doi:10.1046/j.1420-9101.2003.00612.x

A molecular phylogeny of enteric bacteria and implications for a bacterial species concept

J. E. WERTZ,* C. GOLDSTONE,* D. M. GORDON† & M. A. RILEY*

*Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA
†Division of Botany and Zoology, Australian National University, Canberra, ACT 0200, Australia

Keywords:

auxiliary;
core;
Enterobacteriaceae;
genome;
phylo-phenetic;
prokaryotic;
taxonomy.

Abstract

A molecular phylogeny for seven taxa of enteric bacteria (*Citrobacter freundii, Enterobacter cloacae, Escherichia coli, Hafnia alvei, Klebsiella oxytoca, Klebsiella pneumoniae*, and *Serratia plymuthica*) was made from multiple isolates per taxa taken from a collection of environmental enteric bacteria. Sequences from five housekeeping genes (*gapA, groEL, gyrA, ompA*, and *pgi*) and the 16s rRNA gene were used to infer individual gene trees and were concatenated to infer a composite molecular phylogeny for the species. The isolates from each taxa formed tight species clusters in the individual gene trees, suggesting the existence of 'genotypic' clusters that correspond to traditional species designations. These sequence data and the resulting gene trees and consensus tree provide the first data set with which to assess the utility of the recently proposed core genome hypothesis (CGH). The CGH provides a genetically based approach to applying the biological species concept to bacteria.

Introduction

The impact of molecular systematics on bacterial classification has been profound. Indeed, phylogenies based on the sequence of ribosomal RNA genes have forever changed how we view the organization of life on this planet (Fox *et al.*, 1980; Olsen & Woese, 1993). Molecular approaches have revealed three (Archaea, Bacteria and Eukarya), rather than five (Animalia, Plantae, Fungi, Protista and Monera), primary divisions of life and forced an acknowledgement of the extraordinary levels of microbial diversity (Fox *et al.*, 1980; Pace *et al.*, 1985; Woese, 1987; Woese *et al.*, 1990). Further, as additional highly conserved genes, such as those encoding elongation factors and ribosomal proteins are examined, we gain confidence that the 16s rRNA based phylogeny provides a fairly robust description of the major evolutionary lineages (Ludwig *et al.*, 1993; Brown *et al.*, 2001).

Just as molecules appear to have solved some of the outstanding phylogenetic questions, their application has generated an entirely new and unexpected controversy. Molecular phylogenies have revealed that horizontal transfer plays an important and unexpected role in evolution (Kidwell, 1993; Nelson & Selander, 1994; Brown & Doolittle, 1997; Nesbo *et al.*, 2001). Recent observations of possible gene transfer events between some of the deepest branches represented in the 16s-based tree of life have raised the question of whether we should employ networks, rather than dichotomously branching trees, to represent evolutionary lineages over time (Doolittle, 1999). In fact, the importance of lateral gene transfer (LGT) has been elevated to such a degree that it has called into question the very existence of microbial species. If transfer is so rampant, how can we define microbial species and how could they exist?

To address this challenge, Lan & Reeves (2001) revised a proposal first developed by Dykhuizen & Green (1991) that applies the biological species concept (BSC) (Mayr, 1942) to bacteria. Dykhuizen and Green analysed the sequences of three housekeeping genes (*gnd, trp* and *phoA*) (DuBose *et al.*, 1988; Stoltzfus *et al.*, 1988; Dykhuizen & Green, 1991) and, as was previously observed by Milkman for several additional genes (McKane & Milkman, 1995; Milkman, 1997), found high levels of recombination among *Escherichia coli* isolates. Dykhuizen and Green suggest that there is sufficient recombination within the *E. coli* species to ensure a shared gene pool, as required by the BSC (Mayr, 1942).

Correspondence: John E. Wertz, 165 Prospect Street, Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06511, USA.
Tel.: 203 432 3877; fax: 203 432 6533;
e-mail: john.wertz@yale.edu

Lan & Reeves (2001) extended this original proposal of Dykhuizen & Green (1991). They distinguish between what they call core and auxiliary genes. Core genes are essentially housekeeping genes and each isolate of a species possesses a full (or nearly so) compliment of core genes. Lan and Reeves argue that there is no strong selective advantage to acquiring new core genes through LGT and thus these genes comprise the species 'shared genome'. These genes would rarely transfer between taxa and their eventual divergence would serve as barriers to homologous recombination between species. The auxiliary genes, in contrast, might transfer frequently and serve in adapting bacterial isolates to local competitive or environmental pressures (Cohan, 1996; Cohan, 2001). The auxiliary genes might include pathogenicity islands, resistance genes and cassettes, novel metabolic functions, toxin genes, etc. (Dobrindt & Reidl, 2000; Karlin, 2001; White *et al.*, 2001). The core genome hypothesis (CGH) predicts the existence of a barrier to interspecies recombination for core genes, which is not shared by auxiliary genes (Lan & Reeves, 2000). At present, there are no appropriate data sets with which to test this prediction.

In this paper we provide a molecular-based phylogeny of representatives of the enteric family of bacteria and discuss the implications this information has on implementing a tractable bacterial species concept. In order to construct a robust phylogeny for enteric bacteria, we employ a multi-locus DNA sequence approach. Portions of five housekeeping genes (*gapA, groEL, gyrA, ompA*, and *pgi*) and the 16s rRNA gene have been sequenced, and phylogenetic trees inferred for seven taxa of the *Enterobacteriaceae* (*Citrobacter freundii, Enterobacter cloacae, E. coli, Hafnia alvei, Klebsiella oxytoca, Klebsiella pneumoniae*, and *Serratia plymuthica*). Multiple isolates of each taxa were included in our sequence analysis and tree construction to allow an estimate of the within and between species levels of nucleotide diversity.

All bacterial isolates were obtained from a collection of environmental enteric bacteria isolated from Australian mammals (Gordon & FitzGibbon, 1999). This collection is comprised of over 951 strains, representing 24 species of enteric bacteria isolated from 642 wild mammals in Australia and represents the most extensive sample of wild enteric bacteria reported to date. In addition, Gordon and colleagues have generated a large and growing body of phenotypic and genetic data about the collection that continues to increase its value as a research tool (Gordon & Lee, 1999; Okada & Gordon, 2001). For instance, the collection has been screened for bacteriocin production, antibiotic resistance, and plasmid profiles (Sherley *et al.*, 2000; M. Sherley, pers. comm.). The collection has also been examined for correlations between species diversity and host organism or geographic effects (Gordon & FitzGibbon, 1999; Gordon & Lee, 1999).

In this study, housekeeping genes serve as our proxy of the species 'shared, core genome' (as described by Lan & Reeves, 2001). The ideal situation would be to obtain numerous genome sequences for our sample of taxa. However, a more realistic solution (with respect to both financial and time considerations) is to subsample the core genome and then pool these data to generate a molecular phylogeny. Housekeeping genes are an appropriate focus for this study for a number of reasons. They are a class of highly expressed, highly conserved, protein encoding genes that exhibit a high degree of codon bias. These genes evolve more slowly than typical protein encoding genes, but more rapidly than rRNA genes, and are therefore often used to construct gene trees of closely related taxa (Lawrence *et al.*, 1991). We investigate the molecular phylogenies inferred for each locus separately and pooled, and argue that the pooled data provides an appropriate estimate of the enteric phylogeny. In addition, this phylogeny reveals that the phenotypic 'clusters' that have traditionally been used to define bacterial species (Holt, 1994; Rossello-Mora & Amann, 2001) are clearly seen as gene pools, or genome pools, at the DNA sequence level. We argue from these data that, for at least the enteric bacteria, bacterial species do exist and can be defined in similar ways with both phenotypic and genotypic data.

Materials and methods

Strains

DNA sequences from a total of 38 strains were used in this study. DNA sequences from two fully sequenced genomes [*E. coli* MG1655 (Blattner *et al.*, 1997) and *Vibrio cholerae* biotype *El Tor* (Heidelberg *et al.*, 2000)] were obtained from GenBank. A 36 strain subset of a collection of environmental enteric bacteria isolated from wild Australian mammals, which is also used by our lab to study bacteriocin ecology and evolution, was used for DNA sequence determination (Gordon & FitzGibbon, 1999). Information about the strains, including species designation, geographic origin and host is listed in Table 1. Although strain SM1 is identified in the Gordon collection as *Serratia marcescens*, for the genes we examined, it is identical to *S. plymuthica*. For ease of discussion, we will consider it a member of the later species.

Gene selection

Housekeeping genes were selected based on a number of criteria. They must be essential for the cells survival in its natural environment. Other considerations were that they be spaced far enough apart on the *E. coli* chromosome that they not be co-transducible (>100 kb), that sequence exist in GenBank for as many of the target species as possible, that there be no known paralogs and that the genes selected do not over-sample any one physiological process to prevent concordance in the gene trees as an artifact of co-evolution. Gene locations and

Table 1 Enteric bacterial strains employed in this study.

Strain designation	Collection number	Species	Source organism	State*
CF1	M250	*C. freundii*	*Isoodon macrourus*	NT
CF2	M289	*C. freundii*	*Perameles nasuta*	NSW
CF3	M141	*C. freundii*	*Antechinus flavipes*	SA
CF4	M140	*C. freundii*	*Antechinus flavipes*	SA
CF5	M255	*C. freundii*	*Isoodon macrourus*	NT
EB3	M322	*E. cloacae*	*Mus musculus*	VIC
EB8	M338	*E. cloacae*	*Mus musculus*	VIC
EB9	M50	*E. cloacae*	*Mus musculus*	VIC
EB10	M99	*E. cloacae*	*Mus musculus*	VIC
EB11	M90	*E. cloacae*	*Mus musculus*	VIC
EC1	TA57	*E. coli*	*Macropus giganteus*	ACT
EC2	TA79	*E. coli*	*Bettongia penicillata*	WA
EC3	TA157	*E. coli*	*Trichosurus vulpecula*	NT
EC5	TA184	*E. coli*	*Trichosurus caninus*	NSW
EC6	TA234	*E. coli*	*Mus musculus*	VIC
EC7	TA479	*E. coli*	*Mus musculus*	VIC
HA1	M163	*H. alvei*	*Phascogale tapoatafa*	WA
HA2	M230	*H. alvei*	*Antechinus bellus*	NT
HA4	M259	*H. alvei*	*Dasyurus hallucatus*	NT
HA5	M261	*H. alvei*	*Dasyurus hallucatus*	NT
HA6	M690	*H. alvei*	*Homo sapiens*	WA
KO1	M328	*K. oxytoca*	*Trichosurus vulpecula*	TAS
KO2	M499	*K. oxytoca*	*Vespadelus vulturnus*	NSW
KO3	M712	*K. oxytoca*	*Chalinolobus gouldii*	NSW
KO4	M192	*K. oxytoca*	*Zyzomys argurus*	NT
KO5	M151	*K. oxytoca*	*Dasycercus cristicauda*	NT
KP2	M40	*K. pneumoniae*	*Mus musculus*	VIC
KP3	M663	*K. pneumoniae*	*Petaurus gracilis*	QLD
KP4	M208	*K. pneumoniae*	*Parantechinus bilarni*	NT
KP5	M757	*K. pneumoniae*	*Nyctophilus geoffroyi*	NSW
KP6	M758	*K. pneumoniae*	*Zyzomys argurus*	NT
KP7	M47	*K. pneumoniae*	*Mus musculus*	VIC
SP1	M8	*S. plymuthica*	*Potorous tridactylus*	NSW
SP2	M66	*S. plymuthica*	*Antechinus stuartii*	SA
SP3	M297	*S. plymuthica*	*Perameles nasuta*	NSW
SP4	M145	*S. marcescens*	*Antechinus flavipes*	NSW

*State abbreviations are: ACT, Australian capital territory; QLD, Queensland; NSW, New South Wales; NT, Northern territory; SA, South Australia; TAS, Tasmania; VIC, Victoria and WA, Western Australia.

functions for the five housekeeping genes and the 16s rRNA gene employed in this study are listed in Table 2. The gene locations are based on the *E. coli* MG1655 sequence and there is at present no data available for the locations of these genes in the other taxa examined.

Nucleotide sequencing

The nucleotide sequences for portions of the five housekeeping genes (*gapA, groEL, gyrA, ompA* and *pgi*) and the 16s rRNA gene were obtained by direct sequencing of polymerase chain reaction (PCR) products. PCR reaction mixtures (50 μL) were prepared with 1.25 U *Taq* polymerase (Promega, Madison, WI, USA), 1× *Taq* polymerase buffer (10 mM Tris-HCl [pH 8.3], 50 mM KCl, 1.5 mM $MgCl_2$), 0.2 mM (each) deoxynucleoside triphosphate and 0.2 μM (each) primer. Template DNA was obtained by boiling cells from a single isolated colony in 200 μL of sterile distilled water for 10 min. Two microlitre of the suspension were added to each reaction. Primer sequences used for amplification and sequencing are available as supplementary material from the journal web site. PCR products were purified using the QIAquick PCR purification kit (Qiagen Inc., Valencia, CA, USA). DNA sequencing was performed on the ABI 377 DNA Sequencer (Applied Biosystems, Foster City, CA, USA). 5–20 ng of purified PCR product were added to sequencing reactions and sequencing was performed using Big Dye chemistry according to ABI standard protocols (Applied Biosystems). Sequence lengths and accession numbers for each gene are given in Table 2.

Phylogenetic analysis

With the exception of 16s rRNA sequences, nucleotide sequences were translated using the standard genetic

Table 2 Target loci information.

Gene	Gene product	Map position	Gene length	Sequence length	PIB*	GenBank Acc. no.
gapA	Glyceraldehyde 3-phosphate dehydrogenase A	40.11	996	832	194	AY301101–32
groEL	GroEL protein	94.17	1647	1146	245	AY301231–63
gyrA	DNA gyrase subunit A	50.33	2628	660	226	AY301133–67
ompA	Outer membrane protein A	21.95	1041	526	219	AY301168–201
pgi	Glucose-6-phosphate isomerase	91.21	1650	670	210	AY301202–30
16s	16s rRNA	†	1541	291	30	AY301069–100

*Parsimony informative bases.
†There are seven copies of the 16s rRNA sequence in *E. coli* K-12.

code, and protein sequences were aligned using the ClustalW algorithm (Thompson *et al.*, 1994) and the Gonnet series protein weight matrix in Megalign version 4.05 (Dnastar, Inc.). 16s rRNA nucleotide sequences were aligned using the same software and algorithm, but employed the IUB DNA weight matrix. Maximum likelihood trees were inferred for each gene and the concatenated alignment of all genes using PAUP version 4.0b8 (Swofford, 1997). Optimized parameters for the heuristic algorithm used for building maximum likelihood trees in PAUP were generated by the MODELTEST program version 3.06 (Posada & Crandall, 1998) using the default starting parameters (Neighbor-Joining using a Jukes and Cantor model of evolution). Statistical support of the branch points was tested by performing 500 maximum likelihood bootstrap replications using PAUP version 4.0b8 and by using the program MrBayes version 2.01 (Huelsenbeck & Ronquist, 2001) to generate 50 000 trees, of which the first 5000 trees were discarded as 'burnin', and a majority rule consensus tree was constructed from the remaining 45 000 trees. Maximum parsimony was employed during our first pass at phylogeny reconstruction to estimate the phylogenetic information content of each gene. The number of parsimony informative bases for each housekeeping gene was used to determine if additional sequence data needed to be collected so that each gene contributed an approximately equal amount to the composite phylogeny. In the maximum parsimony analysis, heuristic searches were conducted with all positions weighted equally, gaps were treated as missing and the tree-bisection-reconnection branch-swapping algorithm was used.

Gene trees

For each gene, a protein sequence was inferred from the DNA sequences generated and a protein alignment produced. Protein alignments were then converted back to DNA alignments for phylogenetic inference. Sequences from *V. cholerae* biotype *El Tor* were obtained from GenBank and used as the outgroup for rooting each of the gene trees. For some loci, shorter sequences were used for *Hafnia* (*pgi* and *gapA*), *Serratia* (*pgi, gapA* and *groEL*) and *Enterobacter* (*gapA*) because of difficulties in obtaining full-length sequences for these genera. In these instances, initial trees were made with all sequences trimmed down to the length of the shortest sequence. These short sequences were then extended to full length by the addition of N's, and another tree was built with the longer sequences in order to increase the phylogenetic signal enough to resolve the ancestral relationships of the more closely related taxa. In all instances, extending the short sequences to full length with N's did not alter the tree topology with respect to these taxa.

Sequences for *gapA, groEL, gyrA* and *pgi* were aligned without gaps. The *ompA* sequence contained gaps in two variable regions. The first area corresponds to surface exposed loop L4 and is centered at amino acid residue 112 (Pautsch & Schulz, 2000). The second, smaller variable region corresponds to amino acid residues 175–180, which encodes the hinge region of the protein. The 16s sequence alignments had two single residue gaps corresponding to bases 79 and 88 in the *E. coli* 16s sequence.

Composite tree

The composite tree was generated by concatenating the same sequence alignments used to infer all six gene trees. Taxa that were missing from any one alignment were deleted from the other alignments. This resulted in a final data set of 4203 aligned nucleotides from 24 taxa. As in the generation of gene trees, a composite tree was generated from a shortened alignment to insure that missing sequence from *Hafnia* and *Serratia* isolates did not impact the composite tree topology.

Average distance and nucleotide diversity

The average uncorrected pairwise distance for each gene was calculated by averaging all of the pairwise comparisons between isolates of two different species using the PAUP program (Swofford, 1997) (Table 3). Nucleotide diversity, number of polymorphic sites, number of haplotpyes, and haplotype diversity was calculated for each taxa and gene using the DnaSP program version 3.53 (Rozas & Rozas, 1999) and the results are summarized on Table 4.

Table 3 Average percentage pairwise distances for each gene.

		CF*	EB	EC	HA	KO	KP	SP
16s	CF	0.55						
	EB	1.11	0.00					
	EC	3.37	3.06	0.85				
	HA	6.89	6.22	8.05	0.00			
	KO	0.97	1.38	3.65	7.60	0.00		
	KP	2.54	1.50	4.56	4.95	2.83	0.23	
	SP	3.17	2.08	5.79	4.81	3.46	2.19	NA
	VC	10.68	11.42	13.60	13.75	10.72	10.84	11.68
gapA	CF	0.41						
	EB	12.31	0.41					
	EC	7.23	13.18	0.36				
	HA	15.72	15.93	16.36	0.30			
	KO	5.31	10.79	7.64	17.09	0.05		
	KP	8.27	10.71	8.64	15.11	6.05	1.04	
	SP	14.60	14.45	15.18	18.41	13.49	10.55	0.25
	VC	21.71	20.09	20.67	18.67	22.02	20.51	22.94
groEL	CF	2.46						
	EB	9.42	1.64					
	EC	6.15	8.46	1.33				
	HA	10.76	12.02	10.46	0.61			
	KO	7.98	8.11	7.53	10.76	0.75		
	KP	10.41	8.78	9.17	11.85	7.57	3.48	
	SP	11.77	12.57	11.11	10.98	12.74	11.54	0.00
	VC	28.90	28.39	29.00	28.76	28.54	29.20	26.19
gyrA	CF	1.60						
	EB	9.13	0.11					
	EC	10.05	8.65	1.28				
	HA	17.14	17.53	18.38	0.20			
	KO	8.53	12.03	11.41	18.24	0.33		
	KP	11.75	12.27	11.22	19.13	11.45	1.45	
	SP	16.85	17.12	17.10	15.54	17.09	16.80	0.00
	VC	21.72	23.14	23.86	21.55	21.94	23.14	20.58
ompA	CF	1.74						
	EB	13.99	0.69					
	EC	13.88	16.08	1.96				
	HA	24.32	25.62	23.95	1.77			
	KO	17.96	14.95	16.33	28.91	0.45		
	KP	18.15	14.35	15.93	27.98	6.32	5.01	
	SP	26.84	24.33	22.30	21.41	22.45	23.14	0.27
	VC	52.11	52.73	52.58	53.65	53.11	52.61	54.31
pgi	CF	4.37						
	EB	9.66	0.15					
	EC	9.53	12.50	2.00				
	HA	23.07	25.20	25.46	0.49			
	KO	10.82	11.35	11.69	24.88	0.17		
	KP	11.11	11.60	11.20	24.14	9.71	4.43	
	SP	20.13	21.13	23.15	27.68	21.72	21.74	0.00
	VC	28.22	29.50	27.91	26.26	28.54	28.96	30.02

*Taxa names are CF, *Citrobacter freundii*; EB, *Enterobacter cloacae*; EC, *Escherichia coli*; HA, *Hafnia alvei*; KO, *Klebsiella oxytoca*; KP, *Klebsiella pneumoniae*; SP, *Serratia plymuthica* and VC, *Vibrio cholera*.

Results

Gene trees

Figure 1 provides phylogenetic trees inferred by maximum likelihood methods from the nucleotide sequence of each of five housekeeping genes (*gapA, groEL, gyrA, ompA*, and *pgi*) and the 16s rRNA gene. The gene lengths and length of sequence determined for each gene included in this analysis are given in Table 2. Because it was our intention from the outset to construct a composite tree by concatenating the sequences from

Table 4 Nucleotide diversity and number of polymorphic sites.

Gene	Species*	Sequences	Nucleotides	Polymorphic sites	Nucleotide diversity
16s	CF	5	289	3	0.00554
	EB	4	289	0	0
	EC	7	291	6	0.00851
	HA	5	291	0	0
	KO	5	289	0	0
	KP	6	289	2	0.00231
	SM	1	291	NA	NA
gapA	CF	5	831	6	0.00409
	EB	3	660	4	0.00404
	EC	7	660	9	0.0039
	HA	4	660	4	0.00303
	KO	5	832	1	0.00048
	KP	6	832	19	0.01042
	SM	3	540	2	0.00247
groEL	CF	5	1095	45	0.02447
	EB	5	1146	42	0.01626
	EC	7	1146	41	0.01334
	HA	4	1050	11	0.00635
	KO	5	1128	17	0.00762
	KP	4	1081	67	0.03562
	SM	4	444	0	0
gyrA	CF	5	738	20	0.01599
	EB	5	738	2	0.00108
	EC	7	738	21	0.01278
	HA	4	726	3	0.00207
	KO	5	738	5	0.00325
	KP	6	738	22	0.01454
	SM	4	643	0	0
ompA	CF	5	472	14	0.01737
	EB	5	475	7	0.00695
	EC	7	472	24	0.01947
	HA	5	487	19	0.01766
	KO	5	487	5	0.00452
	KP	5	487	57	0.0501
	SM	3	487	2	0.00274
pgi	CF	4	360	24	0.04167
	EB	4	641	2	0.00156
	EC	7	550	35	0.02061
	HA	4	405	4	0.00494
	KO	4	670	2	0.00181
	KP	4	670	52	0.04428
	SM	3	305	0	0

*Species abbreviations are CF, *Citrobacter freundii*; EB, *Enterobacter cloacae*; EC, *Escherichia coli*; HA, *Hafnia alvei*; KO, *Klebsiella oxytoca*; KP, *Klebsiella pneumoniae* and SP, *Serratia plymuthica*. NA: Not applicable.

each housekeeping gene, an effort was made to collect the same amount of phylogenetic information for each gene, so that the composite tree would be a true average of the phylogenetic information from each gene. Phylogenetic information density was estimated by determining the number of parsimony informative bases (Table 2). Although the sequences employed for the five housekeeping genes differ in length, they have very similar levels of phylogenetic information and fall within 12% of the average number of parsimony informative positions averaged over all sequences.

With only a few exceptions (*C. freundii* in the 16s tree and *K. pneumoniae* in the *gapA* and *ompA* trees), all of the species form monophyletic groups in all of the gene trees. In other words, there is a tight clustering of isolates within each species. For example, all isolates identified as *K. oxytoca* through phenotypic methods cluster together genetically as well. This is true even of *E. coli* K-12 strain MG1655, which was included to determine if a laboratory *E. coli* K-12 strain would cluster with natural isolates of *E. coli*.

In contrast, the relationship between species as seen in different trees is not consistent, that is, the precise species

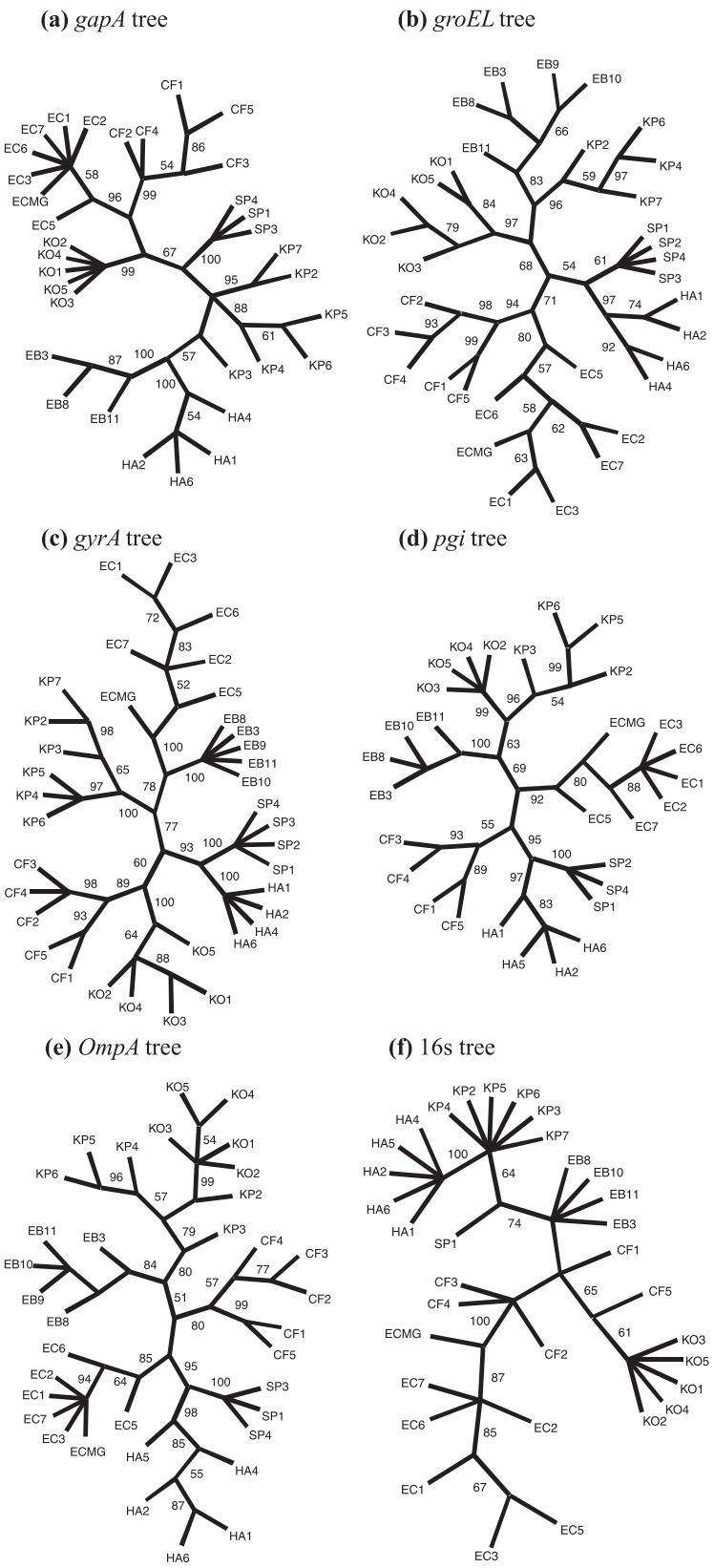

Fig. 1 Cladograms based on maximum likelihood for six putative core genes. (a) *gapA*, (b) *groEL*, (c) *gyrA*, (d) *pgi*, (e) *ompA*, and (f) 16s rRNA. Bootstrap values less than 50% (of 500 replicates) were omitted. Taxa abbreviations are CF: *Citrobacter freundii*; EB: *Enterobacter cloacae*; EC: *Escherichia coli*; HA: *Hafnia alvei*; KO: *Klebsiella oxytoca*; KP: *Klebsiella pneumoniae* and SP: *Serratia plymuthica*. ECMG refers to *E. coli* strain MG1655. Isolate numbers following taxa abbreviations refer to strains described in Table 1.

relationships differ from gene to gene (Fig. 1). Species that appear to be close relatives in one gene tree (for instance *E. coli* and *C. freundii* for *groEL*) fall on quite divergent branches of another tree (in this case *gyrA*). In fact, no two species pairs maintain the same relationship across all gene trees. However, some trends are apparent. For instance, more often than not *Hafnia* and *Serratia* are each other's closest relatives, and are located closest to the root of the tree (rooted trees not shown). *E. coli* and *C. freundii* fall nearest to each other in four out of the six gene trees.

Composite tree

A molecular species phylogeny was inferred from the composite data by concatenating the sequences of the six genes (Fig. 2). Figure 2 provides both an un-rooted tree (Fig. 2a), that is in the same format as that given for the individual gene trees (Fig. 1) and a version of the same tree rooted with *V. cholera* as the out group (Fig. 2b). The composite tree, in close agreement with the gene trees, clusters all of the isolates for each species into monophyletic groups. Even the outliers noted above for the *ompA*, *gapA* and 16s gene trees now fall within the species to which they were originally assigned based upon phenotypic analysis. The concatenated sequence contains enough phylogenetic signal to resolve all of the interspecies nodes with high bootstrap values (>77%).

The concatenated sequence was also used to build a phylogeny using Bayesian analysis. The MrBayes program (Huelsenbeck & Ronquist, 2001) was used to generate a majority rule consensus tree, which is analogous to a bootstrapped maximum likelihood tree (Fig. 2b). The two consensus trees have essentially the same topology. The maximum likelihood tree has a node connecting HA2 and HA3 with a bootstrap support of 51% that is not supported by the MrBayes tree, and the MrBayes tree contains a node connecting EC3 and ECMG, which is unsupported in the maximum likelihood tree. The Bayesian tree has a consistently higher support for most branch points. The Bayesian method is significantly faster than maximum likelihood bootstrap analysis. For example, it required approximately 600, 1GHz processor hours to calculate 500 bootstrap replicates using maximum likelihood, whereas it took approximately three, 1Ghz processor hours to generate the 45 000 trees used to construct the MrBayes tree.

An inherent danger in constructing a phylogeny from concatenated sequences is that the sequences can contribute unequal quantities of phylogenetic information to the composite tree. The phylogenetic information obtained for each housekeeping gene was estimated by calculating the number of parsimony informative bases contained in the sequence (Table 2). The contribution of each housekeeping gene was normalized such that each gene contributed approximately the same number of informative sites, to within 12% of the mean number of informative sites contributed per sequence.

Nucleotide diversity and distance

Table 3 contains estimates of the average within and between species sequence distance for each of the genes examined. Table 4 provides estimates of the within species nucleotide diversity. Estimates of nucleotide distance for this sample of taxa and genes ranges from 0.05 to 0.29 for the housekeeping genes, and 0.01–0.08 for the 16s rRNA gene. Estimates of nucleotide diversity

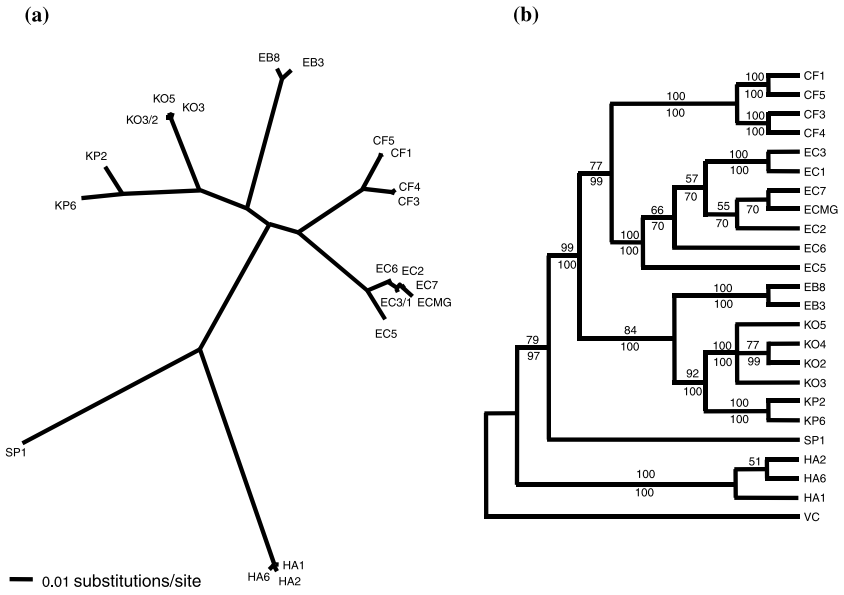

Fig. 2 Composite enteric phylogeny based on maximum likelihood methods. (a) Unrooted phylogram in which branch lengths indicate relative inferred evolutionary distance between isolates. (b) Cladogram of the composite phylogeny rooted using *Vibrio cholerae* as the outgroup. Numbers above branch points indicate Maximum likelihood bootstrap support (of 500 replicates), numbers below branch points indicate the frequency with which they occurred in a majority rule consensus tree made using the MrBayes program. Taxa abbreviations are as in Fig. 1.

ranges from zero to 0.05. *K. pneumoniae, C. freundii* and *E. cloacae* possessed the highest overall within species diversity, with *K. pneumoniae* showing the highest nucleotide diversity in all genes except *gyrA* and 16s rRNA. Overall *S. plymuthica* and *K. oxytoca* have the lowest within species nucleotide diversity. In the case of *S. plymuthica*, the low levels of diversity could be attributed to the smaller sample size and shorter sequences available for this taxon. For all species and all genes examined, the average within species pairwise distance or nucleotide diversity was lower than the average between species pairwise distance. In fact, for each gene, even the most diverse taxon (usually *K. pneumoniae*) had a lower within species distance than any between species comparison for the same gene.

C. freundii has a relatively high level of nucleotide diversity. Upon closer inspection, it is clear that this diversity is due primarily to the existence of two distinct lineages within this taxon that have a deep separation. Isolates CF1 and CF5 compose one distinct lineage whereas CF3 and CF4 compose the other (Fig. 2b). The fact that the division of these lineages based on phylogenic criteria correlates exactly with collection information regarding the host species and geography (Table 1) suggests that these factors are selectively important, although the small sample size prevents any statistical test for significance of this correlation.

Discussion

One thing is immediately obvious from examination of the six gene trees. The species relationships differ for the different genes. In other words, the seven taxa result in different tree topologies with each of the six genes examined. Species that appear to be close relatives in one gene tree [for instance *E. coli* and *C. freundii* for *groEL* (Fig. 1b)] fall on divergent branches of another tree [in this case *gyrA* (Fig. 1c)]. In fact, no two species pairs maintain the same relationship across all gene trees. However, some trends are apparent. For instance, more often than not *H. alvei* and *S. plymuthica* are each other's closest relatives and *E. coli* and *C. freundii* cluster nearest to each other in four of the six gene trees. It is clear that a robust molecular species phylogeny of enteric bacteria cannot be inferred from a single gene.

There are several possible explanations for the lack of consensus between individual gene trees with respect to species relationships. It could be the result of different genes experiencing different evolutionary pressures. For instance, gaps in the sequence alignment suggest that the *ompA* gene has experienced relatively strong selective pressure, presumably from bacteriophage such as K3 and Ox2, which have been shown to bind to the surface exposed loops of OmpA (Manning *et al.*, 1976; Morona *et al.*, 1984). Alternatively, the gene trees may differ because LGT has resulted in different phylogenetic histories for the different genes. None of the genes examined seem *a priori* to be the product of LGT. If incongruence in the gene trees is indeed a product of an ancient or difficult to detect LGT, then it appears to have gone to fixation.

In part, the discrepancies in the gene trees may reflect a low signal to noise ratio. The phylogenetic 'signal' in the sequences is too weak to infer a robust gene genealogy. Indeed, some of the branching patterns for particular gene trees could not be robustly resolved as shown by the low bootstrap values (Fig. 1). Concatenating the sequences of the six genes increases the number of phylogenetically informative characters and thus resolving previously unresolved branches. This increased resolution suggests that the incongruences between gene trees is more likely because of tree reconstruction artifacts resulting from weak phylogenetic signal rather than from LGT. As more sequences are concatenated, the underlying common phylogenetic signal is reinforced as demonstrated by the increase in bootstrap values. Other studies that have relied on concatenated housekeeping sequences have obtained phylogenies consistent with the fossil record or rRNA-based trees (Slade *et al.*, 1994; Brown *et al.*, 2001). This ability to recover phylogenies congruent with other methods was possible even when there were discrepancies in the individual gene tree topologies.(Eernisse and Kluge, 1993; Brochier et al., 2002).

This is the first molecular phylogeny of enteric species constructed from multiple isolates within each taxa. The molecular species phylogeny described here is in concordance with a previously described molecular phylogeny of enteric isolates where they share common taxa (Lawrence *et al.*, 1991). That prior phylogeny was constructed from type strains and clinical isolates and contained only a single isolate per taxa for the species common to both trees. Concordance between the two trees is not unexpected considering that the earlier tree was made with a subset of the housekeeping genes used in this study.

The consensus phylogeny (Fig. 2) and the average pairwise distance and nucleotide diversity estimates (Tables 2 and 3) clearly demonstrate that for this sample of taxa, the distances between species are always higher, and usually much higher, than the levels of diversity segregating within species. The elevated levels of distance observed between, relative to within, taxa suggests a possible mechanism limiting the exchange of genetic information between taxa. The frequency of homologous recombination has been shown to decrease exponentially as sequence divergence increases (Vulic *et al.*, 1997). This log-linear relationship is demonstrated by the observation that isolates of *E. coli* and *S. typhimurium*, which have 16% divergent genomes suffer a 10^5-fold reduction in recombination frequency compared with isogenic crosses (Vulic *et al.*, 1997). This observed difference in within and between species pairwise distances would therefore translate directly to a greater barrier to recombination between species compared with within species.

Implications for a bacterial species concept

There exists no official prokaryotic species definition. The defacto definition involves grouping strains based on overall phenotypic similarity, and then using DNA-DNA hybridization or 16s rRNA based phylogenies to discern the species boundaries (Cohan, 2001; Rossello-Mora & Amann, 2001). In an effort to provide bacterial classification with a theoretical underpinning, Dykhuizen & Green (1991) proposed that a derivation of the BSC (Mayr, 1942) could be employed. Because of the profound differences between eukaryotes and prokaryotes with regard to reproduction, application of the BSC to bacteria requires that the definition of an interbreeding group be revised to include groups participating in LGT, and that barriers to reproduction be considered barriers to LGT. Dykhuizen and Green suggested that groups that freely exchange genetic information can be identified phylogenetically because different genes from isolates of the same species should have significantly different phylogenies because of LGT and subsequent recombination, but phylogenies of genes from isolates of different species should be essentially the same because of interspecies barriers to recombination (Dykhuizen & Green, 1991). Although this was a huge step forward in applying a species concept to bacteria, as Maynard Smith (Maynard Smith, 1995) pointed out it has some practical shortcomings because of the large variation in the levels of clonality detected for different bacterial species. This would lead to a very narrow species definition for highly clonal organisms such as *Salmonella*, and broad species definitions for organisms like *Neisseria*, which have low levels of clonality (Maynard Smith *et al.*, 1993).

Lan and Reeves (Lan & Reeves, 2000, 2001) have proposed a revised method for applying the BSC to bacteria, independent of their level of clonality. They propose a 'Core Genome Hypothesis' (CGH) in which all of the genes found in a species are classified as either 'core genes', (a.k.a. housekeeping genes), which are found in the majority of the members of a species, or as 'auxiliary genes', which are only found in some members of the species. They argue that because there is little or no selective advantage in acquiring core genes from other species, the core genes would tend to diverge between species. This sequence divergence acts as an increasing barrier to horizontal transfer compared with that experienced by auxiliary genes. In contrast to core genes, auxiliary genes are more likely to be selectively retained after transfer because they often encode niche adaptive phenotypic traits that can confer a selective advantage under suitable environments. The frequent transfer and selection of these auxiliary genes reduces their rate of divergence between species, relative to those levels experienced by the core genes. Lan and Reeves thus propose that the species-specific shared, core genes, which comprise the shared species genome, be used to define species boundaries.

The consensus molecular phylogeny for enteric bacteria provided here (Fig. 2) suggests that, for this sample of enteric taxa, the CGH may provide an appropriate division of enteric species. The critical observation is that for this set of five housekeeping genes and the 16s rRNA gene all isolates from within each taxa cluster in genotype space, relative to isolates from any other taxa. In no case do individual sequences from within a taxa fall within the genotypic boundary of a different taxa. Further, this genotypic clustering corresponds precisely to the phenotypic clustering traditionally used to designate enteric species (Holt, 1994).

The CGH can be used to make predictions about auxiliary genes, which it defines as those genes that occur in <95% of the isolates of a species. It predicts that auxiliary genes should be more freely exchanged across species boundaries than core genes. The CGH also predicts that different species will have a different compliment of core genes, and a gene that is part of the core genome of one species may be an auxiliary gene of another species. By mapping phenotypic characters that represent likely auxiliary genes to the enteric molecular phylogeny, it is possible to determine if the patterns of phenotype occurrence are more consistent with vertical or lateral inheritance. Figure 3 shows ten phenotypic characters commonly used to distinguish the taxa used in this study, mapped onto the enteric composite molecular phylogeny. The phenotype character distribution patterns fall into two groups, those that

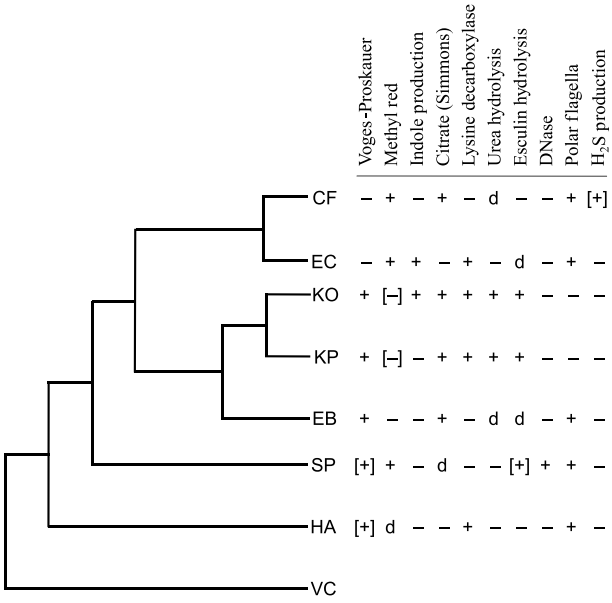

Fig. 3 Phenotypic characters commonly used to differentiate enteric species mapped onto the enteric molecular phylogeny. Taxa abbreviations are as in Fig. 1. Symbols: –, 0–10% positive; [–], 11–25% positive; d, 26–75% positive; [+], 76-89% positive; +, 90–100% positive.

could be the result of a single acquisition (such as H_2S or DNase production) or loss (such as polar flagella) vs. those that require multiple acquisition or loss events. Applying the principle of parsimony we suggest that the latter traits were acquired through horizontal transfer. For instance, in the case of indole production, five independent gene loss events, but only two acquisitions are required to generate the phenotype distribution pattern observed. There are also clear instances where a gene is most likely part of the core genome of one species, but an auxiliary gene in others. For example, citrate utilization is known to occur in almost all isolates of *C. freundii*, with moderate frequency in isolates *S. plymuthica*, and very rarely in *E. coli*.

The taxa included in this study make it particularly useful for exploring a bacterial species concept. By virtue of the collection method employed by Gordon (Gordon & FitzGibbon, 1999), it is known that isolates in this study coexist in mammalian intestines and, thus, the potential for lateral exchange is significant. Several studies suggest that LGT is not only possible, but also frequent among these taxa. Recent work on bacteriocin encoding plasmids isolated from different taxa from this same collection have been shown to be recent chimeras, which are composed of plasmid sequences from multiple enteric genera (Riley *et al.*, 2001; Smajs & Weinstock, 2001). These data indicate that conjugation between these taxa in nature is not uncommon. Estimates of the time since divergence between *E. coli* and *Salmonella typhimurium* range from 120 to 140 million years ago (Ochman & Wilson, 1987). Those two taxa are more closely related than is any pair of taxa in this study. Thus, it is fair to assume that these genera have had ample opportunity for LGT and recombination among the housekeeping genes sampled here. However, we find no evidence for extensive LGT in the sample of core genes examined.

Lan and Reeves have suggested subtractive hybridization and micro array technology as methods to define the core and auxiliary species genomes (Lan & Reeves, 2000). Although these techniques will provide valuable data, vital in understanding the process of genome evolution, they are currently too onerous and expensive to be useful in bacterial taxonomy. We suggest that phylogenies based on composite housekeeping gene sequences can act as an accurate proxy for the species-specific, shared, core genome information required by the CGH. This study shows that from a practical perspective, it is not necessary to determine definitively the scope and boundaries of these two components (core vs. auxiliary genes) of the species genome in order to implement the CGH as a useful concept for prokaryotic taxonomy. We have demonstrated that a robust phylogeny can be constructed from a relatively modest number of housekeeping genes. Enough is known about bacterial physiology that for most groups, assignment of half a dozen or so highly conserved housekeeping genes to the core genome can be done without controversy. As the CGH defines core genes as those that occur in at least 95% of the isolates of a species, using a defined subsample of housekeeping genes to determine species boundaries also prevents the definition of core genes from becoming circular. Care should be taken when selecting presumptive core genes for phylogenetic analysis so that over sampling of a chromosomal region, or any one physiological process does not occur.

Numerous bacterial species concepts have been proposed (Dykhuizen & Green, 1991; Holt, 1994; Lan & Reeves, 2000; Cohan, 2001; Rossello-Mora & Amann, 2001). The current gold standard has recently been renamed the phylo-phenetic species concept (PPSC), which is based upon numerical analysis of independently co-varying phenotypic characters (Rossello-Mora & Amann, 2001). The PPSC defines a bacterial species as '*a monophyletic and genomically coherent cluster of individual organisms that show a high degree of overall similarity with respect to many independent characteristics, and is diagnosable by a discriminative phenotypic property*'. Guidelines to apply this species definition include isolating an adequate collection of strains to account for the within species phenotypic variability, employing 16S rRNA to distinguish the closest relatives of a taxa in question, and extensive characterization of the phenotype. The PPSC proposes the use of 16S rRNA sequence similarity to determine evolutionary relationships among the taxa in question. However, although this molecule has proven invaluable in phylogenetic reconstructions of deep relationships among microbes (Woese, 1987), it is too slowly evolving to provide useful phylogenetic information for closely related bacteria. Further, Guanine+Cytosine content and levels of DNA–DNA similarity are characters proposed for determining a genomic measure of monophyly. Given our expanding sense of the extensive, and highly variable, levels of gene transfer experienced by bacteria, such measures of genome similarity may be quite variable, even for different isolates within the same taxa. For example, it is known that isolates of *E. coli* can vary by over 20% in genome size (Bergthorsson & Ochman, 1998).

This revised method for applying the BSC to bacteria reconciles the fluid mosaic structure found in genomic data with the clustering exhibited in decades of phenotypic studies. The question is not 'does lateral gene transfer occur?' but rather 'does its occurrence obliterate coevolved genomes?' Our ability to construct a robust species phylogeny from a sampling of core genes supports the existence of coevolved genomes that survive through evolutionary time. It is our belief that the role of a species concept in bacterial taxonomy is not to replace the traditionally employed, primarily phenetic method of classification with a purely phylogenetic one. After all, an organism's phenotype defines its interaction with the environment, and therefore its ecological significance. Rather, the role of a biologically based species concept in prokaryotic classification should be twofold; to provide an evolutionary framework for the organization of

higher taxonomic categories, and to provide a theoretical basis to explore the process of speciation. The validity of virtually partitioning the species genome into core and auxiliary genomes requires population level genomic data to test properly. It will be interesting to see how well this distinction holds up as the necessary data become available.

Acknowledgments

We acknowledge support to MAR from the NIH and NSF (GM58433 and DEB9459247). We also acknowledge the technical assistance of Derek Smith.

Supplementary material

The following material is available from: http://www.blackwellpublishing.com/products/journals/suppmat/JEB/JEB612/JEB612sm.htm

Primer sequences used for the amplification of gene fragments. All sequences are listed in a 5′–3′ orientation.

References

Bergthorsson, U. & Ochman, H. 1998. Distribution of chromosome length variation in natural isolates of *Escherichia coli*. *Mol. Biol. Evol.* **15**: 6–16.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao, Y. 1997. The complete genome sequence of *Escherichia coli* K–12. *Science* **277**: 1453–1462.

Brochier, C., Bapteste, E., Moreira, D. & Philippe, H. 2002. Eubacterial phylogeny based on translational apparatus proteins. *Trends Genet.* **18**: 1–5.

Brown, J.R. & Doolittle, W.F. 1997. Archaea and the Prokaryote-to-Eukaryote transition. *Microbiol. Mol. Bio. Rev.* **61**: 456–502.

Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. & Stanhope, M.J. 2001. Universal trees based on large combined protein sequence data sets. *Nature Genet* **28**: 281–285.

Cohan, F. 1996. The role of genetic exchange in bacterial evolution. *ASM News* **62**: 631–636.

Cohan, F.M. 2001. Bacterial species and speciation. *Sys. Biol.* **50**: 513–524.

Dobrindt, U. & Reidl, J. 2000. Pathogenicity islands and phage conversion: evolutionary aspects of bacterial pathogenesis. *Int. J. Med. Microbiol.* **290**: 519–527.

Doolittle, W.F. 1999. Lateral genomics. *Trends Biochem. Sci.* **24**: M5–M8.

DuBose, R.F., Dykuizen, D.E. & Hartl, D.L. 1988. Genetic exchange among natural isolates of bacteria: Recombination within the *phoA* gene of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* **85**: 7036–7040.

Dykhuizen, D.E. & Green, L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J. Bacteriol.* **173**: 7257–7268.

Eernisse, D.J. & Kluge, A.G. 1993. Taxonomic congruence versus total evidence, and Amniote phylogeny inferred from fossils, molecules, and morphology. *Mol. Biol. Evol.* **10**: 1170–1195.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N. & Woese, C.R. 1980. The phylogeny of prokaryotes. *Science* **209**: 457–463.

Gordon, D.M. & FitzGibbon, F. 1999. The distribution of enteric bacteria from Australian mammals: host and geographical effects. *Microbiology.* **145**: 2663–2671.

Gordon, D.M. & Lee, J. 1999. The genetic structure of enteric bacteria from Australian mammals. *Microbiology.* **145**: 2673–2682.

Heidelberg, J.F., Eisen, J.A., Nelson, W.C., Clayton, R.A., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, J.D., Umayam, L., Gill, S.R., Nelson, K.E., Read, T.D., Tettelin, H., Richardson, D., Ermolaeva, M.D., Vamathevan, J., Bass, S., Qin, H.Y., Dragoi, I., Sellers, P., McDonald, L., Utterback, T., Fleishmann, R.D., Nierman, W.C., White, O., Salzberg, S.L., Smith, H.O., Colwell, R.R., Mekalanos, J.J., Venter, J.C. & Fraser, C.M. 2000. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**: 477–483.

Holt, J.G. 1994. *Bergey's Manual of Determinative Bacteriology*. Williams and Wilkins, Baltimore, MD.

Huelsenbeck, J.P. & Ronquist, F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Karlin, S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. *Trends Microbiol.* **9**: 335–343.

Kidwell, M.G. 1993. Lateral transfer in natural populations of eukaryotes. *Annu. Rev. Genet.* **27**: 235–256.

Lan, R.T. & Reeves, P.R. 2000. Intraspecies variation in bacterial genomes: the need for a species genome concept. *Trends Microbiol.* **8**: 396–401.

Lan, R. & Reeves, P.R. 2001. When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol.* **9**: 419–424.

Lawrence, J.G., Ochman, H. & Hartl, D.L. 1991. Molecular and evolutionary relationships among enteric bacteria. *J. Gen. Microbiol.* **137**: 1911–1921.

Ludwig, W., Neumaier, J., Klugbauer, N., Brockmann, E., Roller, C., Jilg, S., Reetz, K., Schachtner, I., Ludvigsen, A., Bachleitner, M., Fischer, U. & Schleifer, K.H. 1993. Phylogenetic-relationships of bacteria based on comparitive sequence-analysis of elongation-factor tu and ATP-synthase Beta-subunit genes. *Antonie Van Leewenhoek* **64**: 285–305.

McKane, M. & Milkman, R. 1995. Transduction, restriction and recombination patterns in *Escherichia coli*. *Genetics* **139**: 35–43.

Manning, P.A., Puspurs, A. & Reeves, P. 1976. Outer membrane of *Escheriichia coli* K12: isolation of mutants with altered protein 3A by using host range mutants of bacteriophage K3. *J. Bacteriol.* **127**: 1080–1084.

Maynard Smith, J. 1995. Do bacteria have population genetics? In: *Population Genetics of Bacteria* (S. Baumberg ed), Cambridge University Press, Cambridge.

Maynard Smith, J., Smith, N.H., O'Rourke, M. & Spratt, B. 1993. How clonal are bacteria? *Proc. Natl. Acad. Sci. USA* **90**: 4384–4388.

Mayr, E. 1942. *Systematics and the Origin of Species*. Columbia University Press, New York.

Milkman, R. 1997. Recombination and population structure in *Escherichia coli*. *Genetics* **146**: 745–750.

Morona, R., Klose, M. & Henning, U. 1984. *Escherichia coli* K-12 Outer Membrane Protein (OmpA) as a Bacteriophage Receptor: Analysis of Mutant Genes Expressing Altered Protein. *J. Bacteriol.* **159**: 570–578.

Nelson, K. & Selander, R.K. 1994. Intergeneric transfer and recombination of the 6-phosphogluconate dehydrogenase gene (*gnd*) in enteric bacteria. *Proc. Natl. Acad. Sci. USA* **91**: 10227–10231.

Nesbo, C.L., L'Haridon, S., Stetter, K.O. & Doolittle, W.F. 2001. Phylogenetic analyses of two 'Archaeal' genes in *Thermotoga maritima* reveal multiple transfers between Archaea and bacteria. *Mol. Biol. Evol.* **18**: 362–375.

Ochman, H. & Wilson, A.C. 1987. Evolutionary history of enteric bacteria. In: *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology*, vol. 2. (F. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter & H. E. Umbarger eds), pp. 1615–1649. American Society for Microbiology, Washngton, DC.

Okada, S. & Gordon, D.M. 2001. Host and geographical factors influence the thermal niche of enteric bacteria isolated from native Australian mammals. *Mol. Ecol.* **10**: 2499–2513.

Olsen, G.J. & Woese, C.J. 1993. Ribosomal RNA: a key to phylogeny. *FASEB J.* **7**: 113–123.

Pace, N.R., Stahl, D.A., Lane, D.J. & Olsen, G.J. 1985. Analyzing natural microbial populations by rRNA sequences. *ASM News* **51**: 4–12.

Pautsch, A. & Schulz, G.E. 2000. High-resolution structure of the OmpA membrane domain. *J. Mol. Biol.* **298**: 273–282.

Posada, D. & Crandall, K.A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Riley, M.A., Pinou, T., Wertz, J.E., Tan, Y. & Valletta, C.M. 2001. Molecular characterization of the Klebicin B plasmid of *Klebsiella pneumoniae*. *Plasmid* **45**: 209–221.

Rossello-Mora, R. & Amann, R. 2001. The species concept for prokaryotes. *FEMS Microbiol. Rev.* **25**: 39–67.

Rozas, J. & Rozas, R. 1999. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**: 174–175.

Sherley, M., Gordon, D.M. & Collignon, P.J. 2000. Variations in antibiotic resistance profile in Enterobacteriaceae isolated from wild Australian mammals. *Environ. Microbiol.* **2**: 620–631.

Slade, R.W., Moritz, C. & Heideman, A. 1994. Multiple nuclear-gene phylogenies: Application to pinnipeds and comparisons with a mitochondrial DNA gene phylogeny. *Mol. Biol. Evol.* **11**: 341–356.

Smajs, D. & Weinstock, G.M. 2001. Genetic organization of plasmid ColJs, encoding colicin Js activity, immunity, and release genes. *J. Bacteriol.* **183**: 3949–3957.

Stoltzfus, A., Leslie, J.F. & Milkman, R. 1988. Molecular evolution of the *Escherichia coli* chromosome. I. Analysis of structure and natural variation in a previously uncharacterized region between *trp* and *tonB*. *Genetics* **120**: 345–358.

Swofford, D. 1997. *PAUP*: Phylogenetic Analysis Using Parsimony (* and other methods)*. Sinauer Associates, Sunderland, MA.

Thompson, J., Higgins, D. & Gibson, T. 1994. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4580.

Vulic, M., Dionisio, F., Taddei, F. & Radman, M. 1997. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc. Natl. Acad. Sci. USA* **94**: 9763–9767.

White, P.A., McIver, C.J. & Rawlinson, W.D. 2001. Integrons and gene cassettes in the Enterobacteriaceae. *Antimicrob. Agents Chemother.* **45**: 2658–2661.

Woese, C.R. 1987. Bacterial evolution. *Microbiol. Rev.* **51**: 221–271.

Woese, C.R., Kandler, O. & Wheelis, M.L. 1990. Towards a natural system of organisms - proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. USA* **87**: 4576–4579.

Received 12 December 2002; revised 21 May 2003; accepted 10 July 2003