Comparison of 61 E. coli genomes
Comparison of 61 Sequenced *Escherichia coli* Genomes

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

**Note:** Include title slide with names of the group, title of the project
**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 publicly available *E. coli* and *Shigella* spp. sequenced genomes are compared, using basic methods to produce phylogenetic 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.
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.

16S rRNA tree

Q: How does E. coli relate?
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*.

**Question:** Why is there poor resolution between *E. coli* genomes?
Question: Does the grouping of the *E. coli* genomes make sense?
Question: What happened when *E. coli* genome #18 was added?
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 E. coli BL21(DE3) E. coli BL21(DE3) E. coli B str. REL606 E. coli 101-1 E. coli K12 str. MG1655 E. coli K12 str. DH1 E. coli K12 str. DH10B E. coli K12 str. W3110 E. coli BW2952 E. coli E24377A Shigella boydii 89Shigella boydii 89 Shigella sonnei Ss046 E. coli E110019 E. coli O26:H11 str. 11368 E. coli O111:H- str. 11128 E. coli SE11 E. coli B7A E. coli O103:H2 str. 12009 E. coli E22 E. coli Oslo O103 E. coli 55989 E. coli IA1 E. coli IA1 E. coli HS E. coli K12 str. ATCC 8739 E. coli UMNO26 E. coli SM5-3-5 E. coli IA19 E. coli SE15 E. coli O127:H6 str. E2348/69 E. coli ED1a E. coli CFT073 E. coli APEC O1 E. coli UT189 E. coli S88 E. coli F11 E. coli 336 Shigella dysenteriae 1012 E. coli BW2952 E. coli K12 str. W3110 E. coli K12 str. DH10B E. coli K12 str. DH1 E. coli K12 str. MG1655 E. coli 101-1 E. coli B str. REL606 E. coli BL21(DE3) E. coli BL21(DE3) E. coli BL21(DE3)

**Question:** Why is there poor resolution between the O157 E. coli genomes?
Comparison of 61 Sequenced *Escherichia coli* Genomes

**Figure 5** BLAST atlas. In the middle, a genome atlas of *E. coli* O157:H7 strain EC4115 is shown, around which BLAST lanes are shown. Every lane corresponds to a genome, with the following colors (going outwards): green *E. coli* O157: H7 (15 lanes); light blue *E. coli* LANL strains (two lanes); dark blue *Shigella* spp. (eight lanes); red *E. coli* K12 and derivatives (six lanes); orange *E. coli* strain B phylogroup (four lanes); followed by all other *E. coli* genomes in different colors. The outermost three lanes represent *E. fergusonii*, *E. albertii*, and *S. enterica* Typhimurium LT2. Lack of color indicates that the genes at that position in strain EC4115 were not found in the genome of that lane. The position of replication origin and terminus is indicated.

**Question:** What is the cluster of highly conserved *E. coli* genes unique to O157?