

Parallel evolution of virulence in pathogenic *Escherichia coli*

Sean D. Reid, Corinne J. Herbelin, Alyssa C. Bumbaugh, Robert K. Selander & Thomas S. Whittam

Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, Pennsylvania 16802, USA

The mechanisms underlying the evolution and emergence of new bacterial pathogens are not well understood. To elucidate the evolution of pathogenic *Escherichia coli* strains, here we sequenced seven housekeeping genes to build a phylogenetic tree and trace the history of the acquisition of virulence genes. Compatibility analysis indicates that more than 70% of the informative sites agree with a single phylogeny, suggesting that recombination has not completely obscured the remnants of ancestral chromosomes¹⁻³. On the basis of the rate of synonymous substitution for *E. coli* and *Salmonella enterica* (4.7×10^{-9} per site per year³), the radiation of clones began about 9 million years ago and the highly virulent pathogen responsible for epidemics of food poisoning, *E. coli* O157:H7, separated from a common ancestor of *E. coli* K-12 as long as 4.5 million years ago. Phylogenetic analysis reveals that old lineages of *E. coli* have acquired the same virulence factors in parallel, including a pathogenicity island involved in intestinal adhesion, a plasmid-borne haemolysin, and phage-encoded Shiga toxins. Such parallel evolution indicates that natural selection has favoured an ordered acquisition of genes and the progressive build-up of molecular mechanisms that increase virulence.

To elucidate the evolution of virulence and pathogenic mechanisms, we studied 14 strains representing common clones of enteropathogenic *E. coli* (EPEC), an important cause of infantile diarrhoea in the developing world, and enterohaemorrhagic *E. coli* (EHEC), one of the primary food-borne pathogens in the industrialized world⁴. In addition, we examined strains of other Shiga-toxin-producing *E. coli* serotypes and laboratory strain K-12. Seven housekeeping genes spaced around the *E. coli* chromosome (Fig. 1) were chosen for nucleotide sequencing. Five of these loci (*icd*, *mdh*, *mtlD*, *pgi* and *aroE*) have been shown to be polymorphic by enzyme electrophoresis⁵. In addition to these genes, we sequenced *rpoS*, a gene encoding sigma factor 38, which is involved in stress response, and *arcA*, a gene encoding the regulatory protein involved in the control of aerobic respiration. In the 21 *E. coli* strains studied, variation per locus ranges between 1% and 8% at both the nucleotide and the amino-acid level (Fig. 1).

Our first step in the phylogenetic analysis was to use the multilocus sequence data to reconstruct the evolutionary history of the pathogenic clones. This is complicated by the fact that recombination, in addition to mutation, contributes to clonal divergence^{1,3}. In nature, as a bacterial lineage accumulates mutations, bits of the genome are replaced by recombination, resulting in DNA sequences that consist of a clonal frame (relic of the ancestral chromosome) interrupted by recombined segments². To detect the effects of recombination on sequence divergence, we analysed the compatibility matrix⁶ of polymorphic nucleotide sites in the housekeeping loci. Two sites are compatible if a phylogeny exists in which all nucleotide changes at the sites can be inferred to have occurred only once. Incompatible sites require several changes, indicating that these sites have experienced recombination or repeated mutation. Figure 2 plots the compatibility matrix between all pairs of 201 binary sites (informative sites that have only two different nucleotides⁶) identified in a total of 5,664 bases sequenced in 21 strains. The observed compatibility is 74%, compared with 37% for

randomized matrices in a star phylogeny, and the neighbourhood similarity (the fraction of adjacent squares of the same colour) is 77%. For comparison, this value significantly exceeds the neighbourhood score (67%) obtained for 1,000 randomized matrices where the positions of sites are shuffled each time.

The total compatibility can be separated into two components: a 'within-locus' component based on the comparison of pairs of sites at the same locus (the fraction of coloured squares within the triangular regions of Fig. 2), and a 'between-locus' component based on the comparison of sites at different loci (the fraction of compatible pairs in the rectangular regions in Fig. 2). In general, there is a positive relationship between the within- and between-locus compatibilities (Fig. 2b). Overall, 85% of the pairs of sites within a locus are compatible, compared with 71% of the pairs at different loci. This observation indicates that even sites that are widely spaced on the bacterial chromosome retain similar phylogenetic information.

The plot of the compatibility matrix reveals two regions with a high concentration of incompatible sites (Fig. 2); blocks such as these are the hallmarks of past recombination⁶. One region includes a 42-nucleotide stretch starting at codon 317 in *mtlD*. The removal of the ten polymorphic sites in this segment increases the average between-locus compatibility to 72% (Fig. 2b). The second region is a cluster of eight polymorphic sites at the 3' end of *icd*. These sites reside on a DNA segment directly downstream from the insertion site of a naturally occurring bacteriophage⁷. When this 129-base pair (bp) segment is excluded from the analysis, the within-locus compatibility for *icd* increases from 73% to 86% (Fig. 2b). The removal of both incompatible regions increases the overall compatibility of sites from 74% to 78%.

What is the quality of the phylogenetic signal? To address this question, we used the method of split decomposition, which does not force the data into a tree-like phylogeny. Instead, this method allows for and thus can detect conflicting phylogenetic information⁸. This method is particularly useful in situations where a bifurcating tree is an inappropriate evolutionary model, such as for viral quasispecies⁹ or bacterial populations with high rates of interstrain gene transfer¹⁰. In these cases, split decomposition results in a mesh-like network of interconnected nodes reflecting the reticulate nature of their evolution.

Split decomposition of the multilocus sequence data shows that the clonal frames of the pathogenic *E. coli* are connected in a network that is more tree-like than mesh-like (Fig. 3a). Six nodes are simple bifurcations and two nodes are multifurcations. There is

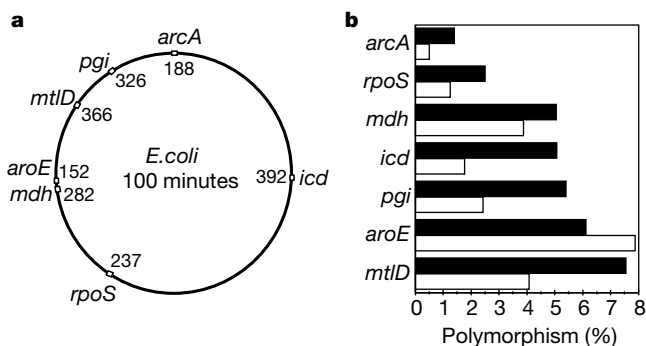


Figure 1 Genomic locations and variability of *E. coli* housekeeping genes. **a**, Location of 7 genes on the 100-minute map of the *E. coli* chromosome. The genes are clockwise as follows (protein, minute on map): *icd* (isocitrate dehydrogenase, 25) *rpoS* (sigma factor 38, 59), *mdh* (malate dehydrogenase, 73), *aroE* (shikimate dehydrogenase, 74), *mtlD* (mannitol 1-phosphate dehydrogenase, *pgi* (phosphoglucose isomerase, 91) and *arcA* (aerobic respiratory control protein, 99). The number of codons sequenced is given next to each locus. **b**, The percentage of variable positions at the nucleotide (black bars) and amino-acid level (white bars).

only one case of parallel paths, indicative of a past recombinational exchange, which involves the relationship among E2348/69, DEC 2a, and the uropathogenic strain 536 (Fig. 3a).

Knowing that the underlying network is tree-like, we inferred a phylogeny for the strains that were based on sequences of the six loci (a combined total of 1,680 codons in 5 kilobases (kb)) that could be rooted with homologous genes from *Salmonella enterica*. A neighbour-joining tree revealed several main branches supported by bootstrap values greater than 85% (Fig. 3b). The same groupings, with trivial changes in topology, were found for the single most parsimonious tree, a unrooted phylogeny based on 7 loci which required a total of 383 mutations.

We compared the rooted phylogeny with the molecular clock hypothesis by the two-cluster method¹¹ to estimate the timescale involved. Nucleotide sequence data from 15 of the 21 strains fit the molecular clock; the most notable exception was strain K-12, which had a significantly longer branch than expected. To infer a timescale for the radiation of clones, we used the rate of synonymous substitution for *E. coli* and *S. enterica* of 4.7×10^{-9} per site per year³, which is based on a speciation date of 100 million years (Myr) ago³. This evolutionary rate for bacterial genes lies between the rates for nuclear genes of mammals (3.51×10^{-9}) and *Drosophila* (15.6×10^{-9})¹². The amount of divergence at synonymous sites indicates that the *E. coli* clonal frames trace back to a common ancestor that existed about 9.0 (8.1–9.9) Myr ago. It also indicates that the lineages destined to give rise to *E. coli* K-12 and O157:H7 separated about 4.5 (3.9–5.0) Myr ago.

The clonal phylogeny derived from multilocus sequencing is concordant with previous findings that were based on enzyme electrophoresis⁵. Three of the four pathogenic groups (EHEC 1, EHEC 2, and EPEC 1) are supported by bootstrap values greater than 85% (Fig. 2b). Members of the EPEC 2 group cluster together, but only two strains (DEC 11a and DEC 12a) have significant bootstrap support. The phylogenetic analysis also shows that the *E. coli* O55:H7 clone shares a recent common ancestor with *E. coli* O157:H7 (Fig. 3b) and is consistent with other features of the stepwise model for the evolution of this pathogen¹³.

Comparison of the distribution of specific genes that mark

Table 1 Virulence factors in pathogenic *E. coli* detected by PCR

Strain	Serotype	Chromosomal*			Plasmid†		
		<i>eae</i>	<i>stx1</i>	<i>stx2</i>	<i>ehly</i>	<i>katP</i>	<i>bfpA</i>
DEC5d	O55:H7	γ	-	-	-	-	-
5905	O55:H7	γ	-	+	-	-	-
493/89	O157:H-	γ	-	+	+	-	-
93-111	O157:H7	γ	+	+	+	+	-
OK-1	O157:H7	γ	+	+	+	+	-
921	O111:H9	+	-	-	-	-	-
2666-74	O26:H-	β	-	-	+	-	-
CL-37	O111:H8	+	+	-	-	-	-
DEC8b	O111:H8	+	+	+	+	-	-
90-1787	X03:H-	-	-	+	-	-	-
CL-3	O113:H21	-	-	+	-	-	-
G5506	O104:H21	-	-	+	+	-	-
B2F1	O91:H21	-	-	+	-	-	-
B170	O111:H-	β	-	-	-	-	+
DEC12a	O111:H2	β	-	-	-	-	+
DEC11a	O128:H2	β	-	-	-	-	+
536	O6:H31	-	-	-	-	-	-
E2348/69	O127:H6	α	-	-	-	-	+
DEC 1a	O55:H6	α	-	-	-	-	+
DEC 2a	O55:H6	α	-	-	-	-	+

* Chromosomal genes include *eae* (intimin) on the LEE pathogenicity island, and *stx1* and *stx2* (Shiga-toxins 1 and 2) encoded by bacteriophages.

† Plasmid encoded genes includes *ehly* (EHEC haemolysin) and *katP* (catalase) found on the EHEC plasmid (pO157) and *bfpA* (pilin subunit of bundle-forming pili) encoded on the EAF plasmid.

‡ Carries a variant *ehly* allele (P. Feng, personal communication).

mobile elements associated with virulence (Table 1) supports the hypothesis that the high virulence of clones is a recent, derived state resulting from acquisition of virulence genes rather than an ancestral condition of primitive *E. coli*. The principal virulence factors and their mobile elements include the intimin gene (*eae*) encoded on a ~35-kb pathogenicity island called LEE (locus of enterocyte effacement); Shiga toxin genes (*stx1* and *stx2*) associated with bacteriophages; an enterohaemolysin gene (*ehly*) and catalase (*katP*), both of which occur on a ~90-kb EHEC plasmid¹⁴; and the main pilin subunit (*bfpA*) encoded on the EAF (EPEC adherence factor) plasmid. *Stx* genes and their bacteriophages are widely disseminated in the *E. coli* population¹⁵ and virtually identical *stx* sequences have been recovered from a variety of strains, indicating

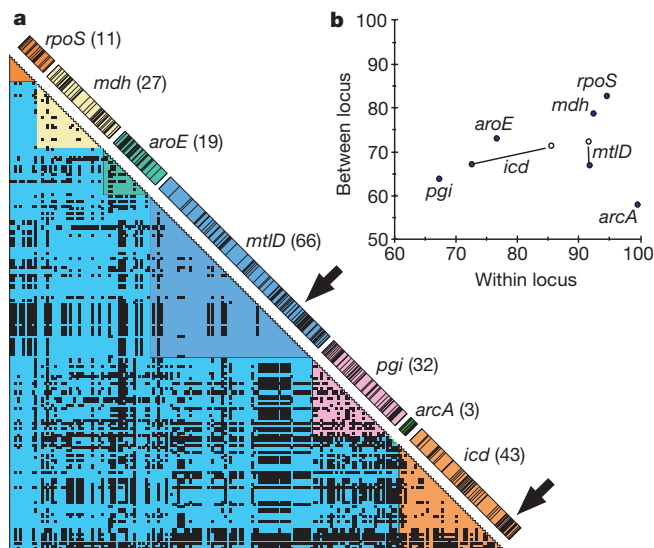


Figure 2 Compatibility of nucleotide polymorphisms. **a**, Lower triangle is a plot of all pairwise comparisons of 201 informative sites that are phylogenetically compatible within loci (solid colours) or between loci (blue). Incompatible pairs of sites are marked with a black square. The locations of the polymorphic sites within a locus are marked on the diagonal elements. For each locus, the number of informative sites is given in parentheses. The arrows mark two regions where incompatible sites within a locus are

clustered along the sequences. These regions include 10 informative sites in *mtlD* and 8 sites in the 3' end of *icd*. **b**, Average level of compatibility within and between loci shows that sites in *mdh* and *rpoS* have the greatest compatibility, whereas sites in *pgi* show the least compatibility. The lines show the effect of the removal of the incompatible regions in *icd* and *mtlD*.

Biosystems 373A automated sequencer. Raw sequences of both DNA strands were analysed and concatenated by DNASTAR with additional internal sequencing primers designed based on the generated sequence data. All conflicting and putative polymorphic sites were sequenced at least three times to reduce sequencing error.

Phylogenetic analysis

Multiple-sequence alignment of the inferred amino-acid sequences was performed with Clustal W²⁶. A supergene was constructed for phylogenetic analysis by concatenating the individual gene sequences in the order: *rpoS*, *mdh*, *aroE*, *mtlD*, *pgi*, *arcA* and *icd*. The program Reticulate was used to identify putative regions of recombination or gene conversion through the construction of compatibility matrices⁶. A Monte Carlo approach was used to evaluate the significance of the matrices. Only binary parsimoniously informative sites were retained with incompatible regions removed from the analysis. Split decomposition was performed with the SplitsTree program²⁷.

Phylogenetic trees were inferred by the neighbour-joining algorithm using MEGA²⁸ or by the method of maximum parsimony using PHYLIP²⁹. The phylogeny based on 6 loci (combined total of 5,042 bp; *aroE* was excluded) was rooted with the homologous sequences from *S. enterica* extracted from the GenBank database. In the combined sequences, there were a total of 726 variable sites, 89 of which involved amino-acid replacements among the 22 sequences. Among the 21 *E. coli* sequences, there were 216 polymorphic sites, 40 of which predict amino-acid differences. Bootstrap confidence values were based on 1,000 simulated trees. Linearized trees and tests of the molecular clock hypothesis were based on the LinTree programs¹¹. The rate of synonymous substitution (d_s) and its standard error was calculated using MEGA²⁸. Confidence intervals for the divergence times were calculated from the standard errors of the genetic distances. The divergence times for nodes in the rooted phylogeny (Fig. 3b) were based on pair-wise comparison of 93-111 and E2348/69 ($d_s \times 100 = 8.46 \pm 0.81$), and 93-111 and CL 37 ($d_s \times 100 = 4.18 \pm 0.56$) for the synonymous sites in 1,888 codons combined from 7 genes.

Received 15 November 1999; accepted 19 April 2000.

1. Guttman, D. S. & Dykhuizen, D. E. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. *Science* **266**, 1380–1383 (1994).
2. Milkman, R. in *Escherichia coli and Salmonella: Cellular and Molecular Biology* 2nd edn (eds Neidhardt, F. C. et al.) 2663–2684 (American Society for Microbiology, Washington DC, 1996).
3. Lawrence, J. G. & Ochman, H. Molecular archaeology of the *Escherichia coli* genome. *Proc. Natl Acad. Sci. USA* **95**, 9413–9417 (1998).
4. Nataro, J. P. & Kaper, J. B. Diarrheagenic *Escherichia coli*. *Clin. Microbiol. Rev.* **11**, 142–201 (1998).
5. Whittam, T. S. et al. Clonal relationships among *Escherichia coli* strains that cause hemorrhagic colitis and infantile diarrhea. *Infect. Immun.* **61**, 1619–1629 (1993).
6. Jakobsen, I. B. & Easteal, S. A program for calculating and displaying compatibility matrices as an aid in determining reticulate evolution in molecular sequences. *CABIOS* **12**, 291–295 (1996).
7. Wang, F. S., Whittam, T. S. & Selander, R. K. Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica*. *J. Bacteriol.* **179**, 6551–6559 (1997).
8. Bandelt, H. J. & Dress, A. W. M. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol. Phylogenet. Evol.* **1**, 242–252 (1992).
9. Dopazo, J., Dress, A. & von Haeseler, A. Split decomposition: a technique to analyze viral evolution. *Proc. Natl Acad. Sci. USA* **90**, 10320–10324 (1993).
10. Holmes, E. C., Urwin, R. & Maiden, M. C. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. *Mol. Biol. Evol.* **16**, 741–749 (1999).
11. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**, 823–833 (1995).
12. Li, W.-H. *Molecular Evolution* (Sinauer Associates, Sunderland, Massachusetts, 1997).
13. Feng, P., Lampel, K. A., Karch, H. & Whittam, T. S. Genetic and phenotypic changes in the emergence of *Escherichia coli* O157:H7. *J. Infect. Dis.* **177**, 1750–1753 (1998).
14. Karch, H. in *Escherichia coli* O157:H7 and Other Shiga Toxin-Producing *E. coli* Strains (eds Kaper, J. B. & O'Brien, A. D.) 183–194 (ASM, Washington DC, 1998).
15. Newland, J. W. & Neill, R. J. DNA probes for Shiga-like toxins I and II and for toxin-converting bacteriophages. *J. Clin. Microbiol.* **26**, 1292–1297 (1988).
16. Schmidt, H., Beutin, L. & Karch, H. Molecular analysis of the plasmid-encoded hemolysin of *Escherichia coli* O157:H7 strain EDL 933. *Infect. Immun.* **63**, 1055–1061 (1995).
17. Schmidt, H. & Karch, H. Enterohemolytic phenotypes and genotypes of shiga toxin-producing *Escherichia coli* O111 strains from patients with diarrhea and hemolytic-uremic syndrome. *J. Clin. Microbiol.* **34**, 2364–2367 (1996).
18. Boyd, E. F. & Hartl, D. L. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. *J. Bacteriol.* **180**, 1159–1165 (1998).
19. Wieler, L. H., McDaniel, T. K., Whittam, T. S. & Kaper, J. B. Insertion site of the locus of enterocyte effacement in enteropathogenic and enterohemorrhagic *Escherichia coli* differs in relation to the clonal phylogeny of strains. *FEMS Microbiol. Lett.* **156**, 49–53 (1997).
20. Sperandio, V. et al. Characterization of the locus of enterocyte effacement (LEE) in different enteropathogenic *Escherichia coli* (EPEC) and Shiga-toxin producing *Escherichia coli* (STEC) serotypes. *FEMS Microbiol. Lett.* **164**, 133–139 (1998).
21. McGraw, E. A., Li, J., Selander, R. K. & Whittam, T. S. Molecular evolution and mosaic structure of α , β , and γ intimins of pathogenic *Escherichia coli*. *Mol. Evol. Biol.* **16**, 12–22 (1999).
22. LeClerc, J. E., Li, B., Payne, W. L. & Cebula, T. A. High mutation frequencies among *Escherichia coli* and *Salmonella* pathogens. *Science* **274**, 1208–1211 (1996).
23. Karalis, D. K., Somara, S., Maneval, D. R., Jr., Johnson, J. A. & Kaper, J. B. A bacteriophage encoding a pathogenicity island, a type-IV pilus and a phage receptor in cholera bacteria. *Nature* **399**, 375–379 (1999).
24. Whittam, T. S. in *Escherichia coli* O157:H7 and other Shiga toxin-producing *E. coli* strains (eds Kaper, J. B. & O'Brien, A. D.) 195–209 (ASM, Washington DC, 1998).

25. Reid, S. D., Betting, D. J. & Whittam, T. S. Molecular detection and identification of intimin alleles in pathogenic *Escherichia coli* by multiplex PCR. *J. Clin. Microbiol.* **37**, 2719–2722 (1999).
26. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).
27. Huson, D. H. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics* **14**, 68–73 (1998).
28. Kumar, S., Tamura, K. & Nei, M. MEGA: Molecular evolutionary genetics analysis, version 1.0. (Pennsylvania State Univ., Univ. Park, Pennsylvania, 1993).
29. Felsenstein, J. PHYLIP—Phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166 (1989).

Acknowledgements

The authors thank S. Plock for technical assistance. This research was supported by NIH grants (to T.S.W. and R.K.S.), and the Enteric Pathogen Research Unit at the University of Maryland Medical School.

Correspondence and requests for materials should be addressed to T.S.W. (e-mail: tsw1@psu.edu). Preliminary sequence data for *S. enterica* Typhimurium was obtained from The Institute for Genomic Research website at <http://www.tigr.org>.

.....
Negative genetic correlation between male sexual attractiveness and survival

Robert Brooks

School of Tropical Biology, James Cook University, Townsville 4811, Queensland, Australia

.....
Indirect selection of female mating preferences may result from a genetic association between male attractiveness and offspring fitness^{1,2}. The offspring of attractive males may have enhanced growth^{3–5}, fecundity^{3,4}, viability^{5–8} or attractiveness^{4,9–11}. However, the extent to which attractive males bear genes that reduce other fitness components has remained unexplored. Here I show that sexual attractiveness in male guppies (*Poecilia reticulata*) is heritable and genetically correlated with ornamentation. Like ornamentation^{12–14}, attractiveness may be substantially Y-linked. The benefit of mating with attractive males, and thus having attractive sons, is opposed by strong negative genetic correlation between attractiveness and both offspring survival and the number of sons maturing. Such correlations suggest either antagonistic pleiotropy between attractiveness and survival or linkage disequilibrium between attractive and deleterious alleles. The presence of many colour pattern genes on or near the non-recombining section of the Y chromosome may facilitate the accumulation of deleterious mutations by genetic hitchhiking^{15,16}. These findings show that genes enhancing sexual attractiveness may be associated with pleiotropic costs or heavy mutational loads.

The genetic relationships between male ornamentation, attractiveness to females and several fitness components in guppies were investigated. Female mate choice in this species is based on polymorphic male colour patterns¹⁷, and body³ and tail size. Male attractiveness was significantly heritable owing to additive genetic contributions from sires (44 sires, 98 dams, 280 offspring; $h_s = 0.596 \pm 0.28$; $P = 0.044$) but not from dams (nested within sire) ($h_D^2 = 0.101 \pm 0.26$; $P = 0.71$; combined sire and dam h^2 estimate, 0.348 ± 0.15 , $CV_A = 29.75$; $h_s^2 > h_D^2$, $P = 0.05$, randomization test). Furthermore, there was a significant genetic correlation between attractiveness and ornamentation (Fig. 1), implying a common genetic basis, probably because attractiveness is based on ornamentation.

Predation is generally thought to be the most important determinant of juvenile guppy mortality in the field (12–24% survive to