Evidence of horizontal gene transfer between human and animal commensal *Escherichia coli* strains identified by microarray

Elena Grasselli¹, Patrice François², Michaela Gutacker¹, Brian Gettler², Cinzia Benagli¹, Maruska Convert¹, Patrick Boerlin³, Jacques Schrenzel² & Jean-Claude Piffaretti¹

¹Istituto Cantonale di Microbiologia, via Mirasole, Bellinzona, Switzerland; ²Hôpital Cantonal Genève, Genomic Research Laboratory 24, Geneva, Switzerland; and ³University of Guelph, Ontario Veterinary College, ON, Canada

**Abstract**

Bacteria exchange genetic material by horizontal gene transfer (HGT). To evaluate the impact of HGT on *Escherichia coli* genome plasticity, 19 commensal strains collected from the intestinal floras of humans and animals were analyzed by microarrays. Strains were hybridized against an oligoarray containing 2700 *E. coli* K12 chromosomal genes. A core (genes shared among compared genomes) and a flexible gene pool (genes unique for each genome) have been identified. Analysis of hybridization signals evidenced 1015 divergent genes among the 19 strains and each strain showed a specific genomic variability pattern. Four hundred and fifty-eight genes were characterized by higher rates of interstrain variation and were considered hyperdivergent. These genes are not randomly distributed onto the chromosome but are clustered in precise regions. Hyperdivergent genes belong to the flexible gene pool and show a specific GC content, differing from that of the chromosome, indicating acquisition by HGT. Among these genes, those involved in defense mechanisms and cell motility as well as intracellular trafficking and secretion were far more represented than others. The observed genome plasticity contributes to the maintenance of genetic diversity and may therefore be a source of evolutionary adaptation and survival.

**Introduction**

Bacterial genome sequencing has so far provided a wealth of new information about inter- and intraspecies bacterial genetic diversity. One fundamental concept markedly emerged from genome sequence comparisons: bacterial genome is a dynamic molecule exposed to variation through loss or acquisition of genetic material (Polissi & Soria, 2005). This phenomenon is known as horizontal gene transfer (HGT) and involves the reversible integration of foreign genes. HGT is one of the most important causes of the mosaic nature of prokaryotic genomes and plays a central role in the diversification and speciation of bacteria (Dobrindt et al., 2003), as well as in their capacity to adapt to a changing environment (Lercher & Pal, 2008). It has been documented that bacterial chromosomes, such as those of *Escherichia coli*, are peppered with genes transferred from different sources (Lawrence & Ochman, 2002; Dobrindt et al., 2004). During the last decade, many methods have been devised to distinguish the HGT gene pool from the core gene pool (Ochman et al., 2000; Koonin et al., 2001; Anjum et al., 2003; Fukiya et al., 2004; Lawrence, 2005). One major approach assumes that the core genome of a specific strain has a number of compositional hallmarks, such as a specific GC content or codon usage resulting in a peculiar genome signature (Lawrence & Ochman, 1997; Tsirigos & Rigoutsos, 2005). Another approach assumes that organisms of the same species share a pool of genes that include all elements necessary for survival in their usual environment. By extension, non-essential genes would be controlled by specific acquisition–excision events such as HGT (Maurelli et al., 1998; Welch et al., 2002; Hacker et al., 2004). Thus, the majority of horizontally transferred DNA constitutes the flexible gene pool, as opposed to the core gene pool, which encodes proteins involved in basic cellular functions (Maurelli et al., 1998; Hacker & Carniel, 2001).

*Escherichia coli* represents a model organism for studying prokaryotic genome evolution (Lawrence & Ochman, 1998). This bacterium colonizes a wide diversity of environments and hosts and is particularly adapted to reside in the
intestines or other organs and tissues of various animals (Dobrindt et al., 2003). In addition, a number of pathogenic E. coli variants are likely to cause a variety of diseases, such as diarrhea, urinary tract infections, septicemia and neonatal meningitis (Kaper et al., 2004). With the first sequenced E. coli genome (MG1655 K12), the compositional hallmarks method has shown that 755 out of 4288 genes have been acquired from exogenous sources during at least 234 HGT events (Blattner et al., 1997; Lawrence & Ochman, 1998).

The successive genome comparison between E. coli K12 and the O157:H7 enterohemorrhagic EDL933 strain evidenced the presence of discrete genomic regions containing strain-specific genes (Perna et al., 2001). The common shared genes of these two E. coli genomes were defined as a backbone genome by Perna et al. (2001). An additional genome comparison between the Sakai O157:H7 (another enterohemorrhagic strain) and K12 strain confirmed the presence of strain-specific sequences that represent hot-spots for integration of foreign DNA (Hayashi et al., 2001). Finally, the three-way genome comparison of the uropathogenic CFT073, enterohemorrhagic EDL933 and K12 laboratory strains clearly showed that these strains shared only 39.2% of their genes (Welch et al., 2002). In contrast, the so-called non-backbone genes included genes present in only one or two of the three strains and represented 60.8% of the genome content (Hacker et al., 2004). These genomic differences would thus determine different phenotypic traits, enabling the bacterium to adapt to disparate environmental niches with the consequent affirmation of distinct strains or serovars. In contrast, a flexible gene pool contributes to the ability of bacteria to colonize and/or invade host tissues, as well as to their ability to produce a variety of toxins (Maurelli et al., 1998; Hacker & Carniel, 2001). However, a clear view of the presence or the absence of such genes in commensal E. coli strains hosted by humans or animals is still lacking.

Comparative genome hybridization (CGH) is a microarray technique allowing the genomes of numerous strains to be compared with reference genome(s) (Salama et al., 2000; Bjorkholm et al., 2001; Lucchini et al., 2001; Kim et al., 2002). CGH has been successfully applied to the analysis of specific subsets of human E. coli isolates (Ochman et al., 2000; Anjum et al., 2003; Dobrindt et al., 2003; Fukiya et al., 2004). The present CGH analysis aimed at experimentally determining the flexible and the backbone gene pool in 19 E. coli strains isolated in Switzerland from healthy humans and animals. In parallel, a GC content analysis was performed to investigate a potential relationship between the gene variability and their possible foreign acquisition. Finally, the function and chromosomal position of these different gene pools was investigated and compared with the core and flexible gene pools defined in silico by Welch et al. (2002), and allowed to identify specific traits on E. coli genome variability linked to the host.

### Materials and methods

**Bacterial cultures and DNA extraction**

Nineteen human and animal E. coli strains were isolated from the stools of healthy hosts living in the region of Geneva or in Ticino (Table 1). Bacteria were grown in Mueller–Hinton broth at 37 °C, with shaking at 200 r.p.m. Total DNA was extracted using the DNeasy tissue kit (Qiagen AG, Basel, Switzerland), according to the manufacturer’s instructions. DNA concentration and purity were evaluated spectrophotometrically in a UVikon 810 photometer (A.G. Kontron, Zurich, Switzerland). The DNA was used for the subsequent array hybridizations as well as for a triplex PCR (Clermont et al., 2000; Dobrindt et al., 2003).

**Triplex PCR**

A triplex PCR according to the protocols of Clermont and colleagues (Clermont et al., 2000; Dobrindt et al., 2003) was performed in order to differentiate what we define as ‘phylogenetic commensal’ from ‘potentially pathogenic’ strains. The PCR allows the allocation of the E. coli strains to one of four phylogenetic groups (A, B1, B2 or D) (Clermont et al., 2000; Johnson & Stell, 2000; Dobrindt et al., 2003).

### Table 1. Host and geographic origin of the 19 human and animal commensal Escherichia coli strains analyzed by CGH

<table>
<thead>
<tr>
<th>Strain</th>
<th>E. coli group*</th>
<th>Host origin</th>
<th>Geographic origin¹</th>
</tr>
</thead>
<tbody>
<tr>
<td>S20</td>
<td>B2</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>S21</td>
<td>D</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>H01</td>
<td>D</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>H02</td>
<td>B1</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>H03</td>
<td>B2</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>H04</td>
<td>A</td>
<td>Human</td>
<td>Geneva</td>
</tr>
<tr>
<td>H05</td>
<td>A</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>H12</td>
<td>A</td>
<td>Human</td>
<td>Geneva</td>
</tr>
<tr>
<td>H23</td>
<td>D</td>
<td>Human</td>
<td>Geneva</td>
</tr>
<tr>
<td>H27</td>
<td>A</td>
<td>Human</td>
<td>Ticino</td>
</tr>
<tr>
<td>F01</td>
<td>B2</td>
<td>Feline</td>
<td>Ticino</td>
</tr>
<tr>
<td>F02</td>
<td>B2</td>
<td>Feline</td>
<td>Ticino</td>
</tr>
<tr>
<td>C01</td>
<td>B2</td>
<td>Canine</td>
<td>Ticino</td>
</tr>
<tr>
<td>C02</td>
<td>A</td>
<td>Canine</td>
<td>Ticino</td>
</tr>
<tr>
<td>B02</td>
<td>B1</td>
<td>Bovine</td>
<td>Ticino</td>
</tr>
<tr>
<td>B05</td>
<td>B1</td>
<td>Bovine</td>
<td>Ticino</td>
</tr>
<tr>
<td>S04</td>
<td>A</td>
<td>Swine</td>
<td>Ticino</td>
</tr>
<tr>
<td>S05</td>
<td>B1</td>
<td>Swine</td>
<td>Ticino</td>
</tr>
<tr>
<td>S13</td>
<td>A</td>
<td>Swine</td>
<td>Ticino</td>
</tr>
</tbody>
</table>

*Strains were assigned to four phylogenetic groups based on a triplex PCR. Strains belonging to group B2 and, to a lesser extent, D were defined as potentially pathogenic (Clermont et al., 2000; Dobrindt et al., 2003).

¹Ticino is a canton (county) of Southern Switzerland. The canton of Geneva is situated in Eastern Switzerland and is separated from Ticino by the Alps.
Oligoarray chip design and hybridization

In order to design the oligoarray (Agilent Technologies, Palo Alto, CA), the *E. coli* MG1655 K12 genome sequence (GenBank ID U00096) was scanned by the OLICHECK software (Charbonnier et al., 2005). This procedure generated 2700 oligonucleotide probes (63% of *E. coli* K12 ORFs), which were covalently linked to a glass slide by the SurePrint Technology (Agilent). Purified DNA (100 ng) extracted from the 19 strains was labelled enzymatically with dCTP-cyanine-5 or dCTP-cyanine-3 (Perkin-Elmer, Boston, MA) by the Klenow enzyme (40 U) in the presence of random primers and a mixture containing dNTPs (Sigma, Basel, Switzerland); 1.25 mM for dATP, dGTP and dTTP; and 0.63 mM of dCTP, 30 mM cyanine-labelled dCTP using the Bioprime kit (Invitrogen, Basel, Switzerland).

The reaction mixture was incubated for 2 h at 37 °C, and purified on Centrisep columns (Princeton Separations, Adelphia). Competitive hybridization on the arrays with 750 ng of dCTP-cyanine-3-labelled K12 reference DNA was performed for 16 h at 60 °C in Agilent proprietary buffer. The arrays were then washed for 10 min with a first washing solution [0.5% Triton-X 102; 5X saline sodium citrate (SSC)] at room temperature and for 5 min with a second washing solution (0.5% Triton-X 102; 0.1X SSC) at 4 °C. Slides were dried under nitrogen, and scanned at 100% photomultiplier tube power for both wavelengths using the Agilent microarray scanner G2565BA.

Microarray data analysis

The FEATURE EXTRACTION software (Agilent) was used to extract fluorescence values. In order to establish the match of each ORF to the *E. coli* K12 reference genome, the signals subtracted of the local background were analyzed by Kim's algorithm (Kim et al., 2002). Consequently, the probability of presence estimation allowed the identification of conserved or divergent spots or regions for each strain. Genes that were either absent or sufficiently different to be hindered from hybridization to the related oligonucleotides were defined as divergent. In a multistrain analysis, considering a single ORF at a time, one can determine in how many strains each gene is found to be present (conserved) or divergent. Thus, the divergence of each gene within the strain collection was scored. With the CGH analysis of 19 strains, we generated a binary matrix showing spots as either present (0) or divergent (1) for a given ORF compared with the K12 genomic content. The sum of 0/1 values for a determined gene defines its variability index (VI) (see Supplementary Table S1). The VI indicates the number of strains showing divergence for each specific ORF (Fig. 1).

![Fig. 1.](image-url) Linear representation of the *Escherichia coli* genome showing the VI for each analyzed gene. The genes are sorted by their position on the bacterial chromosome, starting from the origin of replication. Grey bars (zone C) represent the CGH constant gene pool (VI = 0), while black bars represent divergent genes. The line above the horizontal axis separates moderately divergent (VI = 1–4; MD) from hyperdivergent (VI = 5–19; HD) genes.
AVI of 0 indicates that a given ORF is present in all 19 strains. In contrast, a high number indicates a major divergence in gene content from the K12 genome. Divergent genes were divided into two subcategories: moderately divergent (VI = 1–4; MD) and hyperdivergent (VI ≥ 5; HD). An HD gene is a potential candidate for the flexible gene pool.

**GC percentage calculation**

The GC percentage for each K12 gene was obtained from the HGT database site (Garcia-Vallve *et al.*, 2003). Based on this information, we calculated the median and SD of the GC content for all *E. coli* genes. All genes showing a difference lower or higher than a twofold SD from the median were considered HGT candidates (see Supplementary Table S2).

**Results**

**Hyperdivergent genes and hot-spots**

Analysis of the VI evidenced three gene categories: (1) 1685 ORFs were found to be present in all 19 strains (VI = 0), and are thus considered as constant genes, (2) 557 ORFs had a VI of 1–4 and were considered as MD, and finally (3) 458 ORFs failed to be detected in at least five of the tested strains (VI = 5–19) and were thus considered to be HD (see Supplementary Table S1).

To enrich the information, the function of the genes analyzed were defined using the COGs database (*Tatusov et al.*, 1997, 2003). As shown in Table 2, the COGs categories that were most often correlated to HD belong to (1) defense mechanisms, (2) cell motility, as well as (3) intracellular trafficking and secretion categories (‘cellular process’ group). On the other hand, highly conserved genes were mainly associated with the ‘information storage and processing’ group. Conserved genes were also detected in the ‘cellular processes’ and the ‘metabolism’ categories.

Figure 1 suggests that some regions of the *E. coli* chromosome are more likely to undergo rearrangements than others: the figure clearly shows how the HD genes are grouped in precise regions of the bacterial chromosome, the so-called hot-spots (Hacker *et al.*, 1997; Hayashi *et al.*, 2001). Considering the position of HD genes on the bacterial chromosome, we identified 96 clusters composed of 2–13 HD genes. In order to detect a potential relationship between the presence or the divergence of these genes and specific strain characteristics, we further investigated in which strains these genes were present or divergent, respectively. Among the 96 gene clusters, two major clusters of 10 and 13 genes [b2481–b2491 (*hyf* operon) and b1385 (*feab*)–b1400 (*paaY*)], as well as 14 minor clusters of two to four genes (see supplementary Table S1), were found constant in strains belonging to phylogenetic groups A and B1 (i.e. ≥ 82% of the strains) and divergent in most strains of phylogenetic groups B2 and D (i.e. ≥ 62% of the strains, ≥ 80% B2 strains and ≥ 33% D strains). Each strain was assigned to one of these four phylogenetic groups using a triplex PCR as described previously (Table 1; Clermont *et al.*, 2000; Dobrindt *et al.*, 2003). Many of these clustered genes are part of the same operon and the majority belong to the ‘metabolism’ COGs category.

**Table 2. Percentage of the HD, MD and constant genes associated with the clusters of orthologous groups of proteins from the COG database (*Tatusov et al.*, 1997, 2003)**

<table>
<thead>
<tr>
<th>COG Code</th>
<th>Description</th>
<th>HD</th>
<th>MD</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>Translation</td>
<td>3</td>
<td>29</td>
<td>68</td>
</tr>
<tr>
<td>A</td>
<td>RNA processing and modification</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>K</td>
<td>Transcription</td>
<td>35</td>
<td>33</td>
<td>32</td>
</tr>
<tr>
<td>L</td>
<td>Replication, recombination and repair</td>
<td>35</td>
<td>31</td>
<td>34</td>
</tr>
<tr>
<td>D</td>
<td>Cell cycle control, mitosis and meiosis</td>
<td>23</td>
<td>29</td>
<td>48</td>
</tr>
<tr>
<td>V</td>
<td>Defence mechanisms</td>
<td>45</td>
<td>23</td>
<td>32</td>
</tr>
<tr>
<td>T</td>
<td>Signal-transduction mechanisms</td>
<td>19</td>
<td>50</td>
<td>31</td>
</tr>
<tr>
<td>M</td>
<td>Cell wall/membrane biogenesis</td>
<td>28</td>
<td>35</td>
<td>37</td>
</tr>
<tr>
<td>N</td>
<td>Cell motility</td>
<td>61</td>
<td>26</td>
<td>13</td>
</tr>
<tr>
<td>U</td>
<td>Intracellular trafficking and secretion</td>
<td>60</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>O</td>
<td>Posttranslational modification, protein turnover, chaperones</td>
<td>4</td>
<td>43</td>
<td>53</td>
</tr>
<tr>
<td>P</td>
<td>Inorganic ion transport and metabolism</td>
<td>20</td>
<td>38</td>
<td>42</td>
</tr>
<tr>
<td>C</td>
<td>Energy production and conversion</td>
<td>21</td>
<td>36</td>
<td>43</td>
</tr>
<tr>
<td>G</td>
<td>Carbohydrate transport and metabolism</td>
<td>30</td>
<td>35</td>
<td>36</td>
</tr>
<tr>
<td>E</td>
<td>Amino acid transport and metabolism</td>
<td>17</td>
<td>35</td>
<td>48</td>
</tr>
<tr>
<td>F</td>
<td>Nucleotide transport and metabolism</td>
<td>4</td>
<td>44</td>
<td>52</td>
</tr>
<tr>
<td>H</td>
<td>Coenzyme transport and metabolism</td>
<td>9</td>
<td>43</td>
<td>47</td>
</tr>
<tr>
<td>I</td>
<td>Lipid transport and metabolism</td>
<td>32</td>
<td>34</td>
<td>33</td>
</tr>
<tr>
<td>Q</td>
<td>Secondary metabolites biosynthesis, transport and catabolism</td>
<td>33</td>
<td>31</td>
<td>36</td>
</tr>
<tr>
<td>X</td>
<td>Not in COGs</td>
<td>63</td>
<td>22</td>
<td>15</td>
</tr>
</tbody>
</table>

**Genome and species signature**

Genome analysis performed by Welch and colleagues on three sequenced *E. coli* strains indicated that the genomes consist of a conserved backbone gene pool (species signature), and of a variable flexible gene pool (strain-specific genes) (*Welch et al.*, 2002; Hacker *et al.*, 2004).

A comparison between the percentages of Welch’s class of genes (V. Burland, pers. commun.) and the two CGH gene categories that we have defined previously is shown in Fig. 2. As expected, the 458 genes found to be HD by the CGH method are rarely present among those shared by the three sequenced *E. coli* genomes (8.3%), i.e. among the core or
backbone genes. In contrast, most of them overlap with Welch’s flexible gene pool (91.7%). Conversely, the CGH constant gene pool (VI = 0) is mostly represented by genes falling into Welch’s core genes category (87.5%). Consequently, the flexible gene pool, including all remnant Welch’s categories, is present in a low proportion (12.5%).

Figure 3 shows the two main Welch’s classes in relation to the CGH VI. The number of genes falling into Welch’s core gene pool increases with a decrease of the VI and, vice versa, the number of genes that belong to the flexible gene pool increases with the VI.

The negative correspondence existing between CGH HD genes and Welch’s core gene pool is further emphasized by Fig. 4. In addition, a clear correlation between a different GC content and HD genes can be observed: 66.7% of the lower GC content genes belong to the HD gene pool.

Discussion

Difference in genetic backgrounds within strains of the same species may provide essential characters increasing their ability to survive and develop in various environments (Hacker & Kaper, 2000). Most of the genetic determinants responsible for specific characteristics are likely to be acquired by HGT (Hacker & Kaper, 2000; Hacker & Carniel, 2001; Lawrence, 2002), a major source of variability in bacterial genomes (Hacker et al., 2004). The majority of this horizontally transferred DNA is part of the flexible gene pool, the variable component of microbial genomes (Hacker & Carniel, 2001). Important functions are transferred on these mobile moieties, such as transcription factors or genes involved in secondary metabolism (Price et al., 2008). Even if only a fraction of these transferred genes appear to be translated into proteins (Taoka et al., 2004), they confer advantages to the recipient bacteria and contribute to the observation of bacterial niche composed of organisms sharing common features (Price et al., 2008). Detection of such genes or gene groups is one of the issues assessable by comparative genomics (Koonin et al., 2001; Lindsay & Holden, 2004; Lawrence, 2005). Identification of these mobile regions within a genome is possible on the basis of the GC content or codon usage analysis. Another possibility is to identify genes that are not shared by the majority of strains in a particular species (Lawrence & Ochman, 1997; Welch et al., 2002; Hacker et al., 2004; Tsirigos & Rigoutsos, 2005) using genome-wide strategies. Our data showed that CGH is a potent tool for discovering genome-wide differences that develop during the evolution of a microorganism (Nunes et al., 2003; Wolfgang et al., 2003), even using partial coverage of the microorganism studied.

The oligoarray developed in the present study allowed performing CGH between a representative portion of the E. coli K12 genome (2700 ORFs) and the genomes of 19
commensal *E. coli* strains. The CGH analysis of these strains revealed the bipartition of the genome into constant and divergent gene pools. Moreover, CGH allowed a further subdivision of the divergent gene pool into MD and HD genes. The observed chromosomal clustering of HD genes supports their origin by HGT. It is well known that exogenous genetic material preferentially inserts into precise regions of the *E. coli* chromosome that represent hot-spots for recombination. Furthermore, it has been documented that the HGT mostly implies the transfer of blocks of genes, rather than single genes (Dobrindt & Hacker, 2001; Price et al., 2007). Hence, the observed clustering of HD genes in discrete regions of the bacterial chromosome that was observed in the present study could be an indication of HGT events. Further evidence supporting the present separation of the genome into core, MD and HD genes and for the hypothesis that HD genes arose by HGT, stems from their GC content examination. Our observation, that 66.7% of the genes showing a different GC content than the rest of the *E. coli* genome were HD genes, corroborates the hypothesis that HD genes arose by HGT.

Grouping of HD elements in functional categories revealed that the majority of divergent coding sequences are involved in ‘cellular processes’ (Table 2), which is in accordance with the observation that defense mechanisms, cell motility and intracellular trafficking contribute to increase survival and fitness capacity of the bacteria. In contrast, genes found to be conserved by CGH were more frequently associated with the ‘information storage and processing’ COGs group. It is noteworthy that besides these particular examples, constant, MD and HD genes were more evenly distributed among most COG categories (Table 2), suggesting that HGT occurs to an appreciable extent within all *E. coli* functional gene categories.

All analyzed isolates came from the normal intestinal flora of healthy hosts and were, for this reason, described here as commensal strains. However, it is expected that a number of these strains may present a certain pathogenic potential. A PCR system based on the amplification of three genes is able to identify and group such potentially pathogenic strains (Clermont et al., 2000; Dobrindt et al., 2003; Moreno et al., 2006). Groups A and B1 include mainly phylogenetic commensal strains, while groups B2 and D included nearly all strains potentially responsible for extraintestinal infections. In this context, it is noteworthy that two major (10 and 13 genes) and 14 minor HD gene clusters (two to four genes) are present in the majority of commensal isolates (group A, B1) and divergent in most potentially extraintestinal pathogenic strains (group B2 and D). This observation suggests a correlation between the absence or the divergence of some of these genes and extraintestinal strain pathogenicity. The majority of HD genes belong to metabolic COG categories. Interestingly, one of the above-described HD gene clusters (*hyf* operon) encodes a hydrogenase containing an iron–sulfur cluster. The loss of certain metabolic enzymes containing such clusters has been proposed as a survival strategy to circumvent oxidative stress and has been associated with bacterial pathogenicity in the past (Meyer, 2000; Somerville et al., 2002, 2003; Imlay, 2006). *Hyf*-encoded enzymes are induced under fermentative conditions and their expression is oxygen and pH sensitive (Skibinski et al., 2002; Self et al., 2004). Consequently, one may speculate that loss of this operon may confer the ability to colonize extraintestinal tissues and may enhance pathogenicity, as demonstrated in *Shigella* spp. and enteroinvasive *E. coli* strains (Maurelli et al., 1998). Nevertheless, all strains isolated from healthy hosts should still be considered as phylogenetic commensal and their
Pathogenicity potential may depend on the acquisition of transferable virulence factors (Duriez et al., 2001).

One of our working hypotheses was that Welch's backbone genes (genome signature) should be constant in all the strains belonging to the E. coli species. We also expected the flexible gene pool to be more variable within the strain collection. A comparison between Welch's class of genes and our CGH results validated this hypothesis: the CGH constant genes are mostly represented by Welch's core gene pool, while the CGH HD genes principally matched Welch's flexible gene pool.

The gene positions, GC content and function of encoded proteins suggest that HD genes emerged by HGT. We expect that the acquisition of these genes by HGT may be advantageous for the adaptation of the E. coli genome to changing environments. Our results, obtained with a collection of human and animal commensal strains, showed that flexible genes able to contribute to E. coli pathogenicity are present in commensal strains isolated from the intestinal flora of healthy animals. As a consequence, we speculate that commensal strains isolated from healthy animals may represent an important reservoir of genes that might be transferred to other microorganisms by HGT.

Acknowledgements

This work was supported by grant no 40-63270 to J.-C.P. and P.B. from the Swiss National Research Programme NRP49 on antibiotic resistance. We thank Claudio Gilardi for the information support; Valerie Burland, who supplied the gene list of the backbone and flexible genes; and Rainer Merkl, Christian Garzon and Santiago Garcia Vallée for the constructive discussions on data analysis and HD genes.

References


Supplementary material

The following supplementary material is available for this article online:

Table S1. Oligo hybridization and variability index (VI).
Table S2. GC content of K12 genes respective standard deviation from E. coli core genome GC content.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1574-695X.2008.00434.x (This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.