doi:10.1016/j.jmb.2009.09.052 *J. Mol. Biol.* (2009) **394**, 644–652

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Genome Sequences of *Escherichia coli* B strains REL606 and BL21(DE3)

Haeyoung Jeong[1], Valérie Barbe[2], Choong Hoon Lee[1,3], David Vallenet[2], Dong Su Yu[1], Sang-Haeng Choi[1], Arnaud Couloux[2], Seung-Won Lee[1], Sung Ho Yoon[1], Laurence Cattolico[2], Cheol-Goo Hur[1,4], Hong-Seog Park[1,4], Béatrice Ségurens[2], Sun Chang Kim[3], Tae Kwang Oh[1,5], Richard E. Lenski[6], F. William Studier[7]*, Patrick Daegelen[2,8]* and Jihyun F. Kim[1,4]*

[1]*Korea Research Institute of Bioscience and Biotechnology (KRIBB), 111 Gwahangno, Yuseong, Daejeon 305-806, Korea*

[2]*CNRS UMR 8030, Genoscope (CEA), 2 rue Gaston Crémieux, CP 5706, 91000 Evry Cedex, France*

[3]*Department of Biological Sciences, Korea Advanced Institute of Science and Technology, Daejeon 305-701, Korea*

[4]*Functional Genomics Program, University of Science and Technology, Yuseong, Daejeon 305-333, Korea*

[5]*21C Frontier Microbial Genomics and Applications Center, Yuseong, Daejeon 305-806, Korea*

[6]*Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI 48824, USA*

[7]*Biology Department, Brookhaven National Laboratory, P.O. Box 5000, Upton, NY 11973-5000, USA*

[8]*Inserm, 101 rue de Tolbiac, 75013 Paris, France*

Escherichia coli K-12 and B have been the subjects of classical experiments from which much of our understanding of molecular genetics has emerged. We present here complete genome sequences of two *E. coli* B strains, REL606, used in a long-term evolution experiment, and BL21(DE3), widely used to express recombinant proteins. The two genomes differ in length by 72,304 bp and have 426 single base pair differences, a seemingly large difference for laboratory strains having a common ancestor within the last 67 years. Transpositions by IS*1* and IS*150* have occurred in both lineages. Integration of the DE3 prophage in BL21(DE3) apparently displaced a defective prophage in the λ attachment site of B. As might have been anticipated from the many genetic and biochemical experiments comparing B and K-12 over the years, the B genomes are similar in size and organization to the genome of *E. coli* K-12 MG1655 and have >99% sequence identity over ~92% of their genomes. *E. coli* B and K-12 differ considerably in distribution of IS elements and in location and composition of larger mobile elements. An unexpected difference is the absence of a large cluster of flagella genes in B, due to a 41 kbp IS*1*-mediated deletion. Gene clusters that specify the LPS core, O antigen, and restriction enzymes differ substantially, presumably because of horizontal transfer. Comparative analysis of 32 independently isolated *E. coli* and *Shigella* genomes, both commensals and pathogenic strains, identifies a minimal set of genes in common plus many strain-specific genes that constitute a large *E. coli* pan-genome.

*Corresponding authors. E-mail addresses: jfk@kribb.re.kr; daegelen@genoscope.cns.fr; studier@bnl.gov.

Abbreviations used: SNP, single base pair difference; LPS, lipopolysaccharide.

Received 21 September 2009;
accepted 21 September 2009
Available online
26 September 2009

Edited by M. Gottesman

Keywords: E. coli B; REL606; BL21(DE3); Comparative genomics; Similarity of *E. coli* B and K-12

Introduction

Escherichia coli, first described and isolated by Escherich in 1885,[1] is a ubiquitous inhabitant of the mammalian colon and one of the best studied organisms. *E. coli* strains K-12 and B are apparently both derived from normal commensals of the human gut, and their many derivatives have been in the laboratory since 1922 and before 1918, respectively.[2] Both lines became widely distributed in the 1940s following the use of K-12 derivatives for studies of biochemical genetics by Tatum and Lederberg[3-5] and the choice of the B strain of Delbrück and Luria as a common host for phages T1–T7 by the phage workers who met in the summers at Cold Spring Harbor Laboratory.[6,7] Work with these strains has had tremendous impact on our current understanding of biochemistry, molecular genetics, biotechnology and systems biology as well as on the development of methodologies in these fields.

Whole-genome sequences of K-12 strains MG1655 and W3110, together with powerful computational and molecular tools, have greatly expanded information and insights about the biology of K-12.[8-10] We have now determined the genome sequences of two B strains, REL606, which is being used for a long-term evolution experiment,[11-13] and BL21(DE3), a strain widely used for production of recombinant proteins under the control of T7 RNA polymerase.[14,15] These two B strains last had a common ancestor sometime between 1942 and 1959, and the two lineages have passed through several laboratories and different sets of genetic manipulations to arrive at the strains that were sequenced.[2] The genome sequences of the two B strains reveal many differences between them, and comparison of the genomes of B and K-12 show that they are closely related. We report here an overview of similarities and prominent differences between the genome sequences of B and K-12, and compare their genomes to those of other *E. coli* and *Shigella* strains. A more detailed analysis of the two sequenced B genomes and their relationship to K-12 is given in the accompanying paper.[16]

Results and Discussion

Genomes of REL606 and BL21(DE3)

The genome of REL606 was sequenced by the whole-genome shotgun method,[17] and its sequence was then used as a reference to sequence the genome of BL21(DE3) by a combined approach based on NimbleGen comparative genome sequencing[18] and 454 pyrosequencing,[19] as described in Materials and Methods.

The REL606 and BL21(DE3) genomes are single circular chromosomes of 4,629,812 bp and 4,557,508 bp, respectively, without plasmids present in either strain. The difference in genome length is due to 18 deletions of 274–18,055 bp in BL21(DE3) totaling 81,628 bp; tandem repeats of 113 and 82 bp in REL606; five insertions or deletions (indels) of 5–30 bp; 11 single base pair indels; differences in occupancy of the λ and P2 attachment sites totaling 8689 bp; and an additional IS*1* element in BL21(DE3). The largest deletion in the BL21(DE3) lineage arose spontaneously and removes *ompT*, a desirable result for production of recombinant proteins because OmpT is an outer membrane protease that can degrade proteins during purification.[20] Differences between the two strains are summarized in Table 1 and some of them are illustrated in Fig. 1.

In the 50 or more years since the two B lineages separated, IS*1* has transposed twice in BL21(DE3) and once in REL606, and IS*150* has transposed twice in each strain. The spontaneous mutation selected by Wood in 1966[21] for loss of *EcoB* restriction and modification activity in the BL21(DE3) lineage resulted from transposition of IS*1* into the *hsdS* gene. Transpositions are known to occur readily during storage in sealed agar stabs,[22] and IS*1*, IS*150*, and IS*186* elements all exhibit ongoing transposition in the long-term evolution experiment where REL606 was the founding ancestor.[23-25]

The DE3 prophage was integrated in the λ attachment site of BL21 to provide a source of phage T7 RNA polymerase for producing recombinant proteins.[14] The 42,925 bp prophage sequence we determined is that expected from the λ cloning vector D69[26] plus the DNA fragment inserted to

Table 1. Differences between the genomes of REL606 and BL21(DE3)

	REL606	Combined	BL21(DE3)
Total genome (bp)	4,629,812		4,557,508
SNPs		426	
1 bp indels		11	
274–18,055 bp deletions			18
82, 113 bp duplications	2		
5–30 bp indels		5	
Unique IS*1*	1		2
Unique IS*150*	2		2
Phage in P2 *att*	Defective		0
Phage in λ *att*	Defective		DE3

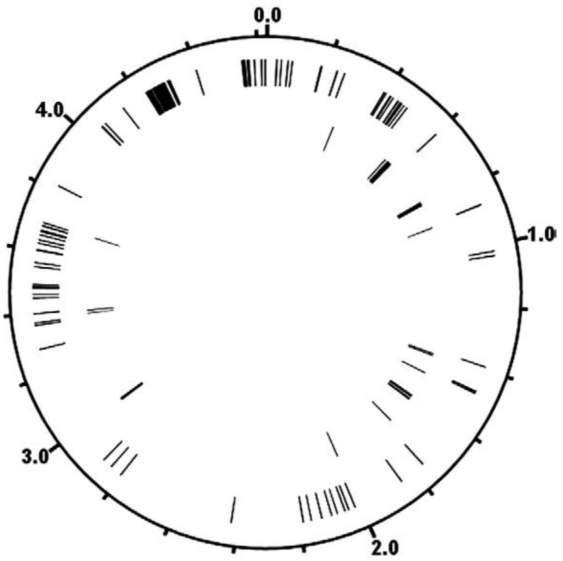

Fig. 1. Distributions of single base pair differences (SNPs) and deletions between BL21(DE3) and REL606. The circle represents the genome of REL606, with position indicated at 0.2×10^6 bp intervals. Positions of the 426 SNPs are marked by the outer set of radial lines and positions of the 18 deletions of 274 bp to 18,055 bp in BL21 (DE3) by the inner set.

make DE3.[14] Unexpectedly, the λ attachment site of REL606 is occupied by a 12,090 bp mobile element that has the *int*, *xis* and *att* insertion module characteristic of λ but little resemblance to the rest of λ. It does, however, have similarities to sequences annotated as phage-related in GenBank and may be the remnant of another prophage. Further analysis[16] found that this element is characteristic of B strains generally and that it was displaced by integration of the DE3 prophage into BL21.

Besides their differences in length, the two B strains have 426 single base pair differences (SNPs) between them (Fig. 1). Strikingly, a region of ~65 kbp at position 4.23–4.30 Mbp in REL606 (~1.4% of the genome) contains 317 of the SNPs (76%) and 9 of the 18 indels of 113 bp or fewer (50%). This number of SNPs seems to be unusually large for two laboratory strains with a common ancestor within the last 67 years, and the concentration of SNPs and indels in a small region was initially perplexing. Detailed analysis reported in the accompanying paper,[16] including extensive comparison with other B strains and with *E. coli* K-12, has led ultimately to a reasonable explanation for every difference between the genomes of the two B strains, almost all of which resulted from known manipulations. In particular, the concentration of differences in the 65 kbp region resulted from disparate integration of K-12 DNA introduced by independent P1-mediated transductions to Mal$^+$ λ^S in the two lineages, unsuspected because the previous literature reported common descent from a single P1 transductant.[2,16]

Comparison of *E. coli* B and K-12

As expected from the cumulative history of comparisons of and genetic recombination between *E. coli* B and K-12 strains, as well as more systematic recent comparisons,[27,28] B and K-12 are similar indeed, from nucleotide sequence to gene content and genome organization. B and K-12 have the same distribution of structural RNAs, except that B is missing *ileY* (encoding the isoleucine tRNA-2 variant) owing to an IS*1*-mediated deletion that also removed eight other genes annotated in K-12 between the end of the cryptic prophage CP4–57 and *csiD*. Large genomic inversions, which are observed sometimes between repetitive sequences (such as *rrn* operons and IS elements) or around the terminus region, were not seen in B relative to K-12 MG1655. The genome sequences could be readily aligned throughout the vast majority of their length: the aligned regions collectively account for >92% of each genome, and the average nucleotide identity is >99% within the aligned regions. Alignment of the REL606 genome with that of K-12 strain MG1655 (leaving out IS elements) is represented in Fig. 2, which illustrates the close similarity of B and K-12. Analysis of the distribution of SNPs in the well-aligned regions of the genomes of B and K-12 is reported in the accompanying paper.[16]

In addition to SNPs, major differences between B and K-12 are in the composition and distribution of large mobile elements (most presumed to be cryptic prophages) and of IS elements. B and K-12 each contain 11 large mobile elements, and 6 of the 15 separate sites occupied by such elements appear to contain a common element diverged to different

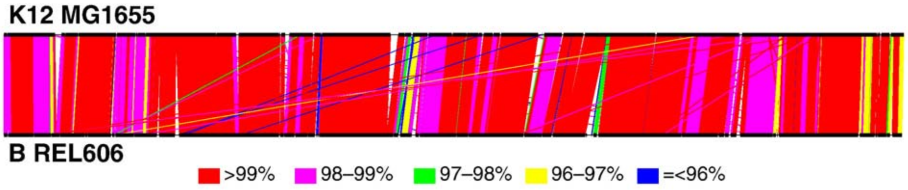

Fig. 2. Whole-genome alignment of K-12 MG1655 and B REL606. Alignments were generated by MUMMER ver. 3.19 using genome sequences screened to eliminate IS elements. Thick horizontal lines indicate aligned segments; rhomboids connecting aligned segments are colored to indicate the percentage identity. Slanted lines connect similar sequences in different parts of the genomes, such as ribosomal RNA genes or other repeated elements.

extents in the two strains.[16] A P2-like prophage inserted between *ybjK* and *ybjL* in B but not K-12 carries the retron Ec86, known to be present in B but not K-12.[29,30] Although the two B strains and K-12 each carry more than 50 copies of IS elements, the number of each type and the sites of insertion are usually different between B and K-12, as noted earlier.[27] The major IS element of B is IS*1*, present in 28 copies in REL606 and 29 in BL21(DE3) but only seven copies in MG1655. The major IS element of MG1655 is IS*5*, present in 11 copies but not present at all in either B strain. Intact or partial copies of at least eight other recognized types of IS elements are present in both B and K-12, usually inserted at different sites.

A frequent effect of IS integration is the inactivation of functional genes, although these elements may also increase expression by providing promoters or altering existing ones. Several phenotypic differences between B and K-12 have previously been associated with IS insertions, including porin expression and Lon protease function.[27,31] The genome sequences also reveal an IS*1*-associated 41 kbp deletion from *uvrY* to *hchA* in the B strains relative to K-12. The deleted segment comprises ~0.9% of the genome and affects 48 annotated genes, including *dcm*, which encodes the DNA cytosine methylase, and 21 *fli* genes encoding flagellar components and the flagella-specific sigma factor (*fliA*, also referred to as *rpoF*). Indeed, B strains are known to lack methylated cytosine in their DNA[32,33] and to be non-motile.[34] B was earlier found to contain a set of genes for group 2 capsule biosynthesis,[35] all of which are lacking in the K-12 genome. These 10 *kps* and 4 *kfi* genes of B would specify a group 2 capsule of the K5 type,[36] but B fails to make a capsule because of an IS*1* insertion in *kfiB*, a gene needed for polymerization of the capsule polysaccharide. Three genes away from the capsule gene cluster, B has a cluster of eight genes for type II secretion that are not found in K-12 (*gspD-K*, located between *yghE* and *yghF*). This gene cluster is in addition to a similar set of 14 type II secretion genes present in both B and K-12 (*gspA-O* between *rpsJ* and *bfr*) and which are known to be cryptic in K-12.[37] The additional set of secretion genes appears to have been lost by a deletion in K-12 that also removed part of the *yghE* coding sequence and the termination codon of *yghF*. The presence of the second set of genes may indicate that B has a secretion capability lacking in K-12. A more comprehensive analysis of similarities and differences between B and K-12 is presented in the accompanying paper.[16]

Restriction-modification system

The classical study of Bertani and Weigle[38] first identified host-controlled restriction and modification by comparing the ability of phages to infect *E. coli* K-12 and B, which have such systems, and other hosts that do not. The restriction–modification complexes are specified by *hsdM, hsdR* and *hsdS*, with specificity determined by HsdS.[39] K-12 and B have different restriction specificities and also have highly diverged *hsdS* genes,[40] although the *hsdM* and *hsdR* genes are quite similar. The region of the genome containing the *hsd* genes and neighboring genes involved in other types of restriction varies considerably among different *E. coli* strains, probably because of horizontal transfer.[41,42] Restriction and modification capacity was eliminated from progenitors of both B strains, by selection of a spontaneous IS*1* insertion in *hsdS* in a progenitor of BL21(DE3),[21] as mentioned above, and by chemical mutagenesis of a progenitor of REL606.[43] REL606 has mutations relative to BL21(DE3) that change one amino acid each in HsdR, HsdM and HsdS, and one or more of these amino-acid changes is presumably responsible for the deficiency.

LPS biosynthesis

The surface of *E. coli* cells is usually formed by lipopolysaccharide (LPS), comprised of lipid A, which forms the outer layer of the lipid bilayer of the outer membrane, linked through a core oligosaccharide to the O-antigen polysaccharide that typically coats the outer surface.[44,45] The LPS core and O antigen, and the gene clusters that encode their synthesis, are highly variable among *E. coli* strains, and horizontal gene transfer contributes to this diversity.[42,46] B and K-12 strains are known to lack O antigen, presumably the result of mutations that occurred in the laboratory after their initial isolation, and the genome sequences reveal the mutations responsible. The gene clusters for synthesis of the LPS core and O antigen of B and K-12 are represented in Fig. 3.

Five different types of LPS core oligosaccharide in *E. coli* have been named K-12 and R1 to R4, and the gene clusters encoding their synthesis have been identified.[47] Gene organization and homology indicates that the LPS core of B belongs to the R1 type, but an IS*1* insertion interrupts *waaT*, which encodes one of the galactosyltransferases involved in synthesis, and thereby truncates the core oligosaccharide from its normal five hexose units to two.[48] As a result, the hexose to which the O-antigen polysaccharide would normally be added is absent, accounting for the lack of O antigen in B. The truncated core may also be responsible for an increased permeability of B relative to K-12.[49,50]

Although the hexose residue to which O antigen would normally attach is lacking in B, a full complement of genes for O-antigen synthesis is present, albeit interrupted by an IS*1* insertion near the end of *wbbD* (Fig. 3). Approximately 170 types of O antigen have been identified in *E. coli* and the O antigen of B is of type O7, as shown by >99% bp identity between the O-antigen gene cluster of B and the sequenced O7 gene cluster of *E. coli* VW187.[51] The IS*1* insertion follows codon 254 of *wbbD* and

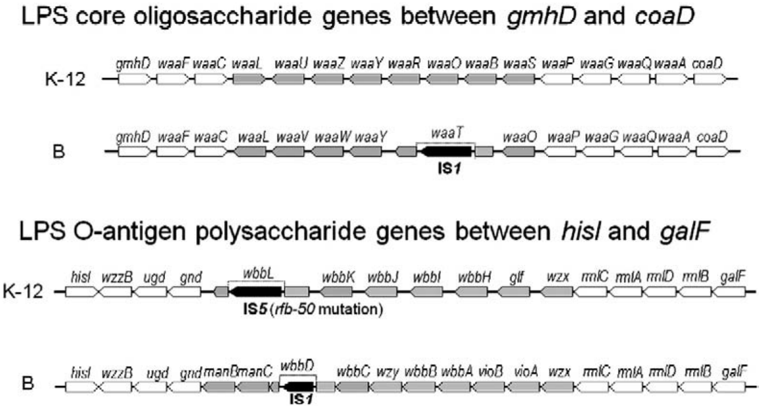
LPS core oligosaccharide genes between gmhD and coaD
K-12
gmhD waaF waaC waaL waaU waaZ waaY waaR waaO waaB waaS waaP waaG waaQ waaA coaD
B
gmhD waaF waaC waaL waaV waaW waaY waaT waaO waaP waaG waaQ waaA coaD
IS1
LPS O-antigen polysaccharide genes between hisI and galF
K-12
hisI wzzB ugd gnd wbbL wbbK wbbJ wbbI wbbH glf wzx rmlC rmlA rmlD rmlB galF
IS5 (rfb-50 mutation)
B
hisI wzzB ugd gnd manB manC wbbD wbbC wzy wbbB wbbA vioB vioA wzx rmlC rmlA rmlD rmlB galF
IS1

Fig. 3. K-12 and B gene clusters for synthesis of LPS core oligosaccharide and O-antigen polysaccharide. The nomenclature is that used by Reeves *et al.*[63] *gmhD* replaces *rfaD*, and *waa* replaces other *rfa* designations for core genes; *wbb* replaces some names for genes of the O7 and O16 O-antigen clusters, and *rmlA-D* replaces *rfbA-D* designations. Arrows indicate direction of transcription, shaded arrows indicate no homology or little similarity between the two genomes, and black arrows indicate IS insertions within the *waaT*, *wbbL* and *wbbD* genes. The gene cluster for the LPS core of B would specify a core of the R1 type, and the gene clusters for the O antigens of B and K would specify the O7 and O16 types, respectively (see the text).

truncates the WbbD protein from 287 amino acids to 254 and an added glycine. WbbD is thought to be a galactosyltransferase catalyzing one of the steps in polymerization of O-antigen polysaccharide, which would not be made if the IS*1* insertion inactivates WbbD. It is probable that the O-antigen polysaccharide can be neither polymerized nor attached to the LPS core in B. Although K-12 strains also lack O antigen, at least two independent mutations are responsible for the deficiency in different laboratory lineages (the more common IS*5* insertion is shown in Fig. 3), and complementation between them can allow the synthesis of O-antigen polysaccharide, which is of type O16.[52] Thus B and K-12, while having closely related basic genomes, have very different clusters of genes for synthesis of both the core oligosaccharide and the O-antigen polysaccharide of the LPS, presumably because of lateral transfer of these gene clusters.

Because the cluster of genes for O7 synthesis in B and the adjacent cluster of genes for colanic acid synthesis both carry genes for phosphomannomutase and mannose-1-phosphate guanyltransferase, B has two pairs of genes with the same function (*manB* and *manC* in the O-antigen cluster and *cpsG* and *cpsB* in the colanic-acid cluster) whereas K-12 has only the pair of genes in the colanic acid cluster. The *manB* and *cpsG* genes of B have 94% bp identity, whereas *manC* and *cpsG* have only 58% amino acid identity, consistent with the switching of O-antigen genes by horizontal transfer. The gene pair in the colanic-acid region of both B and K has 97% bp identity, with 88 bp in the non-coding region between the two genes deleted in K-12 by crossover between 13-bp repeats apparent in B.

Gene content of the *E. coli–Shigella* group

We calculated the minimal core gene set and total gene content of 26 *E. coli* and 6 *Shigella* genomes that represent independent isolates from nature, in much the same way as has been done for other sets of *E. coli* strains.[53,54] *Shigella* strains were included because phylogenetic analyses show they comprise a cluster of related strains embedded within the *E. coli* tree.[55] The *E. coli* strains included six commensals (including B and K-12), seven extra-intestinal pathogenic strains, and 13 intestinal pathogens. The OrthoMCL routine,[56] which uses a Markov clustering algorithm, was applied to compile a list of 9175 putative ortholog clusters of two or more genes and 171 unmatched singletons from 140,162 protein sequences specified by these genomes, including mobile elements and unintegrated plasmids. A total of 1864 orthologous groups, along with 3–36 paralogs per strain, were present in all 32 strains and constitute 36–48% of the gene content of individual strains. The 1867 core genes of MG1655 are listed along with their gene annotations in Table S1 of Supplementary Material.

Using an exponential decay model with five parameters, we estimate the minimal core gene set for the entire *E. coli–Shigella* population to be ~1730 genes. Similar extrapolations estimate considerably larger minimal core gene sets for each of the four subgroups (Fig. 4a). The extrapolated *Shigella*-specific core set of genes is significantly smaller than that of the other three subgroups, which may indicate that the more specialized ecology of *Shigella* promotes accelerated evolution, including gene loss.[57]

A similar extrapolation was used to estimate the total number of genes represented in the entire *E. coli–Shigella* population and for each of the subgroups. In this set of 32 strains, the number of genes added by including successive genomes decreases, on average, from 953 when adding a second strain to 18 when adding a 32nd strain, and the extrapolated pan-genome size is 10,480 genes. The commensal subgroup has the smallest pan-genome (Fig. 4b), presumably reflecting the absence of the wide range of virulence factors found in pathogenic strains. The pan-genome of the intestinal pathogenic subgroup, comprised of enteropathogenic and enterohaemorrhagic strains that employ the type III secretion system, is substantially larger than that of any of the other

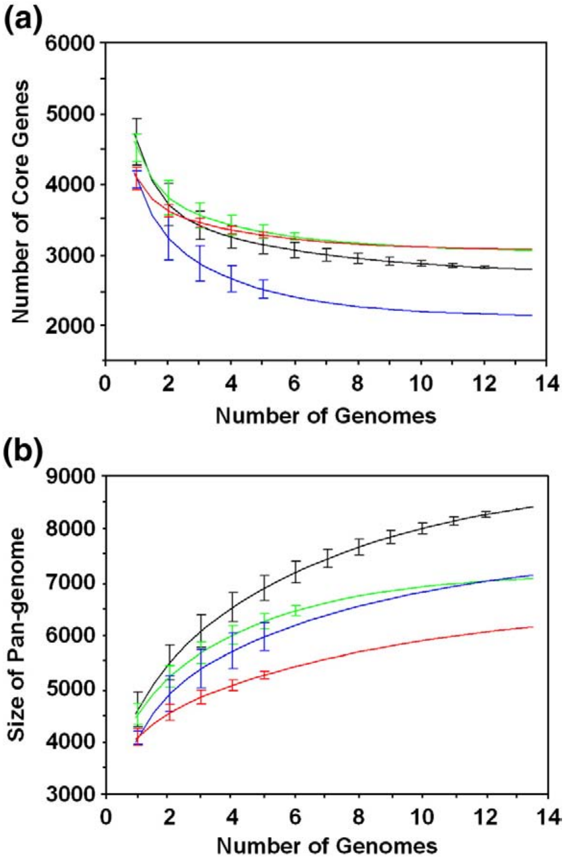

Fig. 4. Analysis of core genes and pan-genomes in four subgroups of *E. coli* strains. The four subgroups comprise six commensal strains (red), 13 intestinal pathogenic strains (black), seven extra-intestinal pathogenic strains (green), and six *Shigella* strains (blue). a, The range of number of genes in common in all possible combinations of a given number of genomes in each subgroup is plotted as a function of number of genomes included. b, The range of total number of different genes is plotted as in a for each subgroup. The curves were fit to the averages of the ranges, as described in Materials and Methods. The last point on each curve is not apparent because it includes every genome of the group and is therefore a single number.

three subgroups, presumably reflecting acquisition of laterally transferred genes that enable more complex and varied life styles.

Materials and Methods

Strains for genome sequencing

The Korean group sequenced REL606 obtained from R.E.L. and the French group sequenced REL606 DNA and culture obtained from Michel Blot and Dominique Schneider, who had obtained the strain from R.E.L. The two sequencing projects were started independently but were combined upon learning of each other. The Korean group sequenced BL21(DE3) obtained from F.W.S.

Genome sequencing and annotation of REL606

The complete genome sequence of *E. coli* B REL606 was determined using a whole-genome shotgun method.[17] Three libraries were constructed; two were made after mechanical shearing of genomic DNA by cloning ~3-kbp and ~10-kbp inserts into plasmids pcDNA2.1 (Invitrogen) (33,151 reads) and the pSU18-derived pCNS[58] (9641 reads), respectively. The third library used DNA fragments of ~20 kbp generated by partial digestion with Sau3AI, which were introduced into the pBeloBac11-derived pBBc[58] (6585 reads). Plasmids were purified, and inserts were end-sequenced by dye-terminator chemistry using ABI 3730 sequencers (Applied Biosystems). The numbers of valid sequences gave ~10-fold coverage on average. The Phred/Phrap/Consed software† was used for genome assembly and quality assessment. About 10,942 additional reactions were necessary to complete the genome (walk plus smoothing, 4862; transposition, 5885; PCR, 195). The integrity of the final assembly was confirmed by comparing the *in silico* restriction map with restriction profiles generated by pulsed-field gel electrophoresis. Putative protein-coding sequences were predicted by AMIGene[59] and ORPHEUS,[60] and submitted for functional annotation by homology search. Removal of overcalled coding sequences and editing of translation start sites by comparison with K-12 homologs were done with Artemis.[61]

Genome sequencing of BL21(DE3)

Draft sequences from the *E. coli* BL21(DE3) genome were collected by NimbleGen CGS[18] with REL606 as the reference sequence and by 454 Life Sciences pyrosequencing (GS 20).[19] Two sets of pyrosequencing contigs with ~10× coverage were produced by *de novo* assembly and by guided assembly using the REL606 genome and the sequence predicted from the D69 vector backbone and DNA insert that constitute the DE3 prophage.[14,26] Discrepancies between the draft and reference sequences were resolved by PCR and conventional sequencing. The resulting 107 high-quality contigs were accommodated in a single scaffold by aligning 2607 fosmid end reads, demonstrating no rearrangement between the BL21(DE3) and REL606 genomes. All remaining gaps were closed by conventional sequencing of PCR-amplified DNA.

Comparative genome analysis

This analysis used amino acid sequences inferred from the REL606 genome and from 31 other sequenced *E. coli* and *Shigella* genomes retrieved from the ColiScope database via the MaGe interface‡. These genomes are from six commensals: REL606; MG1655; ATCC 8739; HS; IAI1; and SE11; seven extra-intestinal pathogens: 536; CFT073; F11; IAI39; S88; UMN026; and UT189; 13 intestinal pathogens: 042; 101-1; 55989; B171; B7A; E110019; E22; E24377A; LF82; O127:H2 E2348/69; O157:H7 EC4115; O157:H7 EDL933; and O157:H7 Sakai; and six *Shigella* strains: *S. flexneri* 2a 2457T; *S. flexneri* 2a 301; *S. flexneri* 5 8401; *S. boydii* Sb227; *S. dysenteriae* Sd197; and *S. sonnei* Ss046. Plasmid-encoded genes were included in the

† http://www.phrap.com/
‡ http://www.genoscope.cns.fr/agc/mage/

analysis. After all-against-all BLASTP searches, putative ortholog clusters (E-value $\le$1E-5, identity $\ge$85%, coverage $\ge$80%) were defined using OrthoMCL.[56] BLAST E-values were then log-transformed to construct a similarity matrix with a maximum value of 300, and a Markov cluster algorithm (MCL) was applied to produce a set of orthologs and recent paralogs using 1.5 as the inflation parameter.[62] The sizes of the minimal core gene sets and the pan genomes were obtained by fitting and extrapolating curvilinear functions:

$$n{=}n_0{+}a \times e^{-bx}{+}c \times e^{-dx}$$

and

$$n{=}n_0{+}a \times \left(1{-}e^{-bx}\right){+}c \times \left(1{-}e^{-dx}\right)$$

respectively, where n is the number of genes, x is the number of genomes, and n_0, a, b, c, and d are variable parameters.

GenBank accession numbers

The genome sequences have been deposited in GenBank: REL606 (GenBank accession no. CP000819), BL21(DE3) (GenBank accession no. CP001509), and the DE3 prophage (GenBank accession no. EU078592). They are also available from the website for the Genome Encyclopedia of Microbes§.

Acknowledgements

We thank Jae-Pil Choi, Hyun-Jin Kim, Soo-Yeon Kim, Benoît Vacherie, and other members of GEM and the KRIBB/CNS sequencing team for technical assistance; the late Michel Blot and Dominique Schneider for samples of REL606 DNA and culture; and Sang Yup Lee, Seung-Hwan Park, and Jean Weissenbach for valuable comments. This work was supported by the 21C Frontier Microbial Genomics and Applications Center Program of the Korean Ministry of Education, Science and Technology, and the KRIBB Research Initiative Program (to J.F.K.); Consortium National de Recherche en Génomique (to P.D.); the GTL Program of the Office of Biological and Environmental Sciences of the U.S. Department of Energy and internal research funding from Brookhaven National Laboratory (to F.W.S.); and the U.S. National Science Foundation and DARPA 'Fun Bio' program (to R.E.L.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.09.052

§ http://www.gem.re.kr

References

1. Escherich, T. (1885). Die Darmbakterien des Neugeborenen und Säuglinge. *Fortschritte der Medizin*, **3**, 515–522.
2. Daegelen, P., Studier, F. W., Lenski, R. E., Cure, S. & Kim, J. F. (2009). Tracing ancestors and relatives of *Escherichia coli* B, and the derivation of B strains REL606 and BL21(DE3). *J. Mol. Biol.* **394**, 634–643.
3. Gray, C. H. & Tatum, E. L. (1944). X-Ray induced growth factor requirements in bacteria. *Proc. Natl Acad. Sci. USA*, **30**, 404–410.
4. Tatum, E. L. (1945). X-ray induced mutant strains of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **31**, 215–219.
5. Tatum, E. L. & Lederberg, J. (1947). Gene recombination in the bacterium *Escherichia coli*. *J. Bacteriol.* **53**, 673–684.
6. Delbrück, M. & Luria, S. E. (1942). Interference between bacterial viruses. I. Interference between two bacterial viruses acting upon the same host, and the mechanism of virus growth. *Arch. Biochem.* **1**, 111–141.
7. Demerec, M. & Fano, U. (1945). Bacteriophage-resistant mutants in *Escherichia coli*. *Genetics*, **30**, 119–136.
8. Blattner, F. R., Plunkett, G., 3rd, Bloch, C. A., Perna, N. T., Burland, V., Riley, M. *et al.* (1997). The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.
9. Hayashi, K., Morooka, N., Yamamoto, Y., Fujita, K., Isono, K., Choi, S. *et al.* (2006). Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. *Mol. Syst. Biol.* **2**, 1–5.
10. Riley, M., Abe, T., Arnaud, M. B., Berlyn, M. K., Blattner, F. R., Chaudhuri, R. R. *et al.* (2006). *Escherichia coli* K-12: a cooperatively developed annotation snapshot–2005. *Nucleic Acids Res.* **34**, 1–9.
11. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. (1991). Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2000 generations. *Am. Nat.* **138**, 1315–1341.
12. Blount, Z. D., Borland, C. Z. & Lenski, R. E. (2008). Historical contingency and the evolution of a key innovation in an experimental population of *Escherichia coli*. *Proc. Natl Acad. Sci. USA*, **105**, 7899–7906.
13. Barrick, J. E., Yu, D. S., Yoon, S. H., Jeong, H., Oh, T. K., Schneider, D. *et al.* (2009). Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature*, doi:10.1038/nature08480.
14. Studier, F. W. & Moffatt, B. A. (1986). Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes. *J. Mol. Biol.* **189**, 113–130.
15. Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol*, **185**, 60–89.
16. Studier, F. W., Daegelen, P., Lenski, R. E., Maslov, S. & Kim, J. F. (2009). Understanding the differences between genome sequences of *Escherichia coli* B strains REL606 and BL21(DE3), and comparison of the *E. coli* B and K-12 genomes. *J. Mol. Biol.* **394**, 653–680.
17. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R. *et al.* (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
18. Albert, T. J., Dailidiene, D., Dailide, G., Norton, J. E., Kalia, A., Richmond, T. A. *et al.* (2005). Mutation

discovery in bacterial genomes: metronidazole resistance in *Helicobacter pylori*. *Nature Methods*, **2**, 951–953.
19. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A. *et al.* (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
20. Grodberg, J. & Dunn, J. J. (1988). *ompT* encodes the *Escherichia coli* outer membrane protease that cleaves T7 RNA polymerase during purification. *J. Bacteriol.* **170**, 1245–1253.
21. Wood, W. B. (1966). Host specificity of DNA produced by *Escherichia coli*: bacterial mutations affecting the restriction and modification of DNA. *J. Mol. Biol.* **16**, 118–133.
22. Naas, T., Blot, M., Fitch, W. M. & Arber, W. (1994). Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics*, **136**, 721–730.
23. Papadopoulos, D., Schneider, D., Meier-Eiss, J., Arber, W., Lenski, R. E. & Blot, M. (1999). Genomic evolution during a 10,000-generation experiment with bacteria. *Proc. Natl Acad. Sci. USA*, **96**, 3807–3812.
24. Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. (2000). Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics*, **156**, 477–488.
25. Rozen, D. E., Schneider, D. & Lenski, R. E. (2005). Long-term experimental evolution in *Escherichia coli*. XIII. Phylogenetic history of a balanced polymorphism. *J. Mol. Evol.* **61**, 171–180.
26. Mizusawa, S. & Ward, D. F. (1982). A bacteriophage lambda vector for cloning with *Bam*HI and *Sau*3A. *Gene*, **20**, 317–322.
27. Schneider, D., Duperchy, E., Depeyrot, J., Coursange, E., Lenski, R. & Blot, M. (2002). Genomic comparisons among *Escherichia coli* strains B, K-12, and O157:H7 using IS elements as molecular markers. *BMC Microbiol.* **2**, 18.
28. Elena, S. F., Whittam, T. S., Winkworth, C. L., Riley, M. A. & Lenski, R. E. (2005). Genomic divergence of *Escherichia coli* strains: evidence for horizontal transfer and variation in mutation rates. *Int. Microbiol.* **8**, 271–278.
29. Lim, D. & Maas, W. K. (1989). Reverse transcriptase-dependent synthesis of a covalently linked, branched DNA-RNA compound in E. coli B. *Cell*, **56**, 891–904.
30. Dodd, I. B. & Egan, J. B. (1996). The *Escherichia coli* retrons Ec67 and Ec86 replace DNA between the *cos* site and a transcription terminator of a 186-related prophage. *Virology*, **219**, 115–124.
31. SaiSree, L., Reddy, M. & Gowrishankar, J. (2001). IS*186* insertion at a hot spot in the *lon* promoter as a basis for Lon protease deficiency of *Escherichia coli* B: identification of a consensus target sequence for IS*186* transposition. *J. Bacteriol.* **183**, 6943–6946.
32. Doskocil, J. & Sormova, Z. (1965). The sequences of 5-methylcytosine in the DNA of *Escherichia coli*. *Biochem. Biophys. Res. Commun.* **20**, 334–339.
33. Fujimoto, D., Srinivasan, P. R. & Borek, E. (1965). On the nature of the deoxyribonucleic acid methylases. Biological evidence for the multiple nature of the enzymes. *Biochemistry*, **4**, 2849–2855.
34. Hershey, A. D., Kalmanson, G. & Bronfenbrenner, J. (1943). Quantitative methods in the study of phage-antiphage reaction. *J. Immunol.* **46**, 267–279.
35. Andreishcheva, E. N. & Vann, W. F. (2006). *Escherichia coli* BL21(DE3) chromosome contains a group II capsular gene cluster. *Gene*, **384**, 113–119.
36. Whitfield, C. & Roberts, I. S. (1999). Structure, assembly and regulation of expression of capsules in *Escherichia coli*. *Mol. Microbiol.* **31**, 1307–1319.
37. Francetic, O., Belin, D., Badaut, C. & Pugsley, A. P. (2000). Expression of the endogenous type II secretion pathway in *Escherichia coli* leads to chitinase secretion. *EMBO J*, **19**, 6697–6703.
38. Bertani, G. & Weigle, J. J. (1953). Host controlled variation in bacterial viruses. *J. Bacteriol.* **65**, 113–121.
39. Murray, N. E. (2000). Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle). *Microbiol. Mol. Biol. Rev.* **64**, 412–434.
40. Gough, J. A. & Murray, N. E. (1983). Sequence diversity among related genes for recognition of specific targets in DNA molecules. *J. Mol. Biol.* **166**, 1–19.
41. Sibley, M. H. & Raleigh, E. A. (2004). Cassette-like variation of restriction enzyme genes in *Escherichia coli* C and relatives. *Nucleic Acids Res.* **32**, 522–534.
42. Milkman, R., Jaeger, E. & McBride, R. D. (2003). Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination. *Genetics*, **163**, 475–483.
43. Lederberg, S. (1966). Genetics of host-controlled restriction and modification of deoxyribonucleic acid in *Escherichia coli*. *J. Bacteriol.* **91**, 1029–1036.
44. Reeves, P. (1995). Role of O-antigen variation in the immune response. *Trends Microbiol.* **3**, 381–386.
45. Raetz, C. R. & Whitfield, C. (2002). Lipopolysaccharide endotoxins. *Annu. Rev. Biochem.* **71**, 635–700.
46. Reeves, P. (1993). Evolution of *Salmonella* O antigen variation by interspecific gene transfer on a large scale. *Trends Genet.* **9**, 17–22.
47. Heinrichs, D. E., Yethon, J. A. & Whitfield, C. (1998). Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*. *Mol. Microbiol.* **30**, 221–232.
48. Jansson, P. E., Lindberg, A. A., Lindberg, B. & Wollin, R. (1981). Structural studies on the hexose region of the core in lipopolysaccharides from Enterobacteriaceae. *Eur. J. Biochem.* **115**, 571–577.
49. Herrera, G., Urios, A., Aleixandre, V. & Blanco, M. (1993). Mutability by polycyclic hydrocarbons is improved in derivatives of *Escherichia coli* WP2 uvrA with increased permeability. *Mutat. Res.* **301**, 1–5.
50. Herrera, G., Martinez, A., Blanco, M. & O'Connor, J. E. (2002). Assessment of *Escherichia coli* B with enhanced permeability to fluorochromes for flow cytometric assays of bacterial cell function. *Cytometry*, **49**, 62–69.
51. Marolda, C. L., Feldman, M. F. & Valvano, M. A. (1999). Genetic organization of the O7-specific lipopolysaccharide biosynthesis cluster of *Escherichia coli* VW187 (O7:K1). *Microbiology*, **145**, 2485–2495.
52. Liu, D. & Reeves, P. R. (1994). *Escherichia coli* K12 regains its O antigen. *Microbiology*, **140**, 49–57.
53. Rasko, D. A., Rosovitz, M. J., Myers, G. S., 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.
54. 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.* **e1000344**, 5.
55. Yang, J., Nie, H., Chen, L., Zhang, X., Yang, F., Xu, X. *et al.* (2007). Revisiting the molecular evolutionary history of *Shigella* spp. *J. Mol. Evol.* **64**, 71–79.

56. Li, L., Stoeckert, C. J., Jr & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.
57. Hershberg, R., Tang, H. & Petrov, D. A. (2007). Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biol.* **8**, R164.
58. Sirand-Pugnet, P., Lartigue, C., Marenda, M., Jacob, D., Barre, A., Barbe, V. *et al.* (2007). Being pathogenic, plastic, and sexual while living with a nearly minimal bacterial genome. *PLos Genet.* **3**, e75.
59. Bocs, S., Cruveiller, S., Vallenet, D., Nuel, G. & Medigue, C. (2003). AMIGene: annotation of microbial genes. *Nucleic Acids Res.* **31**, 3723–3726.
60. Frishman, D., Mironov, A., Mewes, H. W. & Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Res.* **26**, 2941–2947.
61. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
62. van Dongen, S. (2000). *Graph Clustering by Flow Simulation,* PhD thesis, University of Utrecht.
63. Reeves, P. R., Hobbs, M., Valvano, M. A., Skurnik, M., Whitfield, C., Coplin, D. *et al.* (1996). Bacterial polysaccharide synthesis and gene nomenclature. *Trends Microbiol.* **4**, 495–503.

Note added in proof: J.F.K., P.D., R.E.L., and F.W.S. are co-leaders of the International E. coli B Genome Consortium; J.F.K. organized and coordinated the sequencing projects of REL606 and BL21(DE3), and P.D. did so for REL606; H.-S.P., B.S., L.C., and S.-H.C. conducted the shotgun sequencing and pre-assembly steps of REL606; V.B. organized the final steps of sequencing, finishing, assembly, and validating the assembly of REL606; A.C. managed the accumulating sequence data and database for REL606; D.V. contributed to the automatic annotation of REL606, and integrated the final assembly into MaGe; H.J. conducted the mass sequencing and finishing of BL21(DE3), annotated the genomes, and analyzed the sequence data; D.-S.Y. helped with annotation and performed phylogenomic analyses; C.H.L. analyzed the core and pan-genomes; C.-G.H. and S.-W.L. constructed and maintained databases and the web interface; J.F.K., P.D., R.E.L., F.W.S., T.K.O. and S.C.K. conceived analyses and interpreted the biological significance of the results; P.D., D.-S.Y., and S.H.Y. contributed to the manuscript preparation; and H.J., J.F.K., C.H.L., R.E.L., and F.W.S. wrote the paper.