The population genetics of commensal *E. coli* mcc ac

Abstract | The primary habitat of *Escherichia coli* is the vertebrate gut, where it is the predominant aerobic organism. The colonisation of vertebrates and a versatile pathogen, thought to kill more than 2 million humans per year through both intraintestinal and extraintestinal diseases. Indeed, 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.

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%, but it is only 56% in wild mammals, 23% in birds and 10% in reptiles. The concentration per gram of faeces varies in humans from 10^7 to 10^9 colony-forming units (cfu) and is much lower in domestic animals, averaging between 10^4 and 10^6 cfu.

*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. In the wild its total population size has been estimated to be 10^9 [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. As such, it is the perfect candidate to study the transition between commensalism and pathogenicity 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* is a Gram-negative, non-sporulating facultative anaerobe, is an inhabitant of the intestines and faeces of warm-blooded animals and reptiles. *E. coli* is found in the gut microbiota, which consists of more than 500 species of bacteria that total 10^14–10^16 cells per gram of large-intestinal content. Although the anaerobic bacteria in the bowel outnumber *E. coli* by 100 to 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. 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.

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%, but it is only 56% in wild mammals, 23% in birds and 10% in reptiles. The concentration per gram of faeces varies in humans from 10^7 to 10^9 colony-forming units (cfu) and is much lower in domestic animals, averaging between 10^4 and 10^6 cfu.
component and excreted in the faeces. The mucus defines a nutritional ecological niche to which *E. coli* metabolism has adapted. 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 glucuronic acid seems to have a predominant role. Although the concentrations of these sugars in the intestine are low, *E. coli* maximizes its growth by using micro-aerobic and anaerobic respiration in the intestine. This results in a 30-minute generation time in *vitro* on intestinal mucus compared with 40–80 minutes in the intestines of streptomycin-treated mice, in which the cells in the luminal content are static, and 120 minutes when the mice are 'conventionalized' by removing the streptomycin and feeding them with mouse caecal content. 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.

*E. coli* is among the first bacterial species to colonize the intestine during infancy, reaching very high density (higher than 10⁸ cfu per gram of faeces) before the expansion of anaerobes. After 2 years, the density stabilizes and remains at around 10⁸ cfu per gram of faeces until it gradually decreases in the elderly. The initial strains may originate from the maternal faecal microbiota and also from the maternity nursing staff. In fact, increased hygiene in hospitals and in families living in industrial countries has reduced early colonization by *E. coli*.

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, 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.

### A classical model in population genetics

To characterize how *E. coli* adapts to different communal 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. 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. 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 (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 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. Further MLEE studies, complemented by other techniques (such as biotyping, serotyping, outer-membrane protein electrophoretic analysis, random amplified polymorphic DNA and restriction fragment length polymorphism of ribosomal RNA gene regions) 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...
possible recombination events\(^4\). 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 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\(^40,41\). Currently, three MLST schemas\(^42-44\) are available for *E. coli*, using each a different combination of genes. The results obtained with these three schemas are highly correlated\(^21,131\), arguing for the robustness of the clonal structure of the species.

**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 Kaufman\(^45\), 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\(^135\). 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\(^136\). 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\(^137\). 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\(^40,41\). Currently, three MLST schemas\(^42-44\) are available for *E. coli*, using each a different combination of genes. The results obtained with these three schemas are highly correlated\(^21,131\), arguing for the robustness of the clonal structure of the species.
- **Phylogenotyping triple PCR:** this approach allows strains to be assigned to one of the four main phylogenetic groups (A, B1, B2 and D)\(^47\). Since its introduction in 2000, it has become widely used owing to its simplicity and rapidity. The method, based on triple 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\(^48\). The accuracy with which this method assigns strains to their correct MLST-based phylgroup is good (80–85%)\(^49\).

With the arrival of next-generation sequencing technology\(^148\), it will soon be possible to study hundreds of strains to help understand at the whole-genome level the evolutionary processes acting in populations\(^149-151\), opening the era of ‘population genomics’.

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
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.

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 identified four main groups (A, B1, B2 and D) and two accessory groups (C and E) in the species. 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. 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. The use of Escherichia fergusoni, the closest relative of E. coli, instead of Salmonella enterica as the outgroup limited the long-branch attraction artefact 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 E and another branch containing the rest of the species. 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. The B2 group exhibits the highest diversity at both the nucleotide and the gene content level, supporting its early emergence in the species lineage and suggesting that it has subspecies status. 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.

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...
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\(^{19-22}\). The predominant strain and the resident strain tend to be identical for a given individual\(^{19-22}\). 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\(^{23,34}\). Domesticated animals were found to exhibit lower strain diversity than their wild counterparts\(^{35}\).

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 the same household and that sexual partners share strains more commonly than other adults\(^{36-37}\). 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\(^{78,80,84}\), 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\(^{91}\). Interestingly, in the case of germ-free animals, the bacterial strain will persist even after implantation of an indigenous microbiota\(^{92}\). 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\(^{38}\), the observed genetic diversity of *E. coli* exhibits both host taxonomic and environmental components\(^{31,92}\). 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\(^{39,101-109}\)). 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\(^{39,90,101,105}\) [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, whereas colonization is easier to achieve in neonates\(^{89}\). Moreover, an *E. coli* strain that fails to colonize the intestines of mice harbouring...
Some 94–98 integrated, it can propagate through the species if it provides a selective advantage, once emerging from the non-recombining parts. Hence, provided that enough loci are studied, the phylogenetic signal 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) and the phylogeny derived from sequences around the integration site (right tree) is high. Animal hosts are even more complex owing to the herbivores and omnivores seems to favour B2 strains. 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.

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. 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.

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). 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). 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, 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, 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. For example, the physical complexity of the hindgut in the herbivores and omnivores seems to favour B2 strains. 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.

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...
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.

**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) 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.

**Figure 3** | Phylogenetic history of *Escherichia coli*. ClonalFrame analysis is based on the sequences of 8 housekeeping genes (4,095 nucleotides in total), 72 strains from the *Escherichia coli* reference collection (ECOR) (outer circles) and 15 genome reference strains (outer triangles), rooted on *Escherichia fergusonii*. The commensal strains (61 ECOR strains and 5 sequenced strains (that is, strains K-12, 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.
**REVISIONS**

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

<table>
<thead>
<tr>
<th>Population</th>
<th>Phylogenetic group</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B1</td>
</tr>
<tr>
<td>France (1980)</td>
<td>61</td>
<td>12.5</td>
</tr>
<tr>
<td>Croatia</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>Mali</td>
<td>24</td>
<td>58</td>
</tr>
<tr>
<td>Benin</td>
<td>50</td>
<td>32.5</td>
</tr>
<tr>
<td>Pak stan</td>
<td>47</td>
<td>18</td>
</tr>
<tr>
<td>French Guiana (nat ve populons)</td>
<td>63.5</td>
<td>20.5</td>
</tr>
<tr>
<td>Bolivia (nat ve populons)</td>
<td>77</td>
<td>10</td>
</tr>
<tr>
<td>Colombia</td>
<td>57</td>
<td>3.5</td>
</tr>
<tr>
<td>France (2000)</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>Sweden</td>
<td>29</td>
<td>11</td>
</tr>
<tr>
<td>USA</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Japan</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>Austral</td>
<td>19.5</td>
<td>12.5</td>
</tr>
</tbody>
</table>

*Prevalence s g ven as a percentage of the four phylogenetic groups of Escherichia coli in human faecal samples from the nd cated popul ons. †Determine by triplex-PCR.

which is involved in intraintestinal pathogenicity, has been shown to be essential for the colonization of the bovine rectal mucosa. Like-wise, lipopolysaccharide (LPS), Shiga toxins and extraintestinal virulence genes enhance survival by providing protection against predation by protozoa (such as amoebae and Tetrahymena spp.) or nematodes. 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 and have been shown to attack strains differentially according to their O antigen type. Similarly, bacteriophages can attack cells expressing LPS or many antigenic receptors, thereby promoting diversification and the maintenance of the resulting diversity.

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 (Table 1). In animals, the presence of virulence genes increases with body mass, which reflects the gut complexity of larger animals. 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. 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. 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. 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. Integrons are molecular structures that are of great importance for the spread and expression of antibiotic resistance genes. 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 (Table 1). This variability of antibiotic resistance and integron prevalence with host environment was also observed among human populations, 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 and some D group strains are particularly permissive to the development of resistance to third-generation cephalosporins. Conversely, B2 strains are less resistant than the remaining strains, 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 and animal hosts.

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 and that has been largely ignored in defining the commensal niche of E. coli.
Moreover, some efforts are needed to understand the interactions between the host immune system and the commensal microbiota, 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 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.
papers published in the 1950s describing the study of many clone of faecal specimens by O'Tyong and introducing the notion of 'resident' and 'transient' strains.


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 on the E. coli population structure that is still relevant today.


The first paper to convincingly show that the determinants involved in extraintestinal pathogenicity are associated with long-term persistence in the colon.


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
Saccharomyces Genome: http://www.yeastgenome.org
Salmonella enterica

FURTHER INFORMATION
Author homepage: http://shigatox.net/en/cit.top10.html
MLST database at Michigan State University, USA: http://www.mlst.smu.edu/mlst/salmonella_index.html
MLST database at University College Cork, Ireland: http://mlst.ucc.ie

ALL LINKS ARE ACTIVE IN THE ONLINE PDF