Comparison of 61 Sequenced *Escherichia coli* Genomes

Oksana Lukjancenko · Trudy M. Wassenaar · David W. Ussery

Received: 24 February 2010 / Accepted: 23 June 2010 / Published online: 11 July 2010 © The Author(s) 2010. This article is published with open access at Springerlink.com

**Abstract** *Escherichia coli* is an important component of the biosphere and is an ideal model for studies of processes involved in bacterial genome evolution. Sixty-one publically available *E. coli* and *Shigella* spp. sequenced genomes are compared, using basic methods to produce phylogetic and proteomics trees, and to identify the pan- and core genomes of this set of sequenced strains. A hierarchical clustering of variable genes allowed clear separation of the strains into clusters, including known pathotypes; clinically relevant serotypes can also be resolved in this way. In contrast, when in silico MLST was performed, many of the various strains appear jumbled and less well resolved. The predicted pan-genome comprises 15,741 gene families, and only 993 (6%) of the families are represented in every genome, comprising the core genome. The variable or ‘accessory’ genes thus make up more than 90% of the pan-genome and about 80% of a typical genome; some of these variable genes tend to be co-localized on genomic islands. The diversity within the species *E. coli*, and the overlap in gene content between this and related species, suggests a continuum rather than sharp species borders in this group of Enterobacteriaceae.

**Introduction**

The availability of complete genome sequences from multiple isolates of a given species has opened up a whole new range of research strategies. By far the best-studied bacterial species is *Escherichia coli*, and the highest number of individual genome sequences is available for this species, which has been the working horse of bacteriology for as long as the specialization exists. Numerous basic molecular processes have been first characterized and extensively studied in *E. coli*, leading to insights that could subsequently be applied to other bacteria [47]. Despite the vast amount of knowledge already available for *E. coli*, based on decades of experimental research, genetic manipulation and, more recently, observations based on single or multiple genome sequences, comparison of a large number of *E. coli* genome sequences can still provide novel insights, such as the presence of genomic islands, present in some pathogenicity groups, but missing in others. At the time of writing, there are more than 100 *E. coli* genome sequence projects reported, many of which have been deposited to GenBank. Here, we compare 61 publically available genome sequences of *E. coli* and *Shigella* spp. isolates.

*Escherichia* spp. and *Shigella* spp. are Gram-negative, facultative anaerobic, intestinal bacteria belonging to the Enterobacteriaceae, which are taxonomically placed within the gamma subdivision of the Proteobacteria phylum. Although *Shigella* spp. isolates have been rewarded their own genus, which is divided into several species (representing different sero-groups), its separation from *Escherichia* spp. is mainly historical. For example, in Bergey’s Manual of Systematic Bacteriology, the section on *Shigella* phylogeny begins with the following sentence: “Scientific evidence accumulated to date strongly supports that view that *Shigella* species are
biotypes/pathotypes or clones of E. coli" [39]. More than 50 years ago it was observed that Shigella spp. and E. coli have the same fertility system [25]; in 1972, Brenner et al. [3] found that based on DNA/DNA hybridization, that Shigella spp. and E. coli are the same species. Experiments with multilocus enzyme electrophoresis concluded that nearly all of the Shigella species are clones from within E. coli species [35]. Further, analysis of 16S rRNA sequence alignment places Shigella spp. within E. coli [6]. Thus, all current evidence indicates that Shigella spp. should be classified as E. coli [23, 36]. Both genera contain highly diverse species, although Shigella spp. are as related to E. coli as they are to each other. E. coli is a ubiquitous component of the intestinal gut flora of animals including humans, and can survive and multiply in abiotic environments as well. The species comprises both benign and pathogenic variants, whilst Shigella spp. are all enteropathogens in mammals.

E. coli isolates have in the past been divided into subgroups in various ways. Based on established pathogenicity towards the human host, pathogenic versus commensal E. coli have been recognized, although it is acknowledged that ‘pathogenic’ E. coli strains may colonize other animal species asymptptomatically. Pathogenic E. coli have further been subdivided according to their typical site of infection and clinical manifestations in humans, for instance enteropathogenic, uropathogenic, or extra-intestinal pathogenic E. coli, or based on their virulence mechanisms, such as enterohemorrhagic (EHEC), enterotoxigenic, enteroinvasive, and enteroaggregative E. coli [1, 13]. Other divisions that are frequently used are based on serology (e.g., serotypes O127:H7 or K12) or, mainly for population genetic purposes, on phylogenetic properties of particular housekeeping genes, as established by MULTI-ENZYME electrophoresis and later by multilocus sequence typing (MLST) [25]. Finally, some isolates are described simply for their source of isolation, such as environmental isolates or avian pathogenic E. coli.

All these subdivisions have been applied more or less frequently to group isolates that share particular features. We were interested to see if any of these groupings would hold when isolates were compared based on their complete genome sequences, considering some or all of their genes. Isolates from some groups (based on whatever grouping) have been more frequently sequenced than from others, and complete information on all characteristics of interest (pathogenicity, source of isolation, serotype) is not available for all sequenced isolates. Despite these recognized shortcomings in sampling bias and recorded information, comparison of these 61 genome sequences revealed that neither the 16S gene, nor gene fragments usually used for MLST, provides biologically meaningful information on the relatedness of the sequenced isolates. The best way to analyze this is by taking into account all the genomic content, rather than looking at one or a few individual genes. The E. coli core genome has been previously reported to be less than half the genes [13], with more than half the E. coli genes in any given genome being found in some strains, but missing in others. Many of these variable genes can be clustered to specific regions, located on genomic islands in an E. coli chromosome.

Materials and Methods

Bacterial Genomes and Gene Annotations

Sixty-one bacterial genomes of E. coli and Shigella spp. were used in this study (Table 1). Of these, 39 fully sequenced genomes and 19 genomes for which the sequence was still in progress at the time of extraction were obtained from GenBank (1). Sequence from E. coli O103 Oslo was obtained from Norwegian Veterinary Institute and sequences from strains LANL ECA and LANL ECF were obtained from Los Alamos National Lab. Genome sequences of Escherichia albertii, Escherichia fergusonii, and Salmonella enterica Typhimurium LT2 were included for comparison (Table 1). The ‘quality score’ for each genome is given in Table 1, based on the suggested scale by Chain et al. [4]. A completely sequenced genome that has been deposited to GenBank is given a score of ‘1’, with the only exception being E. coli O157:H7 isolate EDL933, which currently has more than 4,000 “N’s” in the DNA sequence of the GenBank file, representing unfilled gaps along the chromosomeal sequence—hence, this genome is given a lower score of ‘2’. The higher scores represent lower quality (and often more contigs, or pieces of the DNA, although sequence quality is not measured only by this, as described in [4]).

16S Ribosomal RNA Analysis

The sequences encoding 16S ribosomal RNA were extracted from the analyzed genomes using RNAmer [22]; sequences with an RNAmer score above 1,400 were considered reliable and were kept for analysis. From every genome, the gene with highest similarity to rrsH of E. coli K12 MG1655 was selected and these sequences were aligned using ClustalX [24]. A phylogenetic tree was generated by ClustalX using the Bootstrap neighborhood-joining method, showing the bootstrap values at branch points, visualized by NJPlot [34].

In Silico MLST

The alleles for seven housekeeping genes used for MLST of various species (www.mlst.net) were analyzed. These were
| GPID | Strain | Size (bp) | No of genes | Gene density (genes/Kbp) | No of contigs | Accession number | Quality score | Pathotype, serotype, other characteristics | Reference  |
|------|--------|-----------|-------------|--------------------------|---------------|-----------------|--------------|-----------------------------------------|------------|
| 225  | *E. coli* K12 MG1655 | 4,639,675 | 4,149 | 0.894 | 1 | U00096 | 1 | Commensal, K12 | [2] |
| 226  | *E. coli* 0157:H7 Sakai | 5,594,477 | 5,230 | 0.934 | 3 | BA000007 | 1 | EHEC, O157:H7 | [15] |
| 259  | *E. coli* 0157:H7 EDL933 | 5,620,522 | 5,312 | 0.945 | 2 | AE005174 | 1 | UPEC, O6:K2:H1 | [45] |
| 313  | *E. coli* CFT073 | 5,231,428 | 5,339 | 1.020 | 1 | AE014075 | 1 |  |  |
| 13959 | *E. coli* HS | 4,643,538 | 4,378 | 0.942 | 1 | CP000802 | 1 | Commensal, O9 | [37] |
| 13960 | *E. coli* E24377A | 5,249,288 | 4,749 | 0.904 | 7 | CP000800 | 1 | ETEC, O139:H2 | [37] |
| 15572 | *E. coli* B7A | 5,300,242 | 4,648 | 0.876 | 119 | AAJU00000000 | 3 | ExPEC, O6:H31 | [37] |
| 15576 | *E. coli* F11 | 5,215,961 | 4,704 | 0.901 | 2 | AAJT00000000 | 4 | EHEC, O157:H7 | [37] |
| 15577 | *E. coli* E22 | 5,528,238 | 5,105 | 0.923 | 127 | AAJW00000000 | 3 | EPEC, O111:H9 | [37] |
| 15578 | *E. coli* E110019 | 5,376,211 | 4,934 | 0.917 | 127 | AAJV00000000 | 3 | EPEC, O111:H9 | [37] |
| 15639 | *E. coli* S3G38 | 5,371,790 | 4,803 | 0.894 | 1 | AAKB00000000 | 2 |  |  |
| 16193 | *E. coli* 101-1 | 4,979,767 | 4,607 | 0.925 | 91 | AAMK00000000 | 3 | EAEC, O-:H10 | [18, 37] |
| 16235 | *E. coli* S36 | 4,938,920 | 4,620 | 0.935 | 2 | CP000243 | 1 | UPEC | [5] |
| 16259 | *E. coli* UTI89 | 5,179,971 | 5,021 | 0.969 | 2 | CP000247 | 1 | UPEC | [7] |
| 16351 | *E. coli* K12 W3110 | 4,646,332 | 4,226 | 0.909 | 1 | AP090948 | 1 | Commensal, K12 | [16] |
| 16718 | *E. coli* APEC01 | 5,497,653 | 4,428 | 0.801 | 3 | CP000468 | 1 | APEC, O1:K1:H7 | [21] |
| 18083 | *E. coli* ATCC8739 | 4,746,218 | 4,200 | 0.907 | 7 | AP099240 | 1 | Commensal, O152:H28 | [32] |
| 18057 | *E. coli* SE11 | 5,155,626 | 4,679 | 0.907 | 91 | AAPK00000000 | 3 | EAEC, O-:H10 | [18, 37] |
| 18281 | *E. coli* B str. REL606 | 4,629,812 | 4,205 | 0.908 | 2 | CP000819 | 1 |  |  |
| 19053 | *E. coli* SE15 | 4,839,683 | 4,488 | 0.927 | 2 | AP099378 | 1 | Commensal, O150:H5 | [42] |
| 19469 | *E. coli* SMS-3-5 | 5,215,377 | 4,743 | 0.909 | 91 | AP009970 | 1 | Environmental isolate | [11] |
| 20079 | *E. coli* K12 DH10B | 4,686,137 | 4,126 | 0.880 | 1 | AP009948 | 1 | K12 derivative | [8] |
| 20713 | *E. coli* BL21(DE3) | 4,557,508 | 4,157 | 0.912 | 1 | AP01509 | 1 | Commensal, strain B phylogroup | Korea Research Institute of Bioscience and Biotechnology | 

**Table 1** Genomes used in this study

*O. Lukjancenko et al.*
Table 1 (continued)

| GPID   | Strain                        | Size (bp) | No of genes | Gene density (genes/Kbp) | No of contigs | Accession number          | Quality score | Pathotype, serotype, other characteristics | Reference                                |
|--------|-------------------------------|-----------|-------------|--------------------------|---------------|---------------------------|---------------|---------------------------------------------|------------------------------------------|
| 27757  | *E. coli* 0157:H7 EC869       | 5,731,065 | 5,220       | 0.910                    | 147           | ABHU00000000              | 3             | EHEC, O157:H7                               | J. Craig Venter Institute                |
| 28847  | *E. coli* 0157:H7 TW14588     | 5,670,297 | 5,803       | 1.023                    | 10            | ABKY00000000              | 2             | EHEC, O157:H7                               | J. Craig Venter Institute                |
| 28965  | *E. coli* BL21 (DE3)          | 4,557,041 | 4,087       | 0.896                    | 1             | AM946981                  | 1             | Commensal, strain B                         | Australian Center for Biopharmaceutical Technology |
| 30031  | *E. coli* DH1                 | 4,630,707 | 4,160       | 1                         | CP001637      | 1 K12 derivative          | DOE JGI-PGF   |                                             |                                          |
| 30681  | *E. coli* BL21(DE3)           | 4,570,938 | 4,228       | 0.949                    | CP001665      | 1 Commensal, strain B     | DOE JGI-PGF   |                                             |                                          |
| 32571  | *E. coli* 027:H6 E2348/69     | 5,069,678 | 4,554       | 0.898                    | 3             | FM180568                  | 1 EPEC, O127:H6 |                                             | [17]                                    |
| 33413  | *E. coli* 55989               | 5,154,862 | 4,763       | 0.923                    | 1             | CU928145                  | 1 EAEC        |                                             | [43]                                    |
| 33737  | *E. coli* IAI1                | 4,700,560 | 4,353       | 0.926                    | 2             | CU928160                  | 1 Commensal, O8 |                                             | [43]                                    |
| 33775  | *E. coli* S88                 | 5,032,268 | 4,696       | 0.933                    | 2             | CU928161                  | 1 ExPEC, O48:K1:H7 |                                             | [43]                                    |
| 33409  | *E. coli* ED1a                | 5,209,548 | 4,915       | 0.943                    | 1             | CU928162                  | 1 Commensal, O81 |                                             | [43]                                    |
| 33411  | *E. coli* IAI39               | 5,132,068 | 4,732       | 0.922                    | 2             | CU928164                  | 1 UPEC, O7:K1  |                                             | [43]                                    |
| 33415  | *E. coli* UMN026              | 5,324,391 | 4,826       | 0.906                    | 3             | CU928163                  | 1 UPEC, O7:K1  |                                             | [43]                                    |
| 33775  | *E. coli* BW2952              | 4,578,159 | 4,084       | 0.892                    | 1             | CP001396                  | 1 K12 derivative |                                             | [10]                                    |
| 41013  | *E. coli* 0103:H2 12009       | 5,524,860 | 5,121       | 0.926                    | 2             | AP010958                  | 1 EHEC, O103:H2 |                                             | [31]                                    |
| 41021  | *E. coli* 026:H11 11368       | 5,851,458 | 5,516       | 0.942                    | 4             | AP010958                  | 1 EHEC, O26:H11 |                                             | [31]                                    |
| 41023  | *E. coli* 0111:H:- 11128      | 5,766,081 | 5,407       | 0.937                    | 1             | CP000266                  | 1 Dysentery, Serogroup 2a |                                             | [20]                                    |
|        | *E. coli* Oslo                 | 4,960,076 | 4,571       | 0.921                    | 153           | Unpublished               | 4 EHEC, O103: | Norwegian Vet Institute                        |
|        | *E. coli* LANL ECA             | 5,489,845 | 4,750       | 0.865                    | 64            | Unpublished               | 3 EHEC, O157:H7 | Los Alamos National Lab                        |
|        | *E. coli* LANL ECF             | 5,556,393 | 4,813       | 0.866                    | 63            | Unpublished               | 3 EHEC, O157:H7 | Los Alamos National Lab                        |
| 310    | *S. flexneri* 2a 301           | 4,828,821 | 4,177       | 0.865                    | 1             | AE005674                  | 1 Dysentery, Serogroup 2a |                                             | [20]                                    |
| 408    | *S. flexneri* 2a 2457T         | 4,599,354 | 4,061       | 0.882                    | 1             | AE014073                  | 1 Dysentery, Serogroup 2a |                                             | [44]                                    |
| 16375  | *S. flexneri* 5 8401           | 4,574,284 | 4,115       | 0.899                    | 1             | CP000266                  | 1 Dysentery, Serogroup 5 |                                             | [29]                                    |
| 15637  | *S. boydii* CDC 3083-94       | 4,874,659 | 4,246       | 0.871                    | 1             | CP001063                  | 1 Dysentery, Serogroup 1 |                                             | [35]                                    |
| 13146  | *S. boydii* Sb227             | 4,646,520 | 4,134       | 0.889                    | 1             | CP000036                  | 1 Dysentery, Serogroup 4 |                                             | [41]                                    |
| 13145  | *S. dysenteriae* Sd197        | 4,560,911 | 4,271       | 0.936                    | 1             | CP000034                  | 1 Dysentery, Serogroup 1 |                                             | [41]                                    |
| 16194  | *S. dysenteriae* 1012         | 5,235,535 | 4,419       | 0.846                    | 189           | AAMJ00000000              | 3 Dysentery, Serotype 4 | TKGR, [34]                                    |
| 13151  | *S. sonneli* Ss046            | 5,055,316 | 4,219       | 0.834                    | 1             | CP000038                  | 1 Dysentery     |                                             | [47]                                    |
| 28851  | *E. albertii* TW07627          | 4,698,533 | 4,386       | 0.933                    | 64            | ABKY01000000              | 3 Enteropathogenic | J. Craig Venter Institute                        |
| 33369  | *E. fergusonii* ATCC 35469    | 4,643,861 | 4,319       | 0.930                    | 2             | CU928158                  | 1 Commensal    |                                             | [43]                                    |
| 241    | *E. enterica* serovar Typhimurium LT2 | 4,951,371 | 4,525       | 0.913                    | 2             | AE006468                  | 1 Enteropathogenic |                                             | [27]                                    |

**GPID** genome project identifier at NCBI (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi).

EPEC enteropathogenic, UPEC uropathogenic, ExPEC extra-intestinal pathogenic, ETEC enterotoxigenic, EIEC enteroinvasive, EAEC enteroaggregative, APEC avian pathogenic *E. coli*
fragments of adk, fumC, icd, gyrB, mdh, purA, and recA. The obtained DNA sequences were extracted from the genome sequence, concatenated and phylogenetically analyzed as described above. Alignments were not manually adjusted to avoid subjective interpretation of the outcome.

Results and Discussion

A number of characteristics of each of the 61 genomes are summarized in Table 1, such as their size, their number of recognized protein genes, and their gene density. Their GC content varies around 50% for all genomes (not shown), but their size and number of genes varies extensively. The smallest E. coli genome included is that of strain BL21 (DE3) sequenced by the Korean consortium, which is only 4.56 Mbp, and the smallest Shigella genome is that of Shigella dysenteriae Sd197, with 4.56 Mbp. The longest genome of the completed genomes so far belongs to E. coli O157:H7 strain EC4115, with 5.70 Mbp. Longer genomes are listed in Table 1, but since those sequences are still in multiple contigs, it is possible that their stated length is overestimated. These size differences mean that around one million nucleotides (approximately 20% of a genome) can be absent in one E. coli or Shigella isolate and present in another. These ‘extra’ sequences are not void, as indicated by the variation in number of genes: the longest E. coli genome has 1,158 more predicted genes than the shortest E. coli genome (5,315 genes for strain EC4115 and 4157 genes for BL21). Further, the observed gene density is relatively constant, at 0.911±0.04 genes per 1,000 base pairs. It should be noted that published proteomes have been defined using different gene prediction programs and definitions, so that the observed slight variation in gene density might be explained by non-standardized gene identification.

Phylogeny of 16S Ribosomal RNA and MLST Genes

A phylogenetic tree based on the 16S ribosomal RNA sequences extracted from a representative set of 20 Enterobacteriacea genomes is shown in panel a of Fig. 1, which is in agreement with the known phylogeny of the family. The tree for the full set of the 61 E. coli and Shigella strains, including two additional species of Escherichia and one from S. enterica is shown in Fig. 1b. From this figure, it is obvious that phylogeny of the 16S rRNA gene does not resolve well within the genus level, as is known, because the rRNA operons are so similar. Although some of the tree nodes are predicted with uncertainty, clearly the genera Shigella and Escherichia are not separated, nor are E. coli genes separated from those of E. fergusonii or E. albertii. This finding was expected, considering the close relatedness between Escherichia spp. and Shigella spp. In general, 16S sequences are not suitable to analyze inter-strain relationships within a species or between closely related species, as illustrated with this set of Enterobacteriaceae genes. This questions the reliability to use 16S as an indicator for the species to which unknown sequenced DNA belongs [45].

Next, it was investigated if conserved housekeeping genes, frequently assessed for MLST, provide a better representation of the relatedness of the investigated genomes. Various MLST schemes are in use for E. coli
Shigella spp. but these are not standardized and the genes assessed in these schemes are not conserved in all genomes. We used the combination of seven housekeeping genes that has been applied to a number of bacterial species \[26\] (www.mlst.net). Since *S. enterica* lacks *fumC* (an observation that somewhat weakens the general applicability of this MLST gene set), that genome was not included in the analysis. The resulting tree, shown in Fig. 2, still mixes *E. coli* with *Shigella* species, and does not separate all pathogenic strains from commensal strains. Some of the phylogroups previously defined by multilocus enzyme electrophoresis are clearly separated, such as the E

---

**Figure 1** Phylogenetic tree based on extracted 16S rRNA sequences. a Comparison of 20 different Enterobacteriaceae, based on extracted 16S rRNA sequences from the GenBank sequence files. *E. coli* and *Shigella* are shown in green. b Tree of 61 sequenced *E. coli* (black) and related species (colored), based on the alignment of the 16S rRNA gene sequence. Apart from *Shigella* spp., the genes from *E. albertii* and *E. fergusonii* are also included (arrows). The 16S rRNA gene of *S. enterica* Typhimurium LT2 was used as the root. Bootstrap values, indicated in red, show that most nodes are predicted with uncertainty; nevertheless, the genera *Escherichia* spp. and *Shigella* spp. are not separated in this tree, and the three *Escherichia* species are also mixed.
cluster containing all O157 strains, the A/B cluster of commensal K12 and B strains, and the B2 cluster containing some of the uropathogenic strains, in accordance to comparisons carried out by others [40]. Other authors concluded that the O157 serotype of EHEC probably evolved in successive evolutionary events [9]; however, that conclusion is not supported by the MLST tree. And although the B phylogroup is known for its commensal isolates, one of which being used by Delbrück and Luria for their famous phage work, this branch also contains the enteroaggregative strain 101-1 (Fig. 2). Moreover, the two S. dysenteriae strains are widely separated from each other. Pupo et al. [36], who used a different set of MLST genes, also found that isolates of the three species Shigella flexneri, Shigella boydii, and S. dysenteriae, could not always be grouped together nor separated from E. coli.
Various enteroinvasive *E. coli* serotypes have been suggested as ancestral to the different *Shigella* serogroups [23], which could explain the lack of differentiation power of MLST in this case. Apparently, neither MLST gene sets are suitable to group these Enterobacteriaceae organisms in a meaningful way. The performance of MLST could in theory be improved by selecting different genes, for instance using a set of genes specifically chosen to produce the desired grouping. However, the strength of MLST analysis should be that a conserved set of genes is able to identify
phylogenetic relationships in any collection of isolates from one species. If one has to select a 'standard' gene set specifically for the species under investigation, it weakens the general application of MLST considerably.

Pan-Genome Comparisons

MLST analyzes allelic differences in genes whose presence has to be conserved in all genomes. However, we hypothesized that genes that are variably present could provide useful information as to the true relatedness of the analyzed genomes. Since the variable fraction contain genes that are present in some, absent in other genomes, a phylogenetic analysis cannot be performed to capture all information. Figure 3 displays a pan-genome clustering tree, based on the gene families that are variably present in the analyzed genomes (gene families comprising singletons were excluded). The hierarchical clustering obtained by this analysis correctly separates the *Shigella* spp. and *S. Typhimurium* from *Escherichia* spp. and, within the latter genus, separates *E. coli* from the other *Escherichia* spp (Fig. 3). Moreover, all *E. coli* O157:H7 genomes now cluster together, as do the K12 derivatives (W3110, MG1655, DH1, BW2952, DH10B, and ATCC8739). The strains belonging to phylogenic group B are also positioned in one cluster, to which the non-pathogenic commensal strain HS also seems to belong. All these are avirulent isolates, and it is quite impressive that all these are positioned close together in the tree. We conclude that this analysis of variable genes identifies inter-strain relationships that can be correlated to the lifestyle of the organisms.

The contribution of every genome to the complete pan-genome of *E. coli* and related organisms is demonstrated in Fig. 4, where the pan-genome and core genome, as defined by other authors [40] of the analyzed sequences is plotted. The number of novel gene families for every added genome is also shown. As can be seen, all genomes contribute to the increase of the pan-genome. This increase is less strong when similar genomes are added (for instance all four K12 genomes, or the B strains). The

![Figure 3](https://example.com/figure3.png)  
*Figure 3* Pan-genome clustering of *E. coli* (black) and related species (colored), based on the alignment of their variable gene content. The genomes now cluster according to species and a relatedness between *E. coli* K12 derivatives (green block) and group B isolates (orange block) is visible.
DH10B, at position 18 in Fig. 4). The core genome of E. coli is shorter genomes are assessed (starting with added, with an expected significant drop when the noted before [42].

instance, six novel E. coli genomes cause a sharp increase. The core genome reduces in size as more genomes are analyzed (53 E. coli genomes identified 1,472 conserved gene families and 13,296 gene families comprising the pan-genome. We prefer to report these findings as gene families, instead of individual genes, using clearly defined criteria for inclusion of genes into a gene family (described in the “Materials and Methods” section).

Where are all these variable genes located in a genome? Gene order is not strongly conserved between the analyzed genomes, so that gene location depends which genome is considered. Nevertheless, by visualizing where a gene, whose presence can vary, is located on a single reference genome provides further information, and this can be visualized in a BLAST atlas [14]. In the BLAST atlas of Fig. 5, it becomes apparent that the variable gene content is not evenly distributed over the reference genome, but appears to be distributed over various islands. The reference chromosome of E. coli O157:H7 EC4115 was chosen, as it is the largest chromosome for which a

addition of Shigella spp. genomes does not alter the shape of the pan-genome curve, but addition of the other Escherichia genomes causes a sharp increase. The contribution of E. fergusonii to the pan-genome has been noted before [42].

The core genome reduces in size as more genomes are added, with an expected significant drop when the shorter genomes are assessed (starting with E. coli K12 DH10B, at position 18 in Fig. 4). The core genome reaches 1,472 gene families conserved in 53 E. coli genomes, which is further reduced to 993 gene families if Shigella spp. are considered as well. The bars show how many novel gene families each genome contributes to the growing pan-genome. It should be noted that the order in which genomes are analyzed influences the number of these reported novel gene families other than for singletons. When novel genes are considered, instead of novel gene families, the findings can be even more dramatic. For instance, six novel E. coli genome sequences identified approximately 10,000 novel genes [42]. Previous work has estimated a core genome of 1,976 genes for 20 E. coli genomes and a pan-genome of 17,831 genes. Our analysis of 53 E. coli genomes identified 1,472 conserved gene families and 13,296 gene families comprising the pan-genome. We prefer to report these findings as gene families, instead of individual genes, using clearly defined criteria for inclusion of genes into a gene family (described in the “Materials and Methods” section).

Where are all these variable genes located in a genome? Gene order is not strongly conserved between the analyzed genomes, so that gene location depends which genome is considered. Nevertheless, by visualizing where a gene, whose presence can vary, is located on a single reference genome provides further information, and this can be visualized in a BLAST atlas [14]. In the BLAST atlas of Fig. 5, it becomes apparent that the variable gene content is not evenly distributed over the reference genome, but appears to be distributed over various islands. The reference chromosome of E. coli O157:H7 EC4115 was chosen, as it is the largest chromosome for which a

Figure 4: Pan- and core genome plot of the analyzed genomes. The blue pan-genome curve connects the cumulative number of gene families present in the analyzed genomes. The red core genome curve connects the conserved number of gene families. The gray bars show the numbers of novel gene families identified in each genome.

1: Escherichia coli O157:H7 str. EC4196
2: Escherichia coli O157:H7 str. EC4113
3: Escherichia coli O157:H7 str. EC4206
4: Escherichia coli O157:H7 str. EC4076
5: Escherichia coli O157:H7 str. EC4151
6: Escherichia coli O157:H7 str. EC4072
7: Escherichia coli O157:H7 str. EC4422
8: Escherichia coli O157:H7 str. EC4886
9: Escherichia coli O157:H7 str. EC4869
10: Escherichia coli O157:H7 str. EC4096
11: Escherichia coli O157:H7 str. EC4001
12: Escherichia coli O157:H7 str. EDL933
13: Escherichia coli O157:H7 str. TW914688
14: Escherichia coli O157:H7 str. Sakai
15: Escherichia coli O157:H7 EC4045
16: Escherichia coli O157:H7 str. NLANL ECF
17: Escherichia coli K12 str. NLANL ECA
18: Escherichia coli K12 str. DH10B
19: Escherichia coli K12 str. MG1655
20: Escherichia coli K12 str. W3110
21: Escherichia coli K12 str. DH1
22: Escherichia coli RW2592
23: Escherichia coli ATCC3739
24: Escherichia coli B REL606
25: Escherichia coli BL21 (DE3 Korea)
26: Escherichia coli BL21 (DE3 AU)
27: Escherichia coli BL21 (DE3 DOE)
28: Escherichia coli HB1
29: Escherichia coli SE11
30: Escherichia coli IA1
31: Escherichia coli E2476
32: Escherichia coli E2477A
33: Escherichia coli D26 H11 str. 11368
34: Escherichia coli D27 H6 str. E2348/69
35: Escherichia coli D27 H2 str. 12009
36: Escherichia coli D111 H str. 11128
37: Escherichia coli D103 Oxy
38: Escherichia coli DMS 3-4
39: Escherichia coli LMM 026
40: Escherichia coli D3638
41: Escherichia coli IA039
42: Escherichia coli UT89
43: Escherichia coli S88
44: Escherichia coli CFT1073
45: Escherichia coli SE118
46: Escherichia coli S36
47: Escherichia coli ED1a
48: Escherichia coli F13
49: Escherichia coli APDC01
50: Escherichia coli B100019
51: Escherichia coli E52
52: Escherichia coli E7A
53: Escherichia coli T0F
54: Shigella flexneri 2a A57T
55: Shigella flexneri 2a S301
56: Shigella flexneri 5 8401
57: Shigella boydii CDD 3083-84
58: Shigella boydii 5027
59: Shigella sonnei 53048
60: Escherichia coli ATCC ATCC 35469
61: Escherichia albens TW07627
62: Salmonella enterica Typhimurium LT2
63: Shigella dysenteriae 50187
64: Shigella dysenteriae 1012
The complete sequence is currently available. Around this, all other genomes are plotted, whereby lack of color indicates that particular gene from EC4115 is missing in the shown genome. The strong conservation of gene presence within the O157 serotype (in green) contrasts with the multiple 'gaps' seen in the other lanes. Every gap represents multiple genes in strain EC4115, illustrating that gene variation is not evenly distributed along the genome, but located in islands.

**Concluding Remarks**

“This gene is not found in *E. coli*”, is an expression often heard in discussions about novel genes in various organisms, and when people are looking for functional matches in databases. It is a sobering thought to realize that any given *E. coli* genome sequenced will have only roughly 20% of its genes part of the *E. coli* core, and the remaining 80% are not found in all other *E. coli* genomes. After a comparison of the diversity with many sequenced *E. coli* genomes, it has become clear such a statement can only be valid when it is specified which *E. coli* genome sequence has been searched. Of the predicted pan-genome comprising about 16,000 gene families, the core (slightly less than a thousand genes) is found to be only about a fifth of a typical *E. coli* genome which contains around 5,000 genes. Many of the accessible or variable genes, making up more than 90% of the pan-genome and roughly four fifth of a typical genome, are often found co-localized on genomic islands. The diversity within the species *E. coli*, and the overlap in gene content between this and related species is far greater than many had anticipated, and represents a broad set of functions for adapting to many different environments. The comparative methods used here are generally applicable to genomes of related species, and are considered a valuable tool to evaluate current insights of species' relatedness and evolutionary history.

**Acknowledgments** We would like to thank the Danish Research Councils and the DTU Globalization funds for financial support.
Comparison of 61 Sequenced Escherichia coli Genomes

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

1. Anjum MF, Lucchini S, Thompson A, Hinton JCD, Woodward MJ (2003) Comparative genomic indexing reveals the phylogeny of Escherichia coli pathogens. Infect Immun 71:4674–4683
2. Blattner FR, Plunkett G3, Bloch CA, Perna NT, Burland V, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia coli K-12. Science 277:1453–1462
3. Brenner DJ, Fanning GR, Skerman FJ, Falkow S (1972) Polynucleotide sequence divergence among strains of Escherichia coli and closely related organisms. J Bact 109:953–965
4. Chain PS, Grafham DV, Fulton RS, Fitzgerald MG, Hostetler J, Fulton RR, Merriman BR, Gridnev I, Magee J, Doolin K, Mong相关信息被省略，不影响自然语言理解。
26. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci USA 95:3140–3145

27. McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK (2001) Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature 413:852–856

28. Moura RA, Sircili MP, Leomil L, Matté MH, Trabulsi LR, Elias WP, Irino K, Pestana de Castro AF (2009) Clonal relationship among atypical enteropathogenic Escherichia coli strains isolated from different animal species and humans. Appl Environ Microbiol 75:7399–7408

29. Nie H, Yang F, Zhang X, Yang J, Chen L, Wang J, Xiong Z, Peng J, Sun L, Dong J, Xue Y, Xu X, Chen S, Yao Z, Shen Y, Jin Q (2006) Complete genome sequence of Shigella flexneri 5b and comparison with Shigella flexneri 2a. BMC Genomics 7:173

30. Nielsen P, Krogh A (2005) Large-scale prokaryotic gene prediction and comparison to genome annotation. Bioinformatics 21:4322–4329

31. Ogura Y, Ooka T, Iguchi A, Toh H, Asadulghani M, Oshima K, Kodama T, Abe H, Nakayama K, Kurokawa K, Tobe T, Hattori M, Hayashi T (2009) Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. Proc Natl Acad Sci USA 106:17939–17944

32. Oshima K, Toh H, Ogura Y, Sasamoto H, Park S, Iyoda S, Kurokawa K, Morita H, Itoh K, Taylor TD, Hayashi T, Hattori M (2010) Complete genome sequence of the wild-type commensal Escherichia coli strain SE15 belonging to phylogenetic group B2. J Bacteriol 192:1165–1166

33. Touchon M, Hoede C, Tenalilhon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Colteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruff CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médiègue C, Roche EPC, Denamur E (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet 5:e1000344

34. Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G, Mayhew GF, Plunkett G3, Rose DJ, Darling A, Mau B, Perna NT, Payne SM, Runyen-Janecky LJ, Zhou S, Schwartz DC, Blattner FR (2003) Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T. Infect Immun 71:2775–2786

35. Welch RA, Burland V, Plunkett G3, Redford P, Roesch P, Rasko D, Buckles EL, Liou S, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HLT, Donnenberg MS, Blattner FR (2002) Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci USA 99:17020–17024

36. Woo PCY, Lau SKP, Teng JLL, Tse H, Yuen K (2008) Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin Microbiol Infect 14:908–934

37. Yang F, Yang J, Zhang X, Chen L, Jiang Y, Yan Y, Tang X, Wang J, Xiong Z, Dong J, Xue Y, Zhu Y, Xu X, Sun L, Chen S, Nie H, Peng J, Xu J, Wang Y, Yuan Z, Wen Y, Yao Z, Shen Y, Qiang B, Hou Y, Yu J, Jin Q (2005) Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. Nucleic Acids Res 33:6445–6458

38. Zimmer C (2008) Microcosm: E. coli and the new science of life. Pantheon books, New York