The decoupling between genetic structure and metabolic phenotypes in Escherichia coli leads to continuous phenotypic diversity

Victor Sabarly, O. Bouvet, J. Glodt, O. Clermont, D. Skurnik, L. Diancourt, D. de Vienne, E. Denamur, C. Dillmann

JOURNAL OF Evolutionary Biology

doi: 10.1111/j.1420-9101.2011.02287.x

The decoupling between genetic structure and metabolic phenotypes in *Escherichia coli* leads to continuous phenotypic diversity

V. SABARLY*†‡, O. BOUVET†, J. GLODT†, O. CLERMONT†, D. SKURNIK†, L. DIANCOURT§, D. DE VIENNE‡, E. DENAMUR† & C. DILLMANN‡

*DGA/CNRS, UMR de Génétique Végétale INRA/CNRS/Univ Paris-Sud, Ferme du Moulon, Gif-sur-Yvette, France
†INSERM U722 and Université Paris-Diderot, Faculté de Médecine, Site Xavier Bichat, Paris, France
‡Univ Paris-Sud, UMR de Génétique Végétale INRA/CNRS/Univ Paris-Sud, Ferme du Moulon, Gif-sur-Yvette, France
§Genotyping of Pathogens and Public Health, Institut Pasteur, Paris, France

Keywords:

carbon source;
E. coli;
genetic distance;
genetic group;
lifestyle;
metabolic pathway;
natural isolate;
phenotypic distance.

Abstract

To assess the extent of intra-species diversity and the links between phylogeny, lifestyle (habitat and pathogenicity) and phenotype, we assayed the growth yield on 95 carbon sources of 168 *Escherichia* strains. We also correlated the growth capacities of 14 *E. coli* strains with the presence/absence of enzyme-coding genes. Globally, we found that the genetic distance, based on multilocus sequence typing data, was a weak indicator of the metabolic phenotypic distance. Besides, lifestyle and phylogroup had almost no impact on the growth yield of non-*Shigella E. coli* strains. In these strains, the presence/absence of the metabolic pathways, which was linked to the phylogeny, explained most of the growth capacities. However, few discrepancies blurred the link between metabolic phenotypic distance and metabolic pathway distance. This study shows that a prokaryotic species structured into well-defined genetic and lifestyle groups can yet exhibit continuous phenotypic diversity, possibly caused by gene regulatory effects.

Introduction

Species have been first differentiated from morphological traits, and nowadays phenotypic criteria are still used to characterize them. Even for bacteria, phenotypic characteristics should agree with phylogenetic relatedness to constitute a species (Wayne *et al.*, 1987; Stackebrandt *et al.*, 2002). The underlying idea is that genetically distinct organisms should also be phenotypically distinct. Several cases, for which phylogeny, phenotype and ecological niche are related, support this view. For instance, in the group of asexual species of bdelloid rotifers, genetic and morphological clusters are the same and result from niche divergence (Fontaneto *et al.*, 2007). In bacteria of the genus *Bacillus*, genetic groups and growth temperature are also linked as a consequence of the ecology of these species (Guinebretière *et al.*, 2008).

However, several studies have revealed that the genetic distances and the phenotypes can be poorly related, as it has been found in eukaryotes species such as *Zea mays* (maize) (Burstin & Charcosset, 1997) and *Lolium perenne* (ryegrass) (Roldán-Ruiz *et al.*, 2001). Similar results have been observed for bacterial species such as members of the genus *Cronobacter* (Baldwin *et al.*, 2009) and strains of *Staphylococcus aureus* (Morandi *et al.*, 2010). A well-known phenomenon that can disrupt the link between genetic distance and phenotype is the phenotypic convergence resulting from similar ecological niches of distinct genetic groups. For instance, life-history strategies are associated with specific habitats in *Saccharomyces cerevisiae*, and genetically distant strains sharing the same habitat have similar life-history strategies (Spor *et al.*, 2009).

Correspondence: Christine Dillmann, UMR de Génétique Végétale, INRA – Univ Paris-Sud – CNRS, Ferme du Moulon, F-91190 Gif-sur-Yvette, France.
Tel.: +33 1 69 33 23 48; fax: +33 1 69 33 23 40;
e-mail: dillmann@moulon.inra.fr

The *E. coli* species is of particular interest to study the relationships between phylogenetic relatedness and phenotypic variation. The evolutionary history of the species (Lecointre *et al.*, 1998) revealed that the strains are distributed among five main phylogroups: A, B1, B2, D and E (Herzer *et al.*, 1990; Escobar-Páramo *et al.*, 2004a). In addition, natural isolates of *E. coli* are found in a variety of habitats, which can be either vertebrate hosts or water or soil (Hartl & Dykhuizen, 1984) and can be commensals (Tenaillon *et al.*, 2010), intra-intestinal pathogens (Intestinal Pathogenic *E. coli* or InPEC) or extra-intestinal pathogens (extra-intestinal pathogenic *E. coli* or ExPEC) (Kaper *et al.*, 2004). We chose to call lifestyles the combinations of habitat and pathogenicity. The prevalence of the different phylogroups varies slightly between lifestyles. For instance, farm animals exhibit a higher proportion of A and B1 strains and a lower proportion of B2 and D strains than wild animals. Likewise, ExPEC strains belong mainly to the phylogroup B2 (Picard *et al.*, 1999). However, there is no clear-cut link between phylogroups and lifestyles, i.e. no lifestyle can be uniquely attributed to a given phylogroup (Gordon & Cowling, 2003; Escobar-Páramo *et al.*, 2006). *E. coli* genome, which encompasses approximately 4700 genes, is highly dynamic: the core-genome, the genes present in all the sequenced genomes, is about 2000 genes, whereas the pan genome, the full set of nonorthologous genes among all genomes, reaches 18 000 genes (Rasko *et al.*, 2008; Touchon *et al.*, 2009).

Based on this large genetic diversity and the various lifestyles, we expect to find a large phenotypic variation within the species. The nonrandom distribution of the phylogroups among different lifestyles may indicate that these groups differ in phenotypes. Besides, as anthropogenic factors such as domestication play a major role in the ecological structure and the level of antimicrobial resistance of *E. coli* (Escobar-Páramo *et al.*, 2006; Skurnik *et al.*, 2006), the exposure of a strain animal host to humans could influence the phenotype of the bacterium. The prevalence of *E. coli* and the relative abundance of the phylogroups depend on the host diet (Gordon & Cowling, 2003), which might also have an impact on the strain phenotype. Finally, the strain phenotype could be globally linked to the pathogenic nature of the bacterium as this has been shown to be the case for a given metabolic phenotypic character. Indeed, the use of deoxyribose constitutes a fitness advantage for the competitiveness of extra-intestinal pathogenic *E. coli* strains (Bernier-Febreau *et al.*, 2004; Martinez-Jéhanne *et al.*, 2009).

To assess the extent of intra-species diversity as well as the links between phylogeny, lifestyle and phenotype, we assayed the growth yield (carbon source utilization) of a panel of genetically diverse *E. coli* natural isolates. We included several phylogenetic outgroups in the study as well as one phenotypic outgroup to test whether our methodology gives a global and representative image of a strain phenotype. Metabolic capacities are conditioned by the occurrence of specific enzymatic reactions in the cell that can be inferred from the strain gene content. Therefore, to go further, we studied in a subset of strains the relationship between growth capacities and metabolic pathways reconstructed from complete genome data. Hence, we were able to analyse the correlations between phylogenetic distance, metabolic phenotypic distance and metabolic pathway presence. Overall, the strain growth yield seemed to present continuous variations around the species average, whereas the pattern of the presence/absence of the metabolic pathways was linked to the species phylogeny. Finally, we discussed the impact of the species life cycle on the metabolic phenotypic diversity and the molecular mechanisms that could account for discrepancies between growth and the presence of metabolic pathways.

Materials and methods

Bacterial strains

The growth experiments were conducted on 168 bacterial strains comprising 159 *E. coli/Shigella* strains, six cryptic *Escherichia* clade strains, two *E. fergusonii* strains and one *E. albertii* strain. *E. fergusonii*, *E. albertii* and cryptic *Escherichia* clade strains were used as phylogenetic outgroups. The cryptic *Escherichia* clades are *Escherichia* lineages that have recently been reported. Strains belonging to these clades are very divergent from *E. coli* based on DNA sequence data; however, no biochemical feature allowed distinguishing them from *E. coli* (Walk *et al.*, 2009). The non-*Shigella E. coli* strains were chosen as representative of the genetic diversity of the species based on the triplex PCR phylogrouping (Clermont *et al.*, 2000) and multilocus sequence typing (MLST) data from more than 4000 isolates from various collections (Picard *et al.*, 1999; Escobar-Páramo *et al.*, 2004a,b, 2006; Clermont *et al.*, 2011). To have four groups of comparable genetic diversity, we chose to make one genetic group, A/B1, from the close A and B1 phylogroups. One hundred and fifty *E. coli* strains belonged to the genetic groups A/B1 (75 strains), B2 (38 strains), D (26 strains) and E (11 strains). Moreover, three strains did not belong to any group and were thus labelled 'ungrouped'. We also included six *Shigella* strains distributed into different *Shigella*-specific phylogroups (two in S1, one in S2, one in S3, one in SD1 and one in SS [Pupo *et al.*, 2000; Escobar-Páramo *et al.*, 2003]). These strains were used as phenotypic outgroup. Indeed, *Shigella* strains are intra-cellular human-specific pathogens that emerged from different *E. coli* phylogroups but present similar distinctive biochemical features as a consequence of their common lifestyle (Pupo *et al.*, 2000; Escobar-Páramo *et al.*, 2003). The 153 non-*Shigella E. coli*

strains were divided into several lifestyle groups: three pathogenic groups (commensal [90 strains], ExPEC [28 strains] and InPEC [35 strains]), four host anthropogenic groups (according to their exposure to humans: humans [53 strains], pet dogs [19 strains], farm animals [49 strains] and wildlife animals [32 strains]) (Skurnik *et al.*, 2006) and four host diets (insectivorous and granivorous birds [19 strains], carnivorous mammals [24 strains], herbivorous mammals [39 strains] and omnivorous mammals [71 strains]). The strains were selected to have comparable genetic diversity in the different lifestyle groups. The study on the relationship between growth capacities and metabolic pathways was conducted on a subset of 13 commensal and pathogenic *E. coli* strains for which the complete genome sequence was available (http://www.genoscope.cns.fr/agc/microscope/) as well as on the laboratory strain K-12. The main characteristics of all the strains are given Table S1. For each strain, the reference stock was conserved at −80 °C with glycerol.

Growth assays

Cells from the stock were grown overnight in Luria–Bertani broth at 37 °C then pelleted and washed once with minimal buffer (100 mM NaCl, 30 mM triethanolamine HCl, 5 mM NH_4Cl, 2 mM NaH_2PO_4, 0.25 mM Na_2SO_4, 0.05 mM $MgCl_2$, 1 mM KCl, 1 μM $FeCl_3$ and pH 7.1) and finally resuspended in minimal buffer. Growth capacities were assayed using commercially available Biolog GN2 microplates (AES Chemunex, Combourg, France). Each of the 96 wells of a Biolog GN2 microplate contains a simple carbon source presented Fig. S1, except one used as control, and a tetrazolium dye, which is an indicator of oxidative carbon metabolism correlated with bacterial growth (MacLean & Bell, 2002, 2003; MacLean *et al.*, 2004; Venail *et al.*, 2008). Each well of the Biolog GN2 microplates was inoculated with 100 μL of cell suspension diluted at an optical density (OD) of 0.03 measured on an Ultrospec 1100 pro spectrophotometer.

We measured the OD at 750 nm with a Tecan Infinite M200 plate reader after 18 h of growth at 37 °C in an incubator where the plates were shaken. We then subtracted the blank value (OD reached in the control well) to the OD after culture in each well. We called this value the growth yield. Experiments were conducted at eight different dates with a block design for the strains and 14 strains were replicated twice. Growth yield was corrected for a date effect using its least-square mean value computed by an analysis of variance (ANOVA) comprising five factors: the date of the assay, the phylogenetic group of the strain, its pathogenic group, the strain host anthropogenic group and its diet. The residual of the model contained both experimental error and genetic variation between strains of the same group. A separate analysis was conducted for each carbon source (see section Statistical analyses). All subsequent analyses were performed on growth yield corrected for the date effect. To determine a threshold above which growth was considered to be positive, we applied Gaussian mixture models to the growth yields (Fraley & Raftery, 2002, 2006). The optimal model according to the Bayesian information criterion (BIC) had three components: two of them with an average OD close to zero and the third one with an average OD close to one (Fig. S2). We chose to consider growth as positive whenever the corrected OD belonged to the third population with a 5% false-positive rate. Hence, positive growth corresponded to a growth yield >0.3388 OD units.

Phylogeny and genetic divergence

To estimate the genetic divergence between strains, we used the MLST data generated from eight partial genes: *din*B (450 bp), *icd*A (516 bp), *pab*B (468 bp), *pol*B (450 bp), *put*P (456 bp), *trp*A (561 bp), *trp*B (594 bp) and *uid*A (600 bp) (Jaureguy *et al.*, 2008; http://www.pasteur.fr/recherche/genopole/PF8/mlst/EColi.html). The phylogenetic tree was inferred with the software PhyML 3.0 (Guindon & Gascuel, 2003) using a generalized time-reversible (GTR) model with optimized equilibrium frequencies, estimated proportion of invariable sites, four substitution rate categories using the mean as the centre of each class and estimated gamma distribution parameter. The tree topology was optimized to maximize the likelihood using the nearest neighbour interchanges (NNIs) tree topology search operation with no random starting tree and a neighbour-joining input tree. The tree was plotted with the R package APE (Paradis *et al.*, 2004). The genetic divergence (d_G) is the distance between strains derived from this phylogenetic tree by the R package APE.

Statistical analyses

Analyses of variance were performed on each substrate for which growth was positive for at least one strain. The growth yield was analysed using a linear model with four main effects: the phylogenetic group of the strain, its pathogenic group, the strain host anthropogenic group and its diet. Type-III ANOVA tables were computed using the R package car (Fox & Weisberg, 2010). *P*-values of the *F*-tests from the ANOVA tables were cumulated, and the effects for which the false-positive discovery rate, FDR (Benjamini & Hochberg, 1995; Strimmer, 2009), was <0.1 % were considered significant. For those effects, we computed the least-square means and corresponding error variance for each group. The growth yield was also used to determine the metabolic phenotypic distance (d_P) between strains, the Euclidean distance between vectors of growth yields and to run a principal component analysis (PCA) computed with the software R (R Development Core Team, 2009). All the

Mantel tests between the different distances were performed using the R package ade4 (Dray & Dufour, 2007).

Metabolic pathways

In the study on the relationship between growth capacities and metabolic pathways, which comprised fewer strains (14 strains for which the complete genome sequence was available), the growth assay procedure was the same than for the other growth experiments except that the OD at 750 nm of each well of the microplates were monitored every 25 min during the 18 h of growth at 37 °C in a Tecan Infinite M200 plate reader where the plates were also shaken. The whole process (overnight growth and microplating) was repeated at least twice on different days. Thus, for each carbon source, growth was represented by two to three curves. The growth yield was estimated from the growth curve after performing a cubic spline interpolation using the software R (R Development Core Team, 2009). It corresponded to the amplitude of growth, i.e. to the OD reached after 18 h of growth minus the initial OD. This procedure minimized the biochemical assay errors to compare the metabolic capabilities of the strains with their gene contents. Growth was considered positive if the growth yield, averaged on the replicates, was greater than the growth threshold (0.2578 OD units), determined using Gaussian mixture models as in the other growth experiments.

The metabolic pathways present in the sequenced strains were recovered using the metabolic profiles from the Microcyc website (http://www.genoscope.cns.fr/agc/microcyc). The process to determine these metabolic profiles is as described in Vieira *et al.*, 2011. For one strain, each pathway was represented by its completion percentage. For example, a pathway for which all the enzyme-coding genes are present in the genome has a completion of 1, if half the enzyme-coding genes are missing, the completion is 0.5, and 0 if the pathway was not inferred in the strain. We defined the metabolic pathway completion distance between two strains (d_M) as the Euclidean distance between their vectors of pathway completions. To link the carbon sources allowing growth of at least one of the 14 sequenced strains to the metabolic pathways specifically involved in their degradation, we first selected all the pathways where the carbon source intervened as substrate or product of a reaction. Then, among this first selection of pathways, we manually removed those not involved in the degradation of the carbon source of interest. To link a maximum of carbon sources to pathways, we manually added four pathways because they involved reactions not classified as part of a pathway or because the reactions were not described yet in the Metacyc 13.0 database. These pathways concerned the following substrates: *N*-acetyl-D-galactosamine (enzymes: N-acetylglucosamine-6-phosphate deacetylase, EC 3.5.1.25; galactosamine-6-phosphate isomerase, no EC; 6-phosphofructokinase I, EC 2.7.1.11; tagatose 6-phosphate aldolase 1, EC 4.1.2.40 [Mukherjee *et al.*, 2008]), lactulose (enzyme: cryptic beta-D-galactosidase, EC 3.2.1.23), D-serine (enzyme: D-serine ammonia-lyase, EC 4.3.1.18) and D-raffinose (enzyme: alpha-galactosidase, EC 3.2.1.22). We successfully matched 43 carbon sources to their degradation pathways (Table S2) but we were unable to relate the consumption of glycyl-L-aspartic acid, L-alanyl-glycine and methylpyruvate to any metabolic pathway.

Phenotypic and metabolic distance models

We implemented a simplified model for the relationship between the metabolic phenotypic distance (d_P) between two strains and their genetic distance (d_G), defined here as the proportion of genes that are not identical by descent between the two strains. Some of the genetic differences can also be because of horizontal gene transfers independently of their phylogeny with a probability Λ. Moreover, only a fraction of the genetic differences cause gene inactivation. We called μ the probability that a genetic difference did not change the gene functionality. Therefore, the probability p_M that two genes had a functional difference was

$$p_M = (1-\mu)(d_G + \Lambda). \quad (1)$$

Genetic differences may not always translate into phenotypic differences. Here, the phenotypic observation is the growth ($P = 1$) or absence of growth ($P = 0$) on a given carbon source. We supposed that all n genes of the pathway needed to be functional for the pathway to be functional ($M = 1$). Hence, the probability that two strains had a functional difference ($\Delta M \neq 0$) for a given carbon source was

$$P(\Delta M \neq 0) = 1 - (1 - p_M)^n. \quad (2)$$

Our lack of knowledge on the metabolic network as well as differences in the gene regulatory network can lead to unexpected phenotypes according to the pathway functionalities. We defined the parameter δ_M as the probability that two strains share a common phenotype ($\Delta P = 0$) on a given substrate while having different pathway functionalities ($\Delta M \neq 0$) concerning this carbon source: $\delta_M = P(\Delta P = 0 \mid \Delta M \neq 0)$. Similarly, δ_P was the probability that two strains have different phenotypes ($\Delta P \neq 0$) while having the same pathway functionalities ($\Delta M = 0$): $\delta_P = P(\Delta P \neq 0 \mid \Delta M = 0)$. Thus, the probability that two strains have different growth capacities on a carbon source was

$$p_P = (1-\delta_M)P(\Delta M \neq 0) + \delta_P P(\Delta M = 0). \quad (3)$$

Monte Carlo simulations were performed to assess the relationship between genetic and phenotypic distances using parameters taken from our experimental data. We used the number of genes implied in each of the

395 metabolic pathways recovered in the 14 sequenced strains we studied. The proportion of genes that are not identical between two strains is proportional to the genetic distance between these strains and were consequently drawn in an uniform distribution on an interval corresponding to the observed values for our data (between zero and 0.25). We computed p_M between 10 000 strain pairs with $\mu = 0.83$ (Patel & Loeb, 2000) and $\Lambda = 0.13$ (Ochman *et al.*, 2000). For each pathway, the number of genes having different functionalities between two strains was drawn in a binomial distribution with a probability p_M and a number of trials equal the number of genes implied in the pathway. The metabolic pathway completion distance between two strains (d_M), defined as the Euclidean distance between vectors of pathway completions, was then calculated, as well as the vector of differences for pathway functionalities ΔM. The metabolic phenotypic distance between two strains (d_P), defined as the Euclidean distance between vectors of qualitative growth status, was computed as the square root of the sum of two random variables following binomial laws: the first one of probability $1 - \delta_M$ on all the carbon sources for which the pathway functionalities differed between the two strains ($\Delta M \neq 0$), and the second one of probability δ_P on all the carbon sources corresponding to pathways having the same functionality ($\Delta M = 0$) (eqn 3). We determined δ_M and δ_P using the metabolic pathway completion and growth data for each strain couple of the 14 sequenced strains and took the average values as estimates: $\delta_M = 0.63$ and $\delta_P = 0.21$. Moderate changes of the parameter values (μ, Λ, δ_M and δ_P) did not significantly change the simulation output (data not shown).

Results

The genetic distance is a weak indicator of the metabolic phenotypic distance

Our strain sample consisted in 159 *E. coli* strains (comprising six *Shigella* strains), six cryptic *Escherichia* clade strains, two *E. fergusonii* strains and one *E. albertii* strain. The phylogenetic tree of these strains shows that the non-*Shigella E. coli* strains constitute four distinct genetic groups (A/B1, B2, D and E) (Fig. 1). To estimate their metabolic phenotypic diversity, we assessed their growth yield on 95 different carbon sources. Figure 2 represents the plot of the metabolic phenotypic distance (d_P) vs. the genetic distance (d_G) between couples of strains. As expected, *E. fergusonii* strains as well as *E. albertii* strains are clearly distant both genetically and phenotypically, whereas cryptic *Escherichia* clade strains are quite

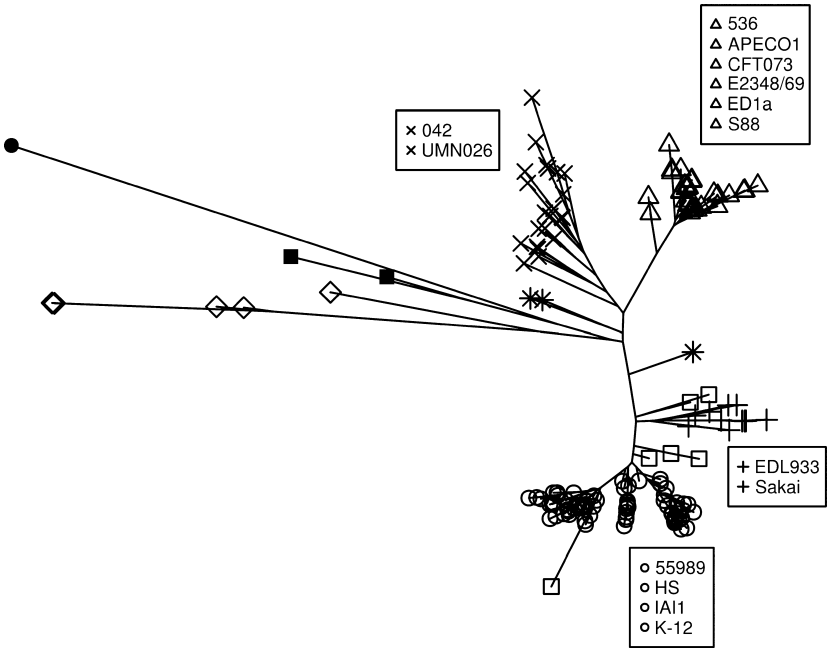

Fig. 1 Phylogenetic tree of 169 *Escherichia* strains reconstructed from the partial sequences of eight housekeeping genes by maximum likelihood. Non-*Shigella E. coli* strains are divided into four genetic groups, A/B1 (o), B2 (Δ), D (×) and E (+), as well as into an ungrouped category (*). The *Shigella* strains, although not monophyletic, are considered as a specific group represented by empty squares (□); they belong to particular phylogroups (S1, S2, S3, SD1 and SS). Cryptic *Escherichia* clade strains are indicated by diamonds (◇), *E. fergusonii* strains by filled squares (■) and the *E. albertii* strain by a filled circle (●). The names of the 14 sequenced strains used in the metabolic pathway study are given in the boxes. This phylogeny is in agreement with the one obtained using complete genome sequences (Touchon *et al.*, 2009).

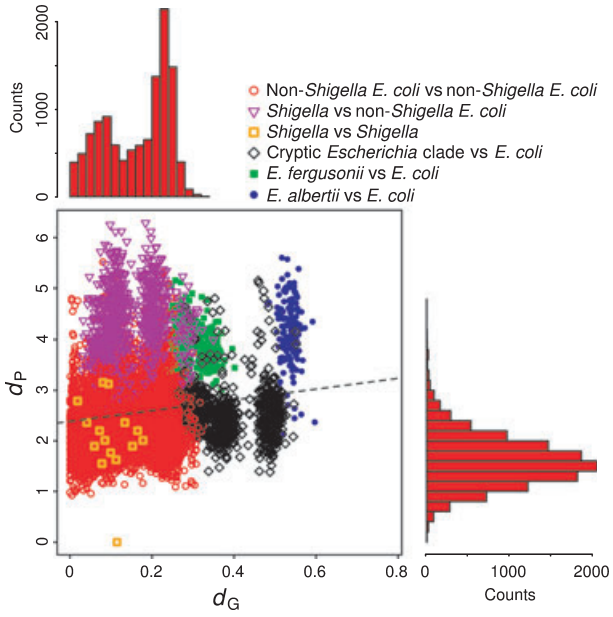

Fig. 2 Relationship between the metabolic phenotypic distance, d_P, and the genetic distance, d_G, resulting from comparisons between 159 *E. coli* strains (comprising six *Shigella* strains), six cryptic *Escherichia* clade strains, two *E. fergusonii* strains and one *E. albertii* strain. Only the comparisons involving at least one *E. coli* strain were considered. The dashed line corresponds to the regression of d_P according to d_G taking into account all represented strain pairs. The histograms represent the distributions of d_G (on top) and d_P (on the right) for the comparisons between two non-*Shigella E. coli* strains corresponding to the red circles on the plot.

divergent genetically but not phenotypically. That is why they have only recently been uncovered although they are genetically very divergent from *E. coli* (Walk *et al.*, 2009). On the contrary, *Shigella* strains, which show d_G of the same order than other couples of *E. coli* strains, are phenotypically distinct when compared to non-*Shigella E. coli* strains. However, two *Shigella* strains present d_P similar to the ones between two non-*Shigella E. coli* strains, which confirms the phenotypic convergence of these strains. Therefore, our phenotypic assay allows for a representative determination of a strain global phenotype.

Overall, there was only a very weak correlation between d_G and d_P (Mantel test $R^2 = 0.02$, P-value = 0.0076). When the *Shigella* and cryptic *Escherichia* clade strains were removed, the correlation increased (Mantel test $R^2 = 0.10$, P-value < 0.0001), showing that these two opposite cases (low d_G, high d_P and high d_G, low d_P) are typical causes of the disruption of the link between d_G and d_P. Within the non-*Shigella E. coli* strains, the correlation is still significant but very weak (Mantel test $R^2 = 0.01$, P-value = 0.0036). Interestingly, the distribution of d_G for the non-*Shigella E. coli* strains exhibited two peaks corresponding to the intra-phylogroup and inter-phylogroup comparisons, whereas the distribution of d_P was unimodal (Fig. 2). Thus, although *E. fergusonii* and *E. albertii* species as well as *Shigella* strains appeared clearly distinct phenotypically from non-*Shigella E. coli* strains, the different *E. coli* phylogroups rather seemed to display continuous phenotypic variations. The structure of the metabolic phenotypic diversity within *E. coli* species is unknown, and thus, in the following analyses, we focused on non-*Shigella E. coli* strain metabolic phenotypes in relation to the phylogroups and lifestyles of these strains.

Most growth yield variation is independent from the strain phylogeny and lifestyle

Of the 95 carbon sources, 40 showed no growth for any strain, seven allowed growth of all 153 non-S*higella E. coli* strains and 48 were variably used among the strains (Fig. 3, see also Fig. S1 for more details). On average, two strains differently used nine substrates. Thus, the growth capacities within the species were highly variable. The genetic diversity in *E. coli* species is highly structured (Escobar-Páramo *et al.*, 2004a). One hundred and fifty strains were classified into one of the four genetic groups (A/B1, B2, D and E). Each strain was also characterized by three lifestyles: its pathogenic group

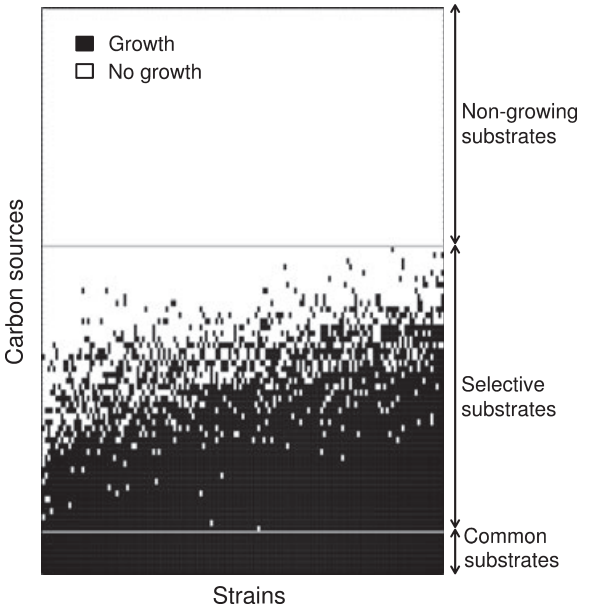

Fig. 3 Diversity of carbon source use by 153 non-*Shigella Escherichia coli* strains. Seven carbon sources allowed the growth of all the 153 strains (common substrates) and 48 of only a fraction of them (selective substrates), whereas 40 did not allow any growth (non-growing substrates). The strains are ordered by the number of substrates they can catabolize. The carbon sources are ordered by the number of strains able to grow on them. See Fig. S1 for a detailed version.

Table 1 Significant grouping effects on the growth yield of 150 non-*Shigella Escherichia coli* strains in the analyses of variance accounting for genetic, pathogenic, host anthropogenic and host diet groups.

	Groups*						
Carbon sources	A/B1 (75)	B2 (38)	D (26)	E (11)	*F*-values†	*P*-values‡	R^2
D-Galactonic acid lactone	0.82 (±0.03)	0.94 (±0.05)	1.01 (±0.06)	0.40 (±0.09)	11.58	8.12×10^{-7}	0.23
D-Serine	0.40 (±0.04)	0.81 (±0.06)	0.40 (±0.07)	0.13 (±0.11)	12.50	2.79×10^{-7}	0.22
Glycyl-L-aspartic acid	0.17 (±0.02)	0.35 (±0.03)	0.20 (±0.03)	0.18 (±0.05)	8.87	2.04×10^{-5}	0.18
Lactulose	0.27 (±0.02)	0.22 (±0.03)	0.13 (±0.03)	0.10 (±0.05)	8.60	2.83×10^{-5}	0.15
p-Hydroxyphenylacetic acid	0.52 (±0.04)	0.09 (±0.05)	0.35 (±0.07)	0.33 (±0.10)	13.15	1.33×10^{-7}	0.30
		Commensal (87)	ExPEC (28)	InPEC (35)			
D,L-Lactic acid		1.10 (±0.02)	0.96 (±0.03)	1.00 (±0.02)	12.26	1.26×10^{-5}	0.18
Uridine		0.51 (±0.02)	0.32 (±0.04)	0.35 (±0.03)	10.87	4.14×10^{-5}	0.20

*Numbers of strains in the groups are indicated in parentheses next to the group label (the three ungrouped strains were discarded). For each carbon source with significant difference between groups, the least-square group mean is given, as well as its corresponding standard error in parentheses.

†The tested *F*-distributions had 3 and 138° of freedom for the genetic group effect and 2 and 138 for the pathogenic group effect.

‡Only the effects for which the FDR was <0.1% were considered significant.

(commensal, ExPEC or InPEC), its host anthropogenic group (according to its exposure to humans: human, pet dog, farm animal or wildlife animal) and its host diet (insectivorous and granivorous bird, carnivorous mammal, herbivorous mammal or omnivorous mammal). To analyse the effect of the genetic group and the lifestyle on the growth yield, we carried out an ANOVA for each carbon source allowing the growth of at least one strain. No significant effect was detected for the host anthropogenic group and the host diet, and only seven of the 55 substrates showed significant grouping effects (Table 1). For instance, the D-serine was differently used among the phylogroups. Members of the B2 group had a higher growth yield on average on this substrate than other strains, which confirmed the results obtained in a study using strains of serotype K1 mainly found in the B2 group (Moritz & Welch, 2006; Bidet *et al.*, 2007). On the contrary, the p-hydroxyphenylacetic acid was almost not used by the strains of the group B2 compared to other strains. Interestingly, the *hca* operon involved in the degradation of this substrate has been found specifically absent in all the group B2 strains (Touchon *et al.*, 2009). However, even in these cases, most of the variance remained unexplained by the model ($R^2 \leq 0.30$). Consequently, on the plots of the PCA based on the growth yield, strains were not grouped by phylogroup or pathogenic group (Fig. 4) or any other lifestyle group (data not shown). Overall, the growth yield diversity did not structure the species into groups, as found previously (Fig. 2), as just a unique cloud of strains emerged from

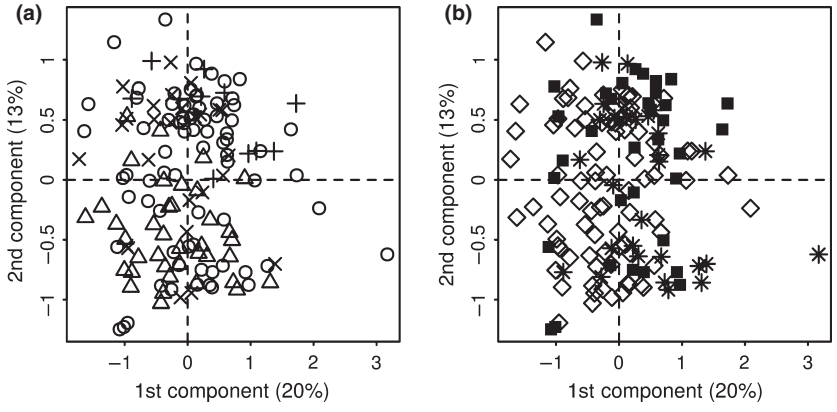

Fig. 4 Principal component analysis (PCA) of 150 non-*Shigella Escherichia coli* strains based on their growth yield on 95 carbon sources. In (A) the symbols correspond to the phylogroups: A/B1 (o), B2 (Δ), D (×) and E (+) (the three ungrouped strains were discarded). In (B) the symbols correspond to the pathogenic groups: commensal (◇), ExPEC (*), InPEC (■). Percentages of total variance explained by the axes are given in parentheses.

the PCA. Thus, the genetic group, as well as the lifestyle group we studied, were very weakly correlated to the growth yield, and, within a given group, phenotypes vary as much as across the whole non-*Shigella E. coli* strains.

Metabolic pathways are distributed according to the species phylogeny

To catabolize a carbon source, a strain must have specific enzymes. Consequently, to check whether the strain growth phenotypes reflect their metabolic gene content, we focused on 14 strains, which had their genome fully sequenced. We recovered the 395 metabolic pathways present in at least one of these 14 strains. About two-thirds (249 pathways) of these 395 pathways were conserved among the strains. A strong correlation (Mantel test $R^2 = 0.56$, P-value < 0.0001) was found between the genetic distance (d_G) and the metabolic pathway completion distance (d_M) (Fig. 5a). On the contrary, the correlation between d_P and d_M was weak (Mantel test $R^2 = 0.11$, P-value = 0.0057, Fig. 5b). Therefore, the presence/absence of metabolic pathways is linked to the genetic distance between strains and thus depends on their genetic group but is only weakly related to their growth yields.

To understand the weak correlation between d_P and d_M, we tried to link the metabolic pathways to the carbon sources they catabolize. Of the 46 carbon sources allowing growth of at least one of the 14 sequenced strains, 43 were successfully linked to one or more pathways. For each strain-by-carbon source combination, we compared the growth status to the presence of the corresponding pathways. Overall, there was quite a good agreement between the presence or absence of metabolic pathways and the growth status as 73% of the cases were coherent, i.e. there was no growth when the pathway was absent or incomplete (53 cases) and growth when the pathway was complete (387 cases). However, we also found inconsistencies in 27 % of the cases, either strains growing while not having the complete required degradation pathway (42 cases) or strains not growing while having the complete degradation pathway (120 cases). Notice that the presence of the degradation pathway does not allow for quantitative predictions. For example, the cases for which the strains grew while not having the complete required degradation pathway did not correspond to particularly low growth yields as they varied in the same range as the growth yields resulting from complete pathways (data not shown).

Few discrepancies between the metabolic pathways and the growth phenotypes are enough to decorrelate phenotypic and metabolic distances

To understand how d_M and d_P can be weakly correlated while the metabolic gene content explain most of the growth capacities, we modelled the phenotypes of a population of strains according to their metabolic gene presence and pathway functionality (all genes must be present for the pathway to be functional). The metabolic pathway completion distance is based on genome sequences and annotations and may not be fully indicative of the phenotypic distance. Indeed, two parameters translate the possible disruption between metabolic pathways and phenotypes. The first, δ_M, is the probability that two strains share a common phenotype on a given substrate (growth or no growth) while having different pathway functionalities concerning this carbon source. The second, δ_P, is the probability that two strains have different phenotypes emerging from the same pathway functionalities. Based on the discrepancies between the observed growth phenotypes and the predicted ones in our data set, we estimated on average $\delta_M = 0.63$ and $\delta_P = 0.21$. Interestingly, the high value for δ_M is mainly because of the cases where the two strains grew while having different pathway functionalities, which means that one strain could grow without the complete corresponding pathway. Using the relationship between the probability that two strains show different phenotypes on a given substrate and the probability that their related pathways have different functionality (eqn 3), we simulated the metabolic pathway distance and the metabolic phenotypic distance between 10 000 strain pairs with a metabolic network composed of 395 pathways (Fig. 6).

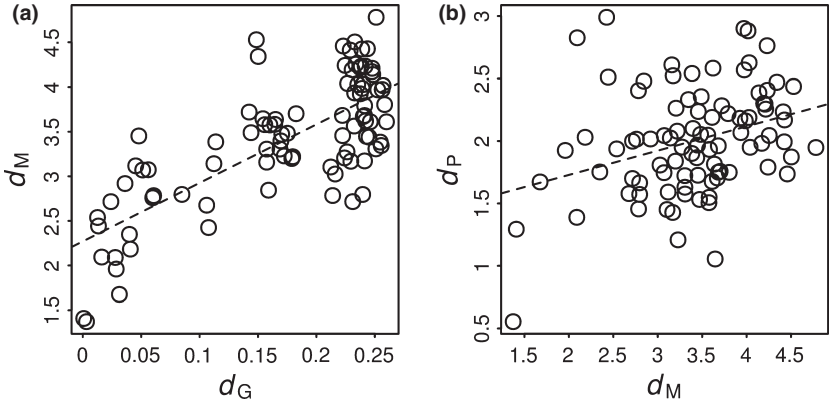

Fig. 5 Relationships between the metabolic pathway completion distance, d_M, and the genetic distance, d_G, (a) and between the metabolic phenotypic distance, d_P, and d_M (b) resulting from comparisons between 14 fully sequenced *E. coli* strains. The dashed lines correspond to the regression of d_M according to d_G (a) and d_P according to d_M (b).

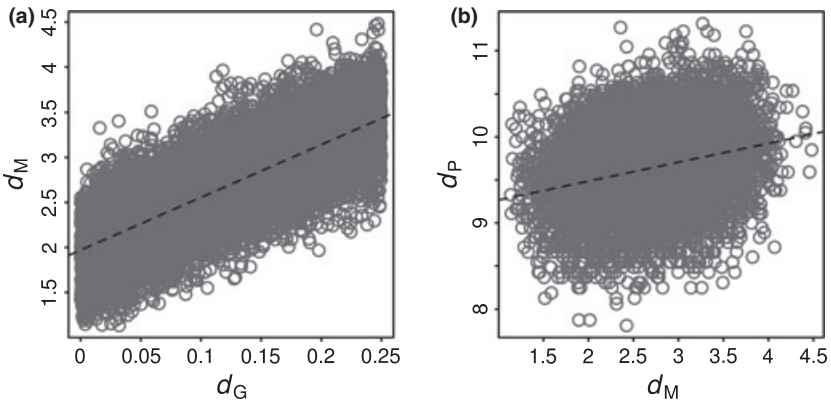

Fig. 6 Relationships between metabolic pathway completion distance, d_M, and genetic distance, d_G, (a) and between metabolic phenotypic distance, d_P, and d_M (b) resulting from the simulation of 10 000 strain pairs with a metabolic network composed of 395 pathways. The dashed lines correspond to the regression of d_M according to d_G (a) and d_P according to d_M (b).

The plot of the metabolic phenotypic distance according to the metabolic pathway distance thus obtained (Fig. 6b) was similar to the one experimentally observed (Fig. 5b). Moreover, the correlation between d_M and d_G (Fig. 6a) was indeed strong ($R^2 = 0.58$), as experimentally observed (Fig. 5a), whereas the one between d_P and d_M was weak ($R^2 = 0.07$). Therefore, the moderate proportion of discrepancies between the presence/absence of metabolic genes and growth phenotypes suffices to blur the link between metabolic phenotypic distance and metabolic pathway distance. The presence/absence of the metabolic pathways explains most of the growth capacities, that is the average strain metabolic phenotype, but is not a good predictor of the phenotypic differences between strains.

Discussion

The global phenotypic structure suggests continuous variations around an average behaviour within the species

We found that on average, in the Biolog GN2 microplate, a strain is able to metabolize 36 carbon sources, seven of which are common to all strains. Beside, several consumed substrates have been shown to be used in the natural habitats, such as L-arabinose, D-galactose, L-fucose, D-gluconic acid, N-acetyl-D-glucosamine, D-glucuronic acid and D-mannose (Chang *et al.*, 2004; Fabich *et al.*, 2008). We also found a great metabolic phenotypic diversity because between two strains nine substrates are differently used on average and globally 48 carbon sources could be used by some strains and not by others. This confirms that it is necessary to study several natural isolates to encompass more aspects of a ubiquitous species such as *E. coli* and that the laboratory model strain K-12 alone is definitely not representative of the whole species (Hobman *et al.*, 2007). The observed diversity is not surprising for microbial species as shown by previous numerical taxonomy studies (Johnson *et al.*, 1975; Sneath *et al.*, 1981). We assessed the effects of the strain phylogroup as well as of different lifestyles (pathogenicity, host exposure to humans and host diet) on the metabolic phenotypes, and we concluded that the metabolic phenotypic diversity of non-*Shigella E. coli* strains is very weakly linked to the strain phylogeny or to their lifestyle. Moreover, the observed variation is unlikely to be explained by other lifestyles as it did not appear to be structured at all. Indeed, the non-*Shigella E. coli* strain growth yield rather seems to present continuous variations around the species average.

The metabolic phenotypes are versatile characters, quickly evolving

In vivo, *E. coli* has a mixed-substrate growth (Harder & Dijkhuizen, 1982; Lendenmann *et al.*, 1996). In environments that contain low concentrations of a variety of substrates, the ability to consume simultaneously several carbon sources even confers a competitive advantage. Indeed, the maximum growth rate of *E. coli* K-12 consuming simultaneously a mixture of two substrates is greater than its maximum growth rate when cultured with either one of the two carbon sources (Narang *et al.*, 1997). In addition, it has been shown that the ability to consume carbon sources impacts on *E. coli* colonization *in vivo* (Chang *et al.*, 2004). For several pathogens, specific metabolic capabilities constitute a fitness advantage or are even necessary for their spread, such as sucrose consumption for *Streptococcus pneumoniae* colonization (Iyer & Camilli, 2007) or lactate uptake for nasopharyngeal colonization by *Neisseria meningitidis* (Exley *et al.*, 2005). Moreover, in a new environment, metabolic capabilities of *E. coli* strains are optimized within a few hundred generations only (Dekel & Alon, 2005). Thus, being able to catabolize and use more than one carbon source is an advantage for both bacterial survival and spread. Therefore, a fraction of the observed metabolic phenotypic diversity might have been selected for and could be the result of the adaptation to slightly different environments. In this respect, the nutrient-niche hypothesis states that several ecological niches correspond to different nutrient availability within the intestine (Freter, 1983). In that case, the growth yield variation would correspond to different

nutritional strategies of the strains adapted to continuous variations in their environment rather than to an environment compartmented into several discrete niches. Accordingly, EDL933 and K-12 have been shown to consume different carbon sources *in vivo* (Fabich *et al.*, 2008). Besides, the ecological niche of a strain is not constant because *E. coli* spends half of its life cycle in its primary habitat (gut of vertebrates) and the other half in the environment (water and soil) (Savageau, 1983). Its geographical spread is rapid and accompanied by frequent ecological niche shifts. For instance, in a farm environment, from the inoculation of a cow, a strain can be recovered from caretakers, mice, pigs, fowls and flies in a few days (Marshall *et al.*, 1990). Other studies showed that *E. coli* can establish and persist for a few days in fish intestines, giving them the opportunity to spread to distant waters (Rio-Rodriguez *et al.*, 1997; Guzmán *et al.*, 2004). Therefore, the continuous variations in metabolic phenotypes can also reflect the adaptation to past niches. Part of the large variability of metabolic phenotypes can also be neutral, having evolved by means of mutations, horizontal gene transfers and genetic drift. Indeed, the high mutational robustness of metabolic networks allows for phenotypic innovations at a low evolutionary cost, as it had been shown from *in silico* analyses (Matias Rodrigues & Wagner, 2009).

Differences in regulatory networks can explain the disruption between genotypes and phenotypes

Less than half of the genome of a strain is shared by all the strains of the species (Rasko *et al.*, 2008; Touchon *et al.*, 2009). Consequently, one expects that part of the observed variation is because of unshared metabolic pathways obtained by horizontal gene transfers or differential gene loss. Accordingly, 73% of the diversity in growth capacities was explained by the presence/absence of degradation pathways. This proportion of explained growth is approximately the same as the level of agreement between experimental and computational results predicted by flux balance analysis calculations of a genome-scale metabolic reconstruction for *E. coli* K-12 (Feist *et al.*, 2007) and falls within the range found in published data on different microorganism species (between 57% and 94%) (Durot *et al.*, 2009). The agreement between growth and metabolic pathway presence in our data is relatively good considering that genome-scale models are more elaborate than our methodology, as they account for the network structure and are often refined with experimental data.

Although *E. coli* core genome represents only 11% of its pan-genome (Touchon *et al.*, 2009), we found that about two-thirds of the pathways present in the species show no difference in completion percentages between the 14 sequenced strains. This observation is in agreement with the fact that *E. coli* core metabolism represents 57% of its pan-metabolism (Vieira *et al.*, 2011). Moreover, 27% of the differences that we observed in metabolic capabilities were not explained by the presence or absence of the corresponding degradation pathways and were used to determine the parameters of the *in silico* simulations of metabolic phenotypes. Growth of strains that do not have the expected metabolic pathways can be caused presumably by unknown or unspecific enzymes that can catalyse several reactions. For instance, no strain had the complete pathway for the degradation of p-hydroxyphenylacetic acid as the enzyme catalysing one of its reactions is not described yet in the database. The cases for which strains had a metabolic pathway but did not grow on the corresponding carbon source can be because of mutations on coding genes or regulatory sequences, which can inactivate a metabolic pathway. Indeed, even if enzyme-coding genes are detected in a genome, missense mutations could still have modified the enzyme activity. Moreover, differences in regulatory networks between strains could affect the expression of the enzyme. For instance, it has been shown that the transcriptome was under selection in the *Shigella* strains (Le Gall *et al.*, 2005). In addition, the evolution of *E. coli* strains in laboratory conditions during a relatively short period revealed that most of the adaptation, i.e. increase in growth rate, was achieved by a transcriptional adjustment (Cooper *et al.*, 2003; Herring *et al.*, 2006). Likewise, the protein expression level has been shown to be rapidly optimized by evolution in *E. coli* (Dekel & Alon, 2005). On the whole, the presence/absence of metabolic pathways is a relatively good predictor of the average growth phenotype although there is an intermediate layer between the metabolic network and the phenotypes. Therefore, the metabolic pathways are distributed according to the phylogroups, but the discrepancies caused by the regulatory layer break this structure and lead to continuous variations in the metabolic phenotypes.

Acknowledgments

We thank Bertrand Picard, Olivier Martin, Meriem El Karoui, Delphine Sicard, Olivier Tenaillon and Thibault Nidelet for many helpful discussions and remarks on the manuscript as well as the Laboratoire de Génomique Comparative and François Le Fèvre for the data on metabolic pathways. We also thank two anonymous reviewers for their constructive comments on our manuscript. This work was partially supported by the Fondation pour la Recherche Médicale, the Délégation Générale pour l'Armement, the Alliance for the Prudent Use of Antibiotics (APUA) in the frame of the Reservoirs of Antibiotic Resistance (ROAR) projects 2006–2007 and the grant ANR-08-SYSC-011 from the Agence Nationale de la Recherche.

References

Baldwin, A., Loughlin, M., Caubilla-Barron, J., Kucerova, E., Manning, G., Dowson, C. *et al.* 2009. Multilocus sequence

typing of *Cronobacter sakazakii* and *Cronobacter malonaticus* reveals stable clonal structures with clinical significance which do not correlate with biotypes. *BMC Microbiol.* **9**: 223.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**: 289–300.

Bernier-Febreau, C., du Merle, L., Turlin, E., Labas, V., Ordonez, J., Gilles, A. *et al.* 2004. Use of deoxyribose by intestinal and extraintestinal pathogenic *Escherichia coli* strains: a metabolic adaptation involved in competitiveness. *Infect. Immun.* **72**: 6151–6156.

Bidet, P., Mahjoub-Messai, F., Blanco, J., Blanco, J., Dehem, M., Aujard, Y. *et al.* 2007. Combined multilocus sequence typing and O serogrouping distinguishes *Escherichia coli* subtypes associated with infant urosepsis and/or meningitis. *J. Infect. Dis.* **196**: 297–303.

Burstin, J. & Charcosset, A. 1997. Relationship between phenotypic and marker distances: theoretical and experimental investigations. *Heredity* **79**: 477–483.

Chang, D., Smalley, D.J., Tucker, D.L., Leatham, M.P., Norris, W.E., Stevenson, S.J. *et al.* 2004. Carbon nutrition of *Escherichia coli* in the mouse intestine. *Proc. Natl Acad. Sci. USA* **101**: 7427–7432.

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

Clermont, O., Olier, M., Hoede, C., Diancourt, L., Brisse, S., Keroudean, M. *et al.* 2011. Animal and human pathogenic *Escherichia coli* strains share common genetic backgrounds. *Infect. Genet. Evol.* **11**: 654–662.

Cooper, T.F., Rozen, D.E. & Lenski, R.E. 2003. Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **100**: 1072–1077.

Dekel, E. & Alon, U. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* **436**: 588–592.

Dray, S. & Dufour, A.B. 2007. The ade4 package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**: 1–20.

Durot, M., Bourguignon, P.Y. & Schachter, V. 2009. Genome-scale models of bacterial metabolism: reconstruction and application. *FEMS Microbiol. Rev.* **33**: 164–190.

Escobar-Páramo, P., Giudicelli, C., Parsot, C. & Denamur, E. 2003. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. *J. Mol. Evol.* **57**: 140–148.

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

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

Escobar-Páramo, P., Le Menac'h, A., Le Gall, T., Amorin, C., Gouriou, S., Picard, B. *et al.* 2006. Identification of forces shaping the commensal *Escherichia coli* genetic structure by comparing animal and human isolates. *Environ. Microbiol.* **8**: 1975–1984.

Exley, R.M., Goodwin, L., Mowe, E., Shaw, J., Smith, H., Read, R.C. *et al.* 2005. *Neisseria meningitidis* lactate permease is required for nasopharyngeal colonization. *Infect. Immun.* **73**: 5762–5766.

Fabich, A.J., Jones, S.A., Chowdhury, F.Z., Cernosek, A., Anderson, A., Smalley, D. *et al.* 2008. Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine. *Infect. Immun.* **76**: 1143–1152.

Feist, A.M., Henry, C.S., Reed, J.L., Krummenacker, M., Joyce, A.R., Karp, P.D. *et al.* 2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol. Syst. Biol.* **3**: 121.

Fontaneto, D., Herniou, E.A., Boschetti, C., Caprioli, M., Melone, G., Ricci, C. *et al.* 2007. Independently evolving species in asexual bdelloid rotifers. *PLoS Biol.* **5**: e87.

Fox, J. & Weisberg, S. 2010. car: Companion to Applied Regression. See http://CRAN.R-project.org/package=car.

Fraley, C. & Raftery, A.E. 2002. Model-based clustering, discriminant analysis and density estimation. *J. Am. Stat. Assoc.* **97**: 611–631.

Fraley, C. & Raftery, A.E. 2006. MCLUST version 3 for R: normal mixture modeling and model-based clustering. See http://www.stat.washington.edu/mclust/.

Freter, R. 1983. Mechanisms that control the microflora in the large intestine. In: *Human Intestinal Microflora in Health and Disease* (D.J. Hentges, ed.), pp. 33–54. Academic Press, Inc., New York, NY.

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

Guindon, S. & Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**: 696–704.

Guinebretière, M., Thompson, F.L., Sorokin, A., Normand, P., Dawyndt, P., Ehling-Schulz, M. *et al.* 2008. Ecological diversification in the *Bacillus cereus* group. *Environ. Microbiol.* **10**: 851–865.

Guzmán, M.C., de los Angeles Bistoni, M., Tamagnini, L.M. & González, R.D. 2004. Recovery of *Escherichia coli* in fresh water fish, *Jenynsia multidentata* and *Bryconamericus iheringi*. *Water Res.* **38**: 2368–2374.

Harder, W. & Dijkhuizen, L. 1982. Strategies of mixed substrate utilization in microorganisms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **297**: 459–480.

Hartl, D.L. & Dykhuizen, D.E. 1984. The population genetics of *Escherichia coli*. *Annu. Rev. Genet.* **18**: 31–68.

Herring, C.D., Raghunathan, A., Honisch, C., Patel, T., Applebee, M.K., Joyce, A.R. *et al.* 2006. Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale. *Nat. Genet.* **38**: 1406–1412.

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

Hobman, J.L., Penn, C.W. & Pallen, M.J. 2007. Laboratory strains of *Escherichia coli*: model citizens or deceitful delinquents growing old disgracefully? *Mol. Microbiol.* **64**: 881–885.

Iyer, R. & Camilli, A. 2007. Sucrose metabolism contributes to *in vivo* fitness of *Streptococcus pneumoniae*. *Mol. Microbiol.* **66**: 1–13.

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

Johnson, R., Colwell, R.R., Sakazaki, R. & Tamura, K. 1975. Numerical taxonomy study of the *Enterobacteriaceae*. *Int. J. Syst. Bacteriol.* **25**: 12–37.

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

Le Gall, T., Darlu, P., Escobar-Páramo, P., Picard, B. & Denamur, E. 2005. Selection-driven transcriptome polymorphism in *Escherichia coli/Shigella* species. *Genome Res.* **15**: 260–268.

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

Lendenmann, U., Snozzi, M. & Egli, T. 1996. Kinetics of the simultaneous utilization of sugar mixtures by *Escherichia coli* in continuous culture. *Appl. Environ. Microbiol.* **62**: 1493–1499.

MacLean, C.R. & Bell, G. 2002. Experimental adaptive radiation in *Pseudomonas*. *Am. Nat.* **160**: 569–581.

MacLean, C.R. & Bell, G. 2003. Divergent evolution during an experimental adaptive radiation. *Proc. Biol. Sci.* **270**: 1645–1650.

MacLean, C.R., Bell, G. & Rainey, P.B. 2004. The evolution of a pleiotropic fitness tradeoff in *Pseudomonas fluorescens*. *Proc. Natl Acad. Sci. USA* **101**: 8072–8077.

Marshall, B., Petrowski, D. & Levy, S.B. 1990. Inter- and intraspecies spread of *Escherichia coli* in a farm environment in the absence of antibiotic usage. *Proc. Natl Acad. Sci. USA* **87**: 6609–6613.

Martinez-Jéhanne, V., du Merle, L., Bernier-Fébreau, C., Usein, C., Gassama-Sow, A., Wane, A. *et al.* 2009. Role of deoxyribose catabolism in colonization of the murine intestine by pathogenic *Escherichia coli* strains. *Infect. Immun.* **77**: 1442–1450.

Matias Rodrigues, J.F. & Wagner, A. 2009. Evolutionary plasticity and innovations in complex metabolic reaction networks. *PLoS Comput. Biol.* **5**: e1000613.

Morandi, S., Brasca, M., Lodi, R., Brusetti, L., Andrighetto, C. & Lombardi, A. 2010. Biochemical profiles, restriction fragment length polymorphism (RFLP), random amplified polymorphic DNA (RAPD) and multilocus variable number tandem repeat analysis (MLVA) for typing *Staphylococcus aureus* isolated from dairy products. *Res. Vet. Sci.* **88**: 427–435.

Moritz, R.L. & Welch, R.A. 2006. The *Escherichia coli argW-dsdCXA* genetic island is highly variable, and *E. coli* K1 strains commonly possess two copies of *dsdCXA*. *J. Clin. Microbiol.* **44**: 4038–4048.

Mukherjee, A., Mammel, M.K., LeClerc, J.E. & Cebula, T.A. 2008. Altered utilization of *N*-acetyl-D-galactosamine by *Escherichia coli* O157:H7 from the 2006 spinach outbreak. *J. Bacteriol.* **190**: 1710–1717.

Narang, A., Konopka, A. & Ramkrishna, D. 1997. New patterns of mixed-substrate utilization during batch growth of *Escherichia coli* K12. *Biotechnol. Bioeng.* **55**: 747–757.

Ochman, H., Lawrence, J.G. & Groisman, E.A. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* **405**: 299–304.

Paradis, E., Claude, J. & Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.

Patel, P.H. & Loeb, L.A. 2000. DNA polymerase active site is highly mutable: evolutionary consequences. *Proc. Natl Acad. Sci. USA* **97**: 5095–5100.

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

Pupo, G.M., Lan, R. & Reeves, P.R. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc. Natl Acad. Sci. USA* **97**: 10567–10572.

R Development Core Team. 2009. R: A Language and Environment for Statistical Computing. Vienna, Austria. See http://www.R-project.org.

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

Rio-Rodriguez, R.E., Inglis, V. & Millar, S.D. 1997. Survival of *Escherichia coli* in the intestine of fish. *Aquac. Res.* **28**: 257–264.

Roldán-Ruiz, I., van Euwijk, F., Gilliland, T.J., Dubreuil, P., Dillmann, C., Lallemand, J. *et al.* 2001. A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (*Lolium perenne* L.) varieties. *Theor. Appl. Genet.* **103**: 1138–1150.

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

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

Sneath, P.H.A., Stevens, M. & Sackin, M.J. 1981. Numerical taxonomy of *Pseudomonas* based on published records of substrate utilization. *Antonie Van Leeuwenhoek* **47**: 423–448.

Spor, A., Nidelet, T., Simon, J., Bourgais, A., de Vienne, D. & Sicard, D. 2009. Niche-driven evolution of metabolic and life-history strategies in natural and domesticated populations of *Saccharomyces cerevisiae*. *BMC Evol. Biol.* **9**: 296.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J. *et al.* 2002. Report of the *ad hoc* committee for the re-evaluation of the species definition in bacteriology. *Int. J. Syst. Evol. Microbiol.* **52**: 1043–1047.

Strimmer, K. 2009. fdrtool: estimation and control of (local) false discovery rates. See http://CRAN.R-project.org/package=fdrtool.

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

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

Venail, P.A., MacLean, R.C., Bouvier, T., Brockhurst, M.A., Hochberg, M.E. & Mouquet, N. 2008. Diversity and productivity peak at intermediate dispersal rate in evolving metacommunities. *Nature* **452**: 210–214.

Vieira, G., Sabarly, V., Bourguignon, P., Durot, M., Le Fèvre, F., Mornico, D. *et al.* 2011. Core and panmetabolism in *Escherichia coli*. *J. Bacteriol.* **193**: 1461–1472.

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

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M. *et al.* 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* **37**: 463–464.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 Diversity of carbon source use by 153 non-*Shigella E. coli* strains.

Figure S2 Representation of the three Gaussian distributions that best fit the growth yield data according to the Bayesian information criterion (BIC) applied to Gaussian mixture models.

Table S1 Characteristics of the *Escherichia* strains used in the study.

Table S2 Metabolic pathways involved in the consumption of 43 carbon sources allowing growth of at least one of the 14 fully sequenced *E. coli* strains.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received 26 February 2011; revised 22 March 2011; accepted 25 March 2011