Genome evolution in enteric bacteria

Howard Ochman and Ulfar Bergthorsson

University of Rochester, Rochester, USA

For more than a decade, the study of bacterial evolution has been dominated by the comparative analysis of nucleotide sequences within and among species. This approach, combined with the characterization of extensive regions of the chromosome by pulsed-field gel electrophoresis, has led to new insights into the dynamics of bacterial genomes.

Current Opinion in Genetics & Development 1995, 5:734–738

Introduction

Variation in bacterial genomes can be mediated by any number of mechanisms, but the resulting changes fall into three general classes: point mutations, which include nucleotide substitutions and frameshift mutations; homologous exchange; and chromosomal rearrangements, involving the acquisition, deletion and reorganization of segments of the genome. The first two types of alteration entail changes to the existing genetic information, and the majority of evolutionary studies have examined variation at this level of genetic organization through comparisons of homologous regions within and among species.

The most widespread application of this comparative approach has been to resolve phylogenetic relationships among bacterial species. Although these analyses do not yield much information about the evolution of bacterial genomes *per se*, they provide the framework necessary for investigating other biological traits. For example, the base compositions of bacterial genomes vary widely across taxa, with guanine plus cytosine (G+C) contents ranging from 25% to 75%; and when examined in a phylogenetic context, very broad groups of related species have similar G+C contents, suggesting that genomic base composition is principally the result of mutational biases rather than a response to environmental factors [1,2•,3•].

The use of sequence information to reconstruct the phylogenetic relationships among bacteria relies on the assumption that a particular molecule has not been subject to genetic exchange — horizontal processes tend to homogenize bacterial species and obscure the true genealogy of a lineage. The relative role of point mutations versus recombination in the generation of allelic variation has been a major focus of research on the population genetics of bacteria, and studies reported over the past year based on comparative sequence analysis and low-resolution physical maps have improved our understanding of the factors generating diversity in bacterial genes and genomes.

Recombination in natural populations

On the basis of electrophoretically detectable variation at polymorphic enzyme loci, natural isolates of *Escherichia coli* and *Salmonella enterica* have been regarded as essentially clonal and, for the most part, unaffected by recombination. Despite very high levels of genic diversity, strains of *E. coli* having particular combinations of alleles over loci (multilocus genotypes) were repeatedly recovered from unrelated mammalian hosts from widely separated geographic locations [4]; these clonal lineages appeared to be fairly stable over time [5], indicating that the rate of recombination in natural populations is very low. Analyses of recent nucleotide sequence data advocate a slightly different, though not inconsistent, view of evolution within natural populations of enteric bacteria, emphasizing the role of homologous exchange in generating variation among strains [6,7••,8••].

Recombination events occurring over an evolutionary timescale can be resolved phylogenetically. Because gene transfer will alter the ancestry of a particular region of the genome, incongruities in the branching orders or relationships of organisms based on different genes (or segments of genes) provide evidence of recombination, an approach most recently applied to enteric bacteria by Guttman and Dykhuizen [9••] and by Nelson and Selander [10•]. Genealogies based on each of four loci situated within a 100 kb segment of the *E. coli* chromosome exhibited several inconsistencies attributable to recombination events, and allelic divergence due to recombination was an order of magnitude greater than that caused by point mutations [9••]. Nelson and Selander [10•] determined the nucleotide sequence of 6-phosphogluconate dehydrogenase (*gnd*) — a highly

Abbreviation

PFGE—pulsed-field gel electrophoresis.

variable locus when assayed by protein electrophoresis—for 87 strains typed to five genera of enteric bacteria. On the basis of *gnd* sequences, certain isolates of *E. coli* emerged as being most closely related to *Citrobacter* or *Klebsiella*, a phylogeny not supported by other housekeeping genes or any alternate schemes of classification. Recombination at the *gnd* locus is presumably the result of its close proximity to genes that mediate antigenic variation, which recombine in response to strong diversifying selection [10•,11•].

Patterns of evolution across the chromosome

As additional genes are sequenced from these taxa, it has become possible to examine the rates and patterns of evolution across genes, and even along the entire bacterial chromosome. In comparisons of homologous genes from *E. coli* K-12 and *S. enterica* serovar Typhimurium LT2, rates of synonymous substitutions—which should be roughly the same in every gene, as these silent changes are presumably under no selective constraints—spanned nearly two orders of magnitude. By plotting the synonymous substitution rates against an index specifying the degree of bias in codon usage for each gene, Sharp and Li [12] found that highly expressed genes—those employing a very restricted set of synonymous codons—displayed the lowest rates of synonymous site evolution.

Not all of the variation in synonymous substitution rates can be explained by codon usage patterns, however: Sharp *et al.* [13] subsequently discovered that synonymous substitution rates increase with distance from the replication origin, an effect attributed to the reduced incidence of recombinational repair near the terminus, as these sequences remain in single copy for a greater portion of the cell cycle. A similar explanation has been invoked to account for the slight decrease in G+C content in genes near the replication terminus, whereby these late replicating sequences are repaired by a mechanism that leads to the preferential incorporation of adenine rather than by a recombinational process which requires a homologous strand [14].

Chromosomes as mosaics

So far, we have only considered the effects of recombination on allelic diversity; however, gene transfer also introduces novel regions to bacterial chromosomes. Various methods are used to detect segments of the chromosome acquired through gene transfer: restricted phylogenetic distribution, atypical base composition and codon usage pattern, or an association with translocatable sequences; a particular region will often manifest several of these properties [6,15]. For example, the gene encoding a non-specific acid phosphatase (*phoN*) in *Salmonella enterica* is confined to very few enteric genera, has a G+C content of 43% (which is much lower than that of the *Salmonella* chromosome, which averages 52%) and resides downstream of a sequence with high levels of similarity to the *oriT* region of incFII plasmids, all of which suggest lateral transfer from a low G+C organism in a plasmid-mediated event [16].

Using these criteria to infer the ancestry of a gene, it is possible to analyse the regions sequenced from a particular species and to estimate the proportion of the chromosome that originated through horizontal transfer. By partitioning the genes of *E. coli* into three classes on the basis of their patterns of codon usage, Médigue *et al.* [17] hypothesized that one class, comprising 16% of the sequenced genes, arose through horizontal transfer. Similarly, approximately 10% of the sequenced genes from Typhimurium LT2 have features which depart from the prevalent characteristics of the genome [3•]. These values may underestimate the actual number of acquired genes because sequences are apt to be transferred from closely related organisms of similar base compositions and codon usage patterns.

On the basis of the amount of gene transfer exposed through the analyses of nucleotide sequences, the chromosomes of enteric bacteria can perhaps be considered mosaics, with portions introduced from diverse sources. However, the collinearity of the genetic maps, as well as the correspondence between the sizes and organization of the *E. coli* and *S. enterica* chromosomes, have led to the view that the bacterial genomes are evolutionarily conserved. But despite the overall similarities, when the linkage maps of *E. coli* K12 and *S. enterica* serovar Typhimurium LT2 are aligned, there are several genetic rearrangements, including an inversion encompassing 10% of the chromosome, and some 30 regions over 25 kb in length—termed 'loops'—that are present in only one of the species [18,19]. Genes that confer several of the traits used to distinguish *E. coli* and *S. enterica*, such as the utilization of lactose and citrate, reside on these loops [19]; recently, Mills *et al.* [20•] defined a 40 kb region in Typhimurium LT2 that is not present in *E. coli* K12 and contains at least 20 genes necessary for the the invasion of mammalian epithelial cells by salmonellae.

Physical structure and mapping

With the advent of pulsed-field gel electrophoresis (PFGE), it is now possible to directly assess the physical size, structure and organization of bacterial chromosomes, even in strains not amenable to genetic manipulation. The impact of PFGE on bacterial genetics and genome analysis has been the subject of several recent reviews [21–23], so we have only considered studies that pertain to the evolution of enteric bacterial genomes.

PFGE was first applied to enteric bacteria by Smith *et al.* [24], who established a low-resolution physical map of *Not*I restriction sites in *E. coli* K-12 strain EMG2. Since then, several other laboratory derivatives of *E. coli* K-12 have been mapped, revealing numerous changes in the chromosome structure, some of which had not been detected by genetic methods [25]. Most of these involve small changes, and a disproportionate number occurred near the replication terminus, suggesting an enhanced high rate of recombination in this region [25,26•]. Although PFGE is useful for detecting variation among laboratory strains of *E. coli*, its application to natural and clinical isolates has principally been limited to epidemiological tracing of pathogenic strains [27–30].

PFGE has most recently been employed to examine genome structure and evolution within a species. In this regard, the most comprehensive studies have examined serovars of *S. enterica* by comparing low-resolution physical maps of the serovars Enteritidis [31], Paratyphi [32], Typhi [33••], and Typhimurium [31,34,35•,36]. Initially, these studies substantiated the overall similarities in the size and organization the *E. coli* and *S. enterica* chromosomes: estimates of chromosome length in *E. coli* K-12 are 4.6 Mb compared with 4.7–4.9 Mb for serovars of *S. enterica* [34]. With regard to the conservation of gene order within *S. enterica*, some serovars contain large-scale changes relative to Typhimurium LT2 — for example, serovar Enteritidis harbors an inversion containing the terminus and involving 18% of the chromosome. This inversion is congruent with one detected from comparisons of *E. coli* K-12 and Typhimurium LT2, but it spans a larger region [31]. The endpoints of the Enteritidis inversion are situated in a non-divisible zone, as ascertained in *E. coli* [37], where the replication fork pauses after moving through the terminus. (Non-divisible zones are regions flanking the terminus that are refractory to genetic inversions.) In experimental populations, inversion endpoints have never been mapped to these non-divisible zones suggesting that such inversions have occurred by mechanisms other than those operating in the experimental populations [31].

One rare-cutting restriction enzyme, I-*Ceu*I, has been used extensively for the characterization of *Salmonella* chromosomes [31]. This enzyme has a 26 bp recognition sequence occurring in rRNA operons of bacterial and organelle genomes. This homing endonuclease has seven sites in the *E. coli* and *Salmonella* chromosomes, corresponding to number of *rrn* genes. Although the number and distribution of I-*Ceu*I restriction sites are well conserved among the chromosomes of different serovars of *S. enterica*, differences in fragment lengths have resulted from the insertions or deletions of regions in independent lineages.

There are some notable exceptions to the conserved distribution of rRNA operons in *S. enterica*, particularly in Typhi and Paratyphi A and C [31,34]. The chromosome of Typhi Ty2 has undergone substantial reorganization, attributable to recombinational events among rRNA operons. Closely-related clones of Typhi have a surprisingly high degree of variation in *rrn* restriction patterns [38,39], which may correspond to the chromosome rearrangements reported by Liu *et al.* [34]. This contrasts the situation in Typhimurium, where 15 of 17 wild-type strains displayed identical I-*Ceu*I restriction patterns and fragment lengths [35•].

Employing PFGE, we have examined the variation in genome size among some dozen strains of *E. coli* from natural sources [40•]. Restriction fragment patterns for two rare-cutting enzymes, *Bln*I and *Not*I, were highly variable among isolates, and estimates of genome size ranged from roughly 4650 kb to 5300 kb, which is several hundred kb larger than the variation detected between the *E. coli* K-12 and *S. enterica* serovar Typhimurium LT2. Differences in genome size increase with the evolutionary genetic distance (as assessed by multilocus enzyme electrophoresis), that is, more closely related strains trend to have similar genome sizes. Although bacterial genomes are commonly thought to be subject to streamlining to assure rapid rates of replication, there was no correlation among these natural isolates of *E. coli* between overall genome size and growth rates in either minimal or nutrient media [40•].

Conclusions

Despite the similarities of the *E. coli* K-12 and *S. enterica* serovar Typhimurium LT2 chromosomes, and the fact that these species are basically clonal, evidence generated over the past year through nucleotide sequencing and pulsed-field gel electrophoresis has considerably altered our view about the rates and patterns of evolution in enteric bacteria. Recombination within and among species has influenced the extent of genic diversity observed at several loci within *E. coli* and *S. enterica*, and regions introduced through horizontal transfer has contributed to large-scale alterations in the structure and organization of bacterial chromosomes. Further research will be directed towards resolving the absolute rate of these processes and their effects on the long-term evolution of enteric species.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Sueoka N: **Directional mutation pressure and neutral molecular evolution.** *Proc Natl Acad Sci USA* 1988, **85**:2653–2657.

2. • Aoyama K, Haase AM, Reeves P: **Evidence for effect of random genetic drift on G+C content after lateral transfer**

of fucose pathway genes to *Escherichia coli*. *Mol Biol Evol* 1994, **11**:829–838.
See annotation [3•].

3. • Ochman H, Lawrence J: **Phylogenetics and the amelioration of bacterial genome.** In *Escherichia coli and Salmonella typhimurium*, vol 2. Edited by Neidhardt FC. Washington DC: ASM Press; 1995: in press.
This paper and [2•] provide cases where homologous genes have been acquired independently by closely related lineages. This allows the authors to examine how genes with atypical G+C contents will change after residence in a genome of a different overall base composition. By considering additional features of bacterial genes, this can be taken a step further to determine the extent of the genome that arose through horizontal transfer.

4. Selander RK, Levin B: **Genic diversity and structure in *Escherichia coli* populations.** *Science* 1980, **210**:545–547.

5. Ochman H, Selander RK: **Standard reference collection of *Escherichia coli* from natural populations.** *J Bacteriol* 1984, **157**:517–24.

6. Whittam TS, Ake SE: **Genetic polymorphisms and recombination in natural populations of *Escherichia coli*.** In *Mechanisms of Molecular Evolution*. Edited by Takahata N, Clark AG. Tokyo: Japan Scientific Societies Press; 1993:223–245.

7. •• Selander RK, Li J, Boyd EF, Wang F-S, Nelson K: **DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*.** In *Bacterial Sytematics*. Edited by Priest FG, A Ramos-Cormenzana A, Tindall R. New York: Plenum Press; 1994:1 –36.
This is arguably the best review of recent nuclelotide sequence data pertaining to the role of mutation and recombination in the evolution within enteric bacteria. Using phylogenetic trees, the authors synthesize information, principally from their laboratory, gathered over the past five years and then show the relationship between sequence diversity and protein function.

8. •• Milkman R: **Recombinational exchange among clonal populations.** In *Escherichia coli and Salmonella typhimurium*, vol 2. Edited by Neidhardt FC. Washington DC: ASM Press; 1995: in press.
An excellent introduction to the theory of clonal frames, examining the extent of exchange among natural isolates of *E. coli*. In some further studies, the author has performed a series of transductions to determine the size and diversity of fragments that can be introduced into a specific genetic background.

9. •• Guttman DS, Dykhuizen DE: **Clonal divergence in *Escherichia coli* as a result of recombination, not mutation.** *Science* 1994, **266**:1380–1383.
This study emphasizes the role of recombination in generating sequence diversity among closely-related natural isolates of *E. coli*, and discusses how recombination can either diversify or homogenize bacterial strains depending on the relationships among isolates.

10. • Nelson K, Selander RK: **Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria.** *Proc Natl Acad Sci USA* 1994, **91**:10227–10231.
See annotation [11•].

11. • Thampapillai G, Lan L, Reeves P: **Molecular evolution in the *gnd* locus of *Salmonella enterica*.** *Mol Biol Evol* 1994, **11**:813–828.
This citation and [10•] both examine the variation at one of the most highly polymorphic loci, *gnd*, in enteric bacteria. The variation is generated by the close linkage of this housekeeping gene to the *rfb* gene cluster, which confers antigenic variation and is subject to diversifying selection.

12. Sharp PM, Li W-H: **The codon adaptation index—a measure of directional synonymous codon usage bias and its potential applications.** *Nucleic Acids Res* 1987, **15**:1281–1295.

13. Sharp PM, Shields DC, Wolfe KH, Li W-H: **Chromosomal location and evolutionary rate variation in Enterobacterial genomes.** *Science* 1989, **246**:808–810.

14. Deschavanne P, Filipski J: **Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes.** *Nucleic Acids Res* 1995, **23**:1350–1353.

15. Arber W: **The generation of variation in bacterial genomes.** *J Mol Evol* 1995, **40**:7–12.

16. Groisman E, Saier MH Jr., Ochman H: **Horizontal transfer of a phosphatase gene as evidence for mosaic structure of the *Salmonella* genome.** *EMBO J* 1992, **11**:1309–1316.

17. Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A: **Evidence for horizontal gene transfer in *Escherichia coli* speciation.** *J Mol Biol* 1991, **222**:851–856.

18. Krawiec S, Riley M: **Organization of the bacterial chromosome.** *Microbiol Rev* 1990, **54**:502–539.

19. Riley M, Sanderson KE: **Comparative genetics of *Escherichia coli and Salmonella typhimurium*.** In *The Bacterial Chromosome*. Edited by Drlica K, Riley M. Washington DC: ASM Press, 1990:85–95.

20. • Mills DM, Bajaj V, Lee CA: **A 40-kb chromosomal fragment encoding *Salmonella typhimurium* invasion genes is absent from the corresponding region of the *Escherichia coli* K-12 chromosome.** *Mol Microbiol* 1995, **15**:749–759.
Although the functions of the majority of sequences contained within chromosome loops are not known, a large number of genes required for the entry of salmonellae into host epithelial cells map to a single region of the chromosome. A large gene cluster—the *inv/spa* complex—resides in this region and shows broad-scale similarities to genes specifying protein export assemblies in other pathogenic bacteria.

21. Cole ST, Saint Girons I: **Bacterial genomics.** *FEMS Microbiol Rev* 1994, **14**:139–160.

22. Romling U, Tummler B: **Bacterial genome mapping.** *J Biotechnol* 1994, **35**:155–164.

23. Fonstein M, Haselkorn R: **Physical mapping of bacterial genomes.** *J Bacteriol* 1995, **177**:3361–3369.

24. Smith CL, Econome JG, Schutt A, Klco S, Cantor CR: **A physical map of the *Escherichia coli* K12 genome.** *Science* 1987, **236**:1448–1453.

25. Perkins JD, Heath JD, Sharma BR, Weinstock GM: ***Xba*I and *Bln*I genomic cleavage maps of *Escherichia coli* K-12 strain MG1655 and comparative analysis of other strains.** *J Mol Biol* 1993, **232**:419–445.

26. • Louarn J, Cornet F, François V, Patte J, Louarn J-M:. **Hyper-recombination in the terminus region of the *Escherichia coli* chromosome: possible relation to nucleoid organization.** *J Bacteriol* 1994, **176**:7524–7531.
This study examines the excision rate of a prophage inserted into several chromosomal locations. Their findings are interesting in light of the distribution of known genes in the bacterial chromosome and relates to the variation observed in physical analyses of genome structure.

27. Arbeit RD, Arthur M, Dunn R, Kim C, Selander RK, Goldstein R: **Resolution of recent evolutionary divergence among *Escherichia coli* from related lineages: the application of pulsed field electrophoresis to molecular epidemiology.** *J Infect Dis* 1990, **161**:230–235.

28. Bohm H, Karch H: **DNA fingerprinting of *Escherichia coli* O157:H7 strains by pulsed-field gel electrophoresis.** *J Clin Microbiol* 1992, **30**:2169–2172.

29. Harsono KD, Kaspar CW, Luchansky JB: **Comparison and genomic sizing of *Escherichia coli* O157:H7 isolates by pulsed-field gel electrophoresis.** *Appl Environ Microbiol* 1993, **59**:3141–3144.

30. Meng J, Zhao S, Zhao T, Doyle MP: **Molecular characterization of *Escherichia coli* O157:H7 isolates by pulsed-field gel electrophoresis and plasmid DNA analysis.** *J Med Microbiol* 1995, **42**:258–263.

31. Liu S-L, Hessel A, Sanderson KE: **The *Xba*I-*Bln*I-*Ceu*I genomic cleavage map of *Salmonella enteritidis* shows an inversion relative to *Salmonella typhimurium* LT2.** *Mol Microbiol* 1993, **10**:655–664.

32. Liu S-L, Hessel A, Cheng HYM, Sanderson KE: **The *Xba*I-*Bln*I-*Ceu*I genomic cleavage map of *Salmonella paratyphi* B.** *J Bacteriol* 1994, **176**:1014–1024.

33. •• Liu S-L, Sanderson KE: **Rearrangements in the genome of the bacterium *Salmonella typhi*.** *Proc Natl Acad Sci USA* 1995, **92**:1018–1022.
The physical map of the chromosome of *S. enterica* serovar Typhi Ty2 has shown several large-scale rearrangements, including inversions due to homologous recombination between *rrn* loci and several insertions of up to 118 kb in length, when compared with other serovars of *S. enterica*. The authors speculate that these rearrangements are related to patterns of virulence.

34. Liu S-L, Hessel A, Sanderson KE: **Genomic mapping with I-*Ceu*I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in *Salmonella* spp., *Escherichia coli*, and other bacteria.** *Proc Natl Acad Sci USA* 1993, **90**:6874–6878.

35. • Liu S-L, Sanderson KE: **I-*Ceu*I reveals conservation of the genome of independent strains of *Salmonella typhimurium*.** *J Bacteriol* 1995, **177**:3355–3357.
The recognition sites of I-*Ceu*I are rare, and well conserved, among clinical isolates of Salmonella. By constructing physical maps of these strains, the authors identify regions that are subject to size variation and determine the source of chromosome rearrangements.

36. Wong KK, McClelland M: **A *Bln*I restriction map of the *Salmonella typhimurium* genome.** *J Bacteriol* 1992, **174**:1656–1661.

37. François V, Louarn J, Rebello JE, Louarn J-M: **Replication termination, non-divisable zones, and structure of the *Escherichia coli* chromosome.** In *The Bacterial Chromosome*. Edited by Drlica K, Riley M. Washington DC: ASM Press, 1990:351–359.

38. Reeves MV, Evins GM, Heiba AA, Plikaytis BD, Farmer JJ III: **Clonal nature of *Salmonella typhi* and its genetic relatedness to other salmonellae as shown by multilocus enzyme electrophoresis and proposal of *Salmonella bongori* comb. nov.** *J Clin Microbiol* 1989, **27**:313–320.

39. Pang T, Altwegg M, Martinetti G, Koh CL, Puthucheary S: **Genetic variation among Malaysian isolates of *Salmonella typhi* as detected by ribosomal RNA gene restriction patterns.** *Microbiol Immunol* 1992, **36**:539–543.

40. • Bergthorsson U, Ochman H: **Heterogeneity of genome sizes among natural isolates of *Escherichia coli*.** *J Bacteriol* 1995, in press.
Early studies based on DNA reassociation procedures indicated that genomes sizes within *E. coli* were highly variable, differing by as much as 30% in length, whereas comparisons among enteric bacteria implied that genomes were well conserved. To reconcile these findings, this study used PFGE to analyze the genomes of *E. coli* strains spanning the range of genetic diversity observed in the species at large.

H Ochman and U Bergthorsson, Department of Biology, Hutchison Hall, University of Rochester, Rochester, NY 14627, USA.

Author for correspondence: H Ochman.
E-mail: ochman@ho.biology.rochester.edu