microbiota in other organisms[85] (TABLE 1). In animals, the presence of virulence genes increases with body mass, which reflects the gut complexity of larger animal[85]. Hence, virulence factors and their change in prevalence among hosts may reflect some local adaptation to commensal habitats rather than virulence *per se*.

Intra-species interactions. Conspecific social interactions can also drive some diversification. To out-compete other clones, the production of colicins can be an efficient strategy in a structured environment[119]. This could allow maladapted strains to colonize the gut and, therefore, allow several clones to coexist in the long term. It may also promote diversification in a clone, as some strains may try to benefit from the production of the colicin but avoid paying the associated cost. Such interactions, which can be compared to the rock–paper–scissors game, have been described using game theory and can explain the maintenance of variable survival strategies[120]. Indeed, the secretion of any metabolite or enzyme can be described as an altruistic behaviour that can benefit some mutants or other strains lacking that component.

Antibiotic resistance. The commensal microbiota, and especially the intestinal microbiota, has been shown to have an important role in the emergence of antibiotic resistance[121]. A high density of bacteria with a large gene pool combined with a high environmental antibiotic exposure, due to the extensive use of antibiotics in both human and veterinary medicine, is an explosive cocktail for the selection of antibiotic resistance in the commensal microbiota.

E. coli isolates from several animal populations that were differentially exposed to human contact have been studied for antibiotic resistance and integron prevalence[101]. Integrons are molecular structures that are of great importance for the spread and expression of antibiotic resistance genes[122]. A clear positive correlation between the antibiotic resistance or integron prevalence in the bacteria and the host exposure to humans and human activities was observed[101] (TABLE 1). This variability of antibiotic resistance and integron prevalence with host environment was also observed among human populations[84,123], suggesting that the exposure of commensal *E. coli* populations to antibiotics shapes their diversity at the molecular level.

In addition to antibiotic exposure, the genetic background of the strain also seems to affect the patterns of antibiotic resistance. A group strains[124] and some D group[125] strains are particularly permissive to the development of resistance to third-generation cephalosporins. Conversely, B2 strains are less resistant than the remaining strains[126–128], regardless of the molecular mechanism involved in the acquisition of resistance (for example, a point mutation or gene acquisition), and have a lower prevalence of integrons in commensal *E. coli* strains from both human hosts[123] and animal hosts[101]. This could explain the relative decrease of B2 strains in domesticated animals in which antibiotics are used extensively.

Conclusions

Multiple factors, from both the host and the environment, shape the genetic structure of commensal *E. coli*. We are only beginning to decipher these factors, and clearly more studies are needed. Future studies should take place in ecologically well-characterized environments and should analyse *E. coli* from all hosts (humans and animals) and environments (water and sediments) together. Furthermore, in addition to the study of complete genomes of numerous isolates, metagenomic approaches should be developed to take into account the vast accompanying intestinal microbiota that is influenced by both host diet and phylogeny[129] and that has been largely ignored in defining the commensal niche of *E. coli*.

Colicin
A protein that is produced by and toxic for some strains of *E. coli*. Colicins result in the rapid elimination of neighbouring cells that are not resistant to their effects.

Moreover, some efforts are needed to understand the interactions between the host immune system and the commensal microbiota[130], as these interactions may vary from one host to another and may shape the *E. coli* commensal diversity. A better characterization of the commensal niche will also be necessary to understand how a useful commensal can become a harmful pathogen.

Note added in proof

A recent publication from the Whittam and Gordon laboratories[145] and data from our laboratory indicate a greater diversity than was previously reported for human and animal commensal '*E. coli*' strains, with at least five clades outside the classical *E. coli* strains that are represented by ECOR. Further work is needed to better characterize these strains.

1. Hobman, J. L., Penn, C. W. & Pallen, M. J. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol. Microbiol.* **64**, 881–885 (2007).
2. Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
3. Kosek, M., Bern, C. & Guerrant, R. L. The global burden of diarrhoeal disease, as estimated from studies published between 1992 and 2000. *Bull. World Health Organ.* **81**, 197–204 (2003).
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. Pinheiro da Silva, F. *et al.* CD16 promotes *Escherichia coli* sepsis through an FcRγ inhibitory pathway that prevents phagocytosis and facilitates inflammation. *Nature Med.* **13**, 1368–1374 (2007).
6. Berg, R. D. The indigenous gastrointestinal microflora. *Trends Microbiol.* **4**, 430–435 (1996).
7. Gordon, D. M. & Cowling, A. The distribution and genetic structure of *Escherichia coli* in Australian vertebrates: host and geographic effects. *Microbiology* **149**, 3575–3586 (2003).
A key study of 2,300 non-domesticated vertebrate hosts, describing the commensal *E. coli* population structure and identifying some of the forces that shape it.
8. Savageau, M. A. *Escherichia coli* habitats, cell types, and molecular mechanisms of gene control. *Am. Nat.* **122**, 732–744 (1983).
9. Solo-Gabriele, H. M., Wolfert, M. A., Desmarais, T. R. & Palmer, C. J. Sources of *Escherichia coli* in a coastal subtropical environment. *Appl. Environ. Microbiol.* **66**, 230–237 (2000).
10. Power, M. L., Littlefield-Wyer, J., Gordon, D. M., Veal, D. A. & Slade, M. B. Phenotypic and genotypic characterization of encapsulated *Escherichia coli* isolated from blooms in two Australian lakes. *Environ. Microbiol.* **7**, 631–640 (2005).
11. Mitsuoka, T. & Hayakawa, K. The fecal flora of man. I. Communication: the composition of the fecal flora of different age groups. *Zentralbl. Bakteriol. Orig. A.* **223**, 333–342 (1972).
12. Penders, J. *et al.* Factors influencing the composition of the intestinal microbiota in early infancy. *Pediatrics* **118**, 511–521 (2006).
13. Slanetz, L. W. & Bartley, C. H. Numbers of enterococci in water, sewage, and feces determined by the membrane filter technique with an improved medium. *J. Bacteriol.* **74**, 591–595 (1957).
14. Poulsen, L. K. *et al.* Spatial distribution of *Escherichia coli* in the mouse large intestine inferred from rRNA *in situ* hybridization. *Infect. Immun.* **62**, 5191–5194 (1994).
15. Freter, R., Brickner, H., Fekete, J., Vickerman, M. M. & Carey, K. E. Survival and implantation of *Escherichia coli* in the intestinal tract. *Infect. Immun.* **39**, 686–703 (1983).
16. Chang, D. E. *et al.* Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proc. Natl Acad. Sci. USA* **101**, 7427–7432 (2004).
A very convincing study using *E. coli* mutants in a mouse model of intestinal colonization, showing the nutritional requirements of *E. coli* in its natural primary habitat.
17. Peekhaus, N. & Conway, T. What's for dinner?: Entner-Doudoroff metabolism in *Escherichia coli*. *J. Bacteriol.* **180**, 3495–3502 (1998).
18. Jones, S. A. *et al.* Respiration of *Escherichia coli* in the mouse intestine. *Infect. Immun.* **75**, 4891–4899 (2007).
19. Licht, T. R., Tolker-Nielsen, T., Holmstrom, K., Krogfelt, K. A. & Molin, S. Inhibition of *Escherichia coli* precursor-16S rRNA processing by mouse intestinal contents. *Environ. Microbiol.* **1**, 23–32 (1999).
20. Rang, C. U. *et al.* Estimation of growth rates of *Escherichia coli* BJ4 in streptomycin-treated and previously germfree mice by *in situ* rRNA hybridization. *Clin. Diagn. Lab. Immunol.* **6**, 434–436 (1999).
21. Syed, S. A., Abrams, G. D. & Freter, R. Efficiency of various intestinal bacteria in assuming normal functions of enteric flora after association with germ-free mice. *Infect. Immun.* **2**, 376–386 (1970).
22. Bettelheim, K. A. & Lennox-King, S. M. The acquisition of *Escherichia coli* by new-born babies. *Infection* **4**, 174–179 (1976).
23. Nowrouzian, F. *et al. Escherichia coli* in infants' intestinal microflora: colonization rate, strain turnover, and virulence gene carriage. *Pediatr. Res.* **54**, 8–14 (2003).
24. Jaureguy, F. *et al.* Effects of intrapartum penicillin prophylaxis on intestinal bacterial colonization in infants. *J. Clin. Microbiol.* **42**, 5184–5188 (2004).
25. Conway, T., Krogfelt, K. A. & Cohen, P. S. The life of commensal *Escherichia coli* in the mammalian intestine. *EcoSal* [online], http://www.ecosal.org/index.php?option=com_content&view=article&id=552&searchword=The life of commensal&searchphrase=exact&Itemid=139 (2004).
26. Rastegar Lari, A., Gold, F., Borderon, J. C., Laugier, J. & Lafont, J. P. Implantation and *in vivo* antagonistic effects of antibiotic-susceptible *Escherichia coli* strains administered to premature newborns. *Biol. Neonate* **58**, 73–78 (1990).
27. Vollaard, E. J. & Clasener, H. A. Colonization resistance. *Antimicrob. Agents Chemother.* **38**, 409–414 (1994).
28. Hudault, S., Guignot, J. & Servin, A. L. *Escherichia coli* strains colonising the gastrointestinal tract protect germfree mice against *Salmonella typhimurium* infection. *Gut* **49**, 47–55 (2001).
29. Schamberger, G. P., Phillips, R. L., Jacobs, J. L. & Diez-Gonzalez, F. Reduction of *Escherichia coli* O157:H7 populations in cattle by addition of colicin E7-producing *E. coli* to feed. *Appl. Environ. Microbiol.* **70**, 6053–6060 (2004).
30. Smith, J. M., Smith, N. H., O'Rourke, M. & Spratt, B. G. How clonal are bacteria? *Proc. Natl Acad. Sci. USA* **90**, 4384–4388 (1993).
31. Orskov, F. *et al.* Special *Escherichia coli* serotypes among enterotoxigenic strains from diarrhoea in adults and children. *Med. Microbiol. Immunol.* **162**, 73–80 (1976).
32. Milkman, R. Electrophoretic variation in *Escherichia coli* from natural sources. *Science* **182**, 1024–1026 (1973).
33. Selander, R. K. & Levin, B. R. Genetic diversity and structure in *Escherichia coli* populations. *Science* **210**, 545–547 (1980).
A seminal work demonstrating the clonal structure of the *E. coli* population by MLEE analysis.
34. Miller, R. D. & Hartl, D. L. Biotyping confirms a nearly clonal population structure in *Escherichia coli*. *Evolution* **40**, 1–12 (1986).
35. Ochman, H. & Selander, R. K. Evidence for clonal population structure in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **81**, 198–201 (1984).
36. Caugant, D. A. *et al.* Genetic diversity in relation to serotype in *Escherichia coli*. *Infect. Immun.* **49**, 407–413 (1985).
37. 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).
38. Milkman, R. & Crawford, I. P. Clustered third-base substitutions among wild strains of *Escherichia coli*. *Science* **221**, 378–380 (1983).
39. Ochman, H. & Selander, R. K. Standard reference strains of *Escherichia coli* from natural populations. *J. Bacteriol.* **157**, 690–693 (1984).
The establishment of a reference collection that is representative of *E. coli* diversity and that has been and is still widely used around the word.
40. DuBose, R. F., Dykhuizen, D. E. & Hartl, D. L. Genetic exchange among natural isolates of bacteria: recombination within the *phoA* gene of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **85**, 7036–7040 (1988).
41. Bisercic, M., Feutrier, J. Y. & Reeves, P. R. Nucleotide sequences of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. *J. Bacteriol.* **173**, 3894–3900 (1991).
42. Dykhuizen, D. E. & Green, L. Recombination in *Escherichia coli* and the definition of biological species. *J. Bacteriol.* **173**, 7257–7268 (1991).
A cornerstone paper that demonstrates the presence of recombination in the *E. coli* species and proposes a definition of the bacterial species that is based on the biological species definition.
43. Milkman, R. & Bridges, M. M. Molecular evolution of the *Escherichia coli* chromosome. IV. Sequence comparisons. *Genetics* **133**, 455–468 (1993).
44. Milkman, R. & Stoltzfus, A. Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics* **120**, 359–366 (1988).
45. Milkman, R. & Bridges, M. M. Molecular evolution of the *Escherichia coli* chromosome. III. Clonal frames. *Genetics* **126**, 505–517 (1990).
46. McKane, M. & Milkman, R. Transduction, restriction and recombination patterns in *Escherichia coli*. *Genetics* **139**, 35–43 (1995).
47. Milkman, R. *et al.* Molecular evolution of the *Escherichia coli* chromosome. V. Recombination patterns among strains of diverse origin. *Genetics* **153**, 539–554 (1999).
48. Guttman, D. S. & Dykhuizen, D. E. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383 (1994).
49. Nelson, K., Whittam, T. S. & Selander, R. K. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **88**, 6667–6671 (1991).
50. Hall, B. G. & Sharp, P. M. Molecular population genetics of *Escherichia coli*: DNA sequence diversity at the *celC*, *crr*, and *gutB* loci of natural isolates. *Mol. Biol. Evol.* **9**, 654–665 (1992).
51. 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).
52. Barcus, V. A., Titheradge, A. J. & Murray, N. E. The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli*. *Genetics* **140**, 1187–1197 (1995).
53. Bergthorsson, U. & Ochman, H. Heterogeneity of genome sizes among natural isolates of *Escherichia coli*. *J. Bacteriol.* **177**, 5784–5789 (1995).
54. Bergthorsson, U. & Ochman, H. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* **15**, 6–16 (1998).
55. Hendrickson, H. Order and disorder during *Escherichia coli* divergence. *PLoS Genet.* **5**, e1000335 (2009).
56. 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).
57. Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *PLoS Genet.* **5**, e1000344 (2009).
A comprehensive paper that uses whole-genome sequences of *E. coli* to reconcile the occurrence of recombination events and the observed clonal struture of the population, thus allowing phylogenetic analyses to be carried out.

58. Ochman, H., Lawrence, J. G. & Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**, 299–304 (2000).
59. Beaumont, M. A., Zhang, W. & Balding, D. J. Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035 (2002).
60. Mercier, R. *et al.* The MatP/*matS* site-specific system organizes the terminus region of the *E. coli* chromosome into a macrodomain. *Cell* **135**, 475–485 (2008).
61. Milkman, R. Recombination and population structure in *Escherichia coli*. *Genetics* **146**, 745–750 (1997).
An insightful perspective by a visionary of the effect of recombination on the *E. coli* population structure that is still relevant today.
62. Schubert, S. *et al.* Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species. *PLoS Pathog.* **5**, e1000257 (2009).
63. Selander, R. K. *et al.* Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. *Appl. Environ. Microbiol.* **51**, 873–884 (1986).
64. Goullet, P. & Picard, B. Comparative electrophoretic polymorphism of esterases and other enzymes in *Escherichia coli*. *J. Gen. Microbiol.* **135**, 135–143 (1989).
65. Selander, R. K., Caugant, D. A. & Whittam, T. S. in Escherichia coli *and* Salmonella typhimurium. *Cellular and Molecular Biology*. (eds Neidhardt, F. C. *et al.*) 1625–1648 (American Society for Microbiology, Washington, DC, 1987).
66. Herzer, P. J., Inouye, S., Inouye, M. & Whittam, T. S. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. *J. Bacteriol.* **172**, 6175–6181 (1990).
67. 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).
68. Escobar-Paramo, P. *et al.* Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study. *Mol. Phylogenet. Evol.* **30**, 243–250 (2004).
69. Wirth, T. *et al.* Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol. Microbiol.* **60**, 1136–1151 (2006).
70. Johnson, J. R., Owens, K. L., Clabots, C. R., Weissman, S. J. & Cannon, S. B. Phylogenetic relationships among clonal groups of extraintestinal pathogenic *Escherichia coli* as assessed by multi-locus sequence analysis. *Microbes Infect.* **8**, 1702–1713 (2006).
71. Gordon, D. M., Clermont, O., Tolley, H. & Denamur, E. Assigning *Escherichia coli* strains to phylogenetic groups: multi-locus sequence typing versus the PCR triplex method. *Environ. Microbiol.* **10**, 2484–2496 (2008).
72. Lawrence, J. G., Ochman, H. & Hartl, D. L. Molecular and evolutionary relationships among enteric bacteria. *J. Gen. Microbiol.* **137**, 1911–1921 (1991).
73. Lecointre, G., Rachdi, L., Darlu, P. & Denamur, E. *Escherichia coli* molecular phylogeny using the incongruence length difference test. *Mol. Biol. Evol.* **15**, 1685–1695 (1998).
74. Felsenstein, J. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410 (1978).
75. Jaureguy, F. *et al.* Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* **9**, 560 (2008).
76. Lescat, M. *et al. aes*, the gene encoding the esterase B in *Escherichia coli*, is a powerful phylogenetic marker of the species. *BMC Microbiol.* **9**, 273 (2009).
77. 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).
78. Sears, H. J., Brownlee, I. & Uchiyama, J. K. Persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *J. Bacteriol.* **59**, 293–301 (1950).
79. Sears, H. J. & Brownlee, I. Further observations on the persistence of individual strains of *Escherichia coli* in the intestinal tract of man. *J. Bacteriol.* **63**, 47–57 (1952).
80. Sears, H. J., Janes, H., Saloum, R., Brownlee, I. & Lamoreaux, L. F. Persistence of individual strains of *Escherichia coli* in man and dog under varying conditions. *J. Bacteriol.* **71**, 370–372 (1956).
This and references 78 and 79 are a series of papers published in the 1950s describing the study of many clones from faecal specimens by O tying and introducing the notion of 'resident' and 'transient' strains.
81. Bettelheim, K. A., Faiers, M. & Shooter, R. A. Serotypes of *Escherichia coli* in normal stools. *Lancet* **2**, 1223–1224 (1972).
82. Caugant, D. A., Levin, B. R. & Selander, R. K. Genetic diversity and temporal variation in the *E. coli* population of a human host. *Genetics* **98**, 467–490 (1981).
The first (and unfortunately unique) paper to have used MLEE to thoroughly study the genetic structure of the commensal *E. coli* population in a human host over a 1 year period.
83. Escobar-Paramo, P. *et al.* Large-scale population structure of human commensal *Escherichia coli* isolates. *Appl. Environ. Microbiol.* **70**, 5698–5700 (2004).
84. Skurnik, D. *et al.* Characteristics of human intestinal *Escherichia coli* with changing environments. *Environ. Microbiol.* **10**, 2132–2137 (2008).
An elegant study using controlled human migration to show the effects of the environment on the *E. coli* microbiota.
85. Escobar-Paramo, P. *et al.* Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environ. Microbiol.* **8**, 1975–1984 (2006).
86. Caugant, D. A., Levin, B. R. & Selander, R. K. Distribution of multilocus genotypes of *Escherichia coli* within and between host families. *J. Hyg. (Lond.)* **92**, 377–384 (1984).
87. Johnson, J. R., Owens, K., Gajewski, A. & Clabots, C. *Escherichia coli* colonization patterns among human household members and pets, with attention to acute urinary tract infection. *J. Infect. Dis.* **197**, 218–224 (2008).
88. Cooke, E. M., Hettiaratchy, I. G. & Buck, A. C. Fate of ingested *Escherichia coli* in normal persons. *J. Med. Microbiol.* **5**, 361–369 (1972).
89. Poisson, D. M., Borderon, J. C., Amorim-Sena, J. C. & Laugier, J. Evolution of the barrier effects against an exogenous drug-sensitive *Escherichia coli* strain after single or repeated oral administration to newborns and infants aged up to three months admitted to an intensive-care unit. *Biol. Neonate* **49**, 1–7 (1986).
90. Myhal, M. L., Laux, D. C. & Cohen, P. S. Relative colonizing abilities of human fecal and K 12 strains of *Escherichia coli* in the large intestines of streptomycin-treated mice. *Eur. J. Clin. Microbiol.* **1**, 186–192 (1982).
91. Goullet, P. & Picard, B. Comparative esterase electrophoretic polymorphism of *Escherichia coli* isolates obtained from animal and human sources. *J. Gen. Microbiol.* **132**, 1843–1851 (1986).
92. Souza, V., Rocha, M., Valera, A. & Eguiarte, L. E. Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents. *Appl. Environ. Microbiol.* **65**, 3373–3385 (1999).
93. Duriez, P. *et al.* Commensal *Escherichia coli* isolates are phylogenetically distributed among geographically distinct human populations. *Microbiology* **147**, 1671–1676 (2001).
94. Zhang, L., Foxman, B. & Marrs, C. Both urinary and rectal *Escherichia coli* isolates are dominated by strains of phylogenetic group B2. *J. Clin. Microbiol.* **40**, 3951–3955 (2002).
95. Obata-Yasuoka, M., Ba-Thein, W., Tsukamoto, T., Yoshikawa, H. & Hayashi, H. Vaginal *Escherichia coli* share common virulence factor profiles, serotypes and phylogeny with other extraintestinal *E. coli*. *Microbiology* **148**, 2745–2752 (2002).
96. Watt, S. *et al. Escherichia coli* strains from pregnant women and neonates: intraspecies genetic distribution and prevalence of virulence factors. *J. Clin. Microbiol.* **41**, 1929–1935 (2003).
97. 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).
98. Gordon, D. M., Stern, S. E. & Collignon, P. J. Influence of the age and sex of human hosts on the distribution of *Escherichia coli* ECOR groups and virulence traits. *Microbiology* **151**, 15–23 (2005).
99. Pallecchi, L. *et al.* Population structure and resistance genes in antibiotic-resistant bacteria from a remote community with minimal antibiotic exposure. *Antimicrob. Agents Chemother.* **51**, 1179–1184 (2007).
100. Nowrouzian, F. L., Ostblom, A. E., Wold, A. E. & Adlerberth, I. Phylogenetic group B2 *Escherichia coli* strains from the bowel microbiota of Pakistani infants carry few virulence genes and lack the capacity for long-term persistence. *Clin. Microbiol. Infect.* **15**, 466–472 (2009).
101. Skurnik, D. *et al.* Effect of human vicinity on antimicrobial resistance and integrons in animal faecal *Escherichia coli*. *J. Antimicrob. Chemother.* **57**, 1215–1219 (2006).
A clear demonstration of the role of human contact in the emergence of antibiotic resistance in the *E. coli* microbiota of wild and domesticated animals.
102. Baldy-Chudzik, K., Mackiewicz, P. & Stosik, M. Phylogenetic background, virulence gene profiles, and genomic diversity in commensal *Escherichia coli* isolated from ten mammal species living in one zoo. *Vet. Microbiol.* **131**, 173–184 (2008).
103. Clermont, O. *et al.* Evidence for a human-specific *Escherichia coli* clone. *Environ. Microbiol.* **10**, 1000–1006 (2008).
104. Dixit, S. M. *et al.* Diversity analysis of commensal porcine *Escherichia coli* — associations between genotypes and habitat in the porcine gastrointestinal tract. *Microbiology* **150**, 1735–1740 (2004).
105. Hacker, J. & Kaper, J. B. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54**, 641–679 (2000).
106. Levin, B. R. The evolution and maintenance of virulence in microparasites. *Emerg. Infect. Dis.* **2**, 93–102 (1996).
107. Wold, A. E., Caugant, D. A., Lidin-Janson, G., de Man, P. & Svanborg, C. Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics. *J. Infect. Dis.* **165**, 46–52 (1992).
The first paper to convincingly show that the determinants involved in extraintestinal pathogenicity are associated with long-term persistence in the colon.
108. 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).
109. Moreno, E. *et al.* Structure and urovirulence characteristics of the fecal *Escherichia coli* population among healthy women. *Microbes Infect.* **11**, 274–280 (2009).
110. 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).
111. 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).
112. 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).
113. Alsam, S. *et al. Escherichia coli* interactions with *Acanthamoeba*: a symbiosis with environmental and clinical implications. *J. Med. Microbiol.* **55**, 689–694 (2006).
114. 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).
115. Diard, M. *et al. Caenorhabditis elegans* as a simple model to study phenotypic and genetic virulence determinants of extraintestinal pathogenic *Escherichia coli*. *Microbes Infect.* **9**, 214–223 (2007).
116. Samie, A. *et al.* Prevalence and species distribution of *E. histolytica* and *E. dispar* in the Venda region, Limpopo, South Africa. *Am. J. Trop. Med. Hyg.* **75**, 565–571 (2006).
117. Wildschutte, H., Wolfe, D. M., Tamewitz, A. & Lawrence, J. G. Protozoan predation, diversifying selection, and the evolution of antigenic diversity in *Salmonella*. *Proc. Natl Acad. Sci. USA* **101**, 10644–10649 (2004).
118. Calendar, R. (ed.) *The Bacteriophages* (Oxford Univ. Press, Oxford, UK, 2006).
119. Chao, L. & Levin, B. R. Structured habitats and the evolution of anticompetitor toxins in bacteria. *Proc. Natl Acad. Sci. USA* **78**, 6324–6328 (1981).
120. Kerr, B., Riley, M. A., Feldman, M. W. & Bohannan, B. J. Local dispersal promotes biodiversity in a real-life game of rock-paper-scissors. *Nature* **418**, 171–174 (2002).

121. Andremont, A. Commensal flora may play key role in spreading antibiotic resistance. *ASM News* **63**, 601–607 (2003).
122. Mazel, D. Integrons: agents of bacterial evolution. *Nature Rev. Microbiol.* **4**, 608–620 (2006).
123. Skurnik, D. *et al.* Integron-associated antibiotic resistance and phylogenetic grouping of *Escherichia coli* isolates from healthy subjects free of recent antibiotic exposure. *Antimicrob. Agents Chemother.* **49**, 3062–3065 (2005).
124. Mammeri, H., Galleni, M. & Nordmann, P. Role of the Ser-287-Asn replacement in the hydrolysis spectrum extension of AmpC β-lactamases in *Escherichia coli*. *Antimicrob. Agents Chemother.* **53**, 323–326 (2009).
125. Deschamps, C. *et al.* Multiple acquisitions of CTX-M plasmids in the rare D_2 genotype of *Escherichia coli* provide evidence for convergent evolution. *Microbiology* **155**, 1656–1668 (2009).
126. Picard, B. & Goullet, P. Correlation between electrophoretic types B_1 and B_2 of carboxylesterase B and sex of patients in *Escherichia coli* urinary tract infections. *Epidemiol. Infect.* **103**, 97–103 (1989).
127. Johnson, J. R. *et al.* Association of carboxylesterase B electrophoretic pattern with presence and expression of urovirulence factor determinants and antimicrobial resistance among strains of *Escherichia coli* that cause urosepsis. *Infect. Immun.* **59**, 2311–2315 (1991).
128. Johnson, J. R. *et al.* O, K, and H antigens predict virulence factors, carboxylesterase B pattern, antimicrobial resistance, and host compromise among *Escherichia coli* strains causing urosepsis. *J. Infect. Dis.* **169**, 119–126 (1994).
129. Ley, R. E. *et al.* Evolution of mammals and their gut microbes. *Science* **320**, 1647–1651 (2008).
130. Macpherson, A. J. *et al.* A primitive T cell-independent mechanism of intestinal mucosal IgA responses to commensal bacteria. *Science* **288**, 2222–2226 (2000).
131. Kauffmann, F. The serology of the coli group. *J. Immunol.* **57**, 71–100 (1947).
132. Orskov, F. & Orskov, I. in *Methods in Microbiology* (ed. Bergan, T.) 43–112 (Academic, London, 1984).
133. Orskov, F. & Orskov, I. *Escherichia coli* serotyping and disease in man and animals. *Can. J. Microbiol.* **38**, 699–704 (1992).
134. Clermont, O., Johnson, J. R., Menard, M. & Denamur, E. Determination of *Escherichia coli* O types by allele-specific polymerase chain reaction: application to the O types involved in human septicemia. *Diagn. Microbiol. Infect. Dis.* **57**, 129–136 (2007).
135. Enright, M. C. & Spratt, B. G. Multilocus sequence typing. *Trends Microbiol.* **7**, 482–487 (1999).
136. 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).
137. Clermont, O., Bonacorsi, S. & Bingen, E. Rapid and simple determination of the *Escherichia coli* phylogenetic group. *Appl. Environ. Microbiol.* **66**, 4555–4558 (2000).
138. Mardis, E. R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **24**, 133–141 (2008).
139. Liti, G. *et al.* Population genomics of domestic and wild yeasts. *Nature* **458**, 337–341 (2009).
140. Schacherer, J., Shapiro, J. A., Ruderfer, D. M. & Kruglyak, L. Comprehensive polymorphism survey elucidates population structure of *Saccharomyces cerevisiae*. *Nature* **458**, 342–345 (2009).
141. MacLean, D., Jones, J. D. & Studholme, D. J. Application of 'next-generation' sequencing technologies to microbial genetics. *Nature Rev. Microbiol.* **7**, 287–296 (2009).
142. Didelot, X. & Falush, D. Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**, 1251–1266 (2007).
143. Oshima, K. *et al.* Complete genome sequence and comparative analysis of the wild-type commensal *Escherichia coli* strain SE11 isolated from a healthy adult. *DNA Res.* **15**, 375–386 (2008).
144. Iguchi, A. *et al.* Complete genome sequence and comparative genome analysis of enteropathogenic *Escherichia coli* O127:H6 strain E2348/69. *J. Bacteriol.* **191**, 347–354 (2009).
145. Walk, S. T. *et al.* Cryptic lineages of the genus *Escherichia*. *Appl. Environ. Microbiol.* **75**, 6534–6544 (2009).

Acknowledgements
We are grateful to everyone who has helped us gather our strain collections over the years and continents and to all the members of our laboratory who have analysed these strains, especially P. Escobar-Páramo, T. Le Gall and O. Clermont. E.D. is partially funded by the Fondation pour la Recherche Médicale and O.T. is supported by the Agence Nationale de la Recherche. This review is dedicated to the memory of Thomas S. Whittam, a pioneer in *E. coli* population genetics, who died on 5 December 2008.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
Entrez Gene: http://www.ncbi.nlm.nih.gov/gene
chbA | *crr* | *gapA* | *gnd* | *gutB*
Entrez Genome Project: http://www.ncbi.nlm.nih.gov/genomeproj
Escherichia coli | *Escherichia coli* K-12 | *Escherichia fergusonii* | *Salmonella enterica*

FURTHER INFORMATION
Authors' homepage: http://www.bichat.inserm.fr/equipes/emi0339/u722.html
ECOR collection: http://foodsafe.msu.edu/whittam/ecor/index.html
MLST database at Michigan State University, USA: http://www.shigatox.net/ecmlst/cgi-bin/index
MLST database at Institut Pasteur Paris, France: http://www.pasteur.fr/recherche/genopole/PF8/mlst/EColi.html
MLST database at University College Cork, Ireland: http://mlst.ucc.ie

ALL LINKS ARE ACTIVE IN THE ONLINE PDF