Escherichia coli Molecular Phylogeny Using the Incongruence Length Difference Test

Guillaume Lecointre, Latif Rachdi,† Pierre Darlu,‡ and Erick Denamur†*

*Service de Systématique Moléculaire, Muséum National d'Histoire Naturelle, Paris, France; †INSERM V458, Hôpital Robert Debré, Paris, France; and ‡INSERM V155, Université Denis Diderot, Paris, France

Molecular phylogeny of the species *Escherichia coli* using the *E. coli* reference (ECOR) collection strains has been hampered by (1) the absence of rooting in the commonly used phenogram obtained from multilocus enzyme electrophoresis (MLEE) data and (2) the existence of recombination events between strains that scramble phylogenetic trees reconstructed from the nucleotide sequences of genes. We attempted to determine the phylogeny for *E. coli* based on the ECOR strain data by extracting from GenBank the nucleotide sequences of 11 chromosomal structural and 2 plasmid genes for which the *Salmonella enterica* homologous gene sequences were available. For each of the 13 DNA data sets studied, incongruence with a nonnucleotide whole-genome data set including MLEE, random amplified polymorphic DNA, and *rrn* restriction fragment length polymorphism data was measured using the incongruence length difference (ILD) test of Farris et al. As previously reported, the incongruence observed between the *gnd* and plasmid gene data and the whole-genome data was multiple, indicating numerous horizontal transfer and/or recombination events. In five cases, the incongruence detected by the ILD test was punctual, and the donor group was identified. Congruence was not rejected for the remaining data sets. The strains responsible for incongruences with the whole-genome data set were removed, leading to a ''prior-agreement'' approach, i.e., the determination of a phylogeny for *E. coli* based on several genes, excluding (1) the genes with multiple incongruences with the whole genome data, (2) the strains responsible for punctual incongruences, and (3) the genes incongruent with each other. The obtained phylogeny shows that the most basal group of *E. coli* strains is the B2 group rather than the A group, as generally thought. The D group then emerges as the sister group of the rest. Finally, the A and B1 groups are sister groups. Interestingly, the most primitive taxon within *E. coli* in terms of branching pattern, i.e., the B2 group, includes highly virulent extraintestinal strains with derived characters (extraintestinal virulence determinants) occurring on its own branch.

Introduction

Escherichia coli is one of the most studied bacteria. It has been used as a model system in the development of molecular biology, and the complete *E. coli* nucleotide sequence is available. The *E. coli* reference (ECOR) collection of 72 strains from diverse natural origins is thought to represent the genetic diversity of the species (Ochman and Selander 1984). It has made it possible to generate substantial amounts of comparative data for many strains of the species. Multilocus enzyme electrophoresis (MLEE) was initially used to produce a topology for the ECOR strains (Herzer et al. 1990) which was consistent with total DNA-based analyses such as random amplified polymorphic DNA (RAPD) or *rrn* restriction fragment length polymorphism (RFLP) (Desjardins et al. 1995). Four main groups (A, B1, B2, and D) and ungrouped strains (UG) were defined in the ECOR collection (Herzer et al. 1990). However, the topology obtained from these data could not be interpreted as a phylogeny, because there was no outgroup to root the tree. The level of recombination and its effect on population structure in *E. coli* are still a matter of debate (see Guttman [1997] and Milkman [1997] for reviews). So far, analysis of individual genes or clusters of genes have shown that recombination events are rare except in two regions, around the O antigen complex and the *hsd* locus, where high levels of polymorphism confer a selective advantage (Milkman 1997). Phylogenetic analysis of these data is possible, because an outgroup is often available (e.g., the homologous genes in the species *Salmonella enterica*). However, horizontal gene transfer disturbs the strain phylogenetic signal, and the branching patterns obtained are valid for the gene but cannot be extrapolated to the strains themselves.

We attempted to determine the strain phylogeny within the *E. coli* species based on the ECOR strain data by extracting from GenBank the sequences of several *E. coli* genes for which homologous genes from *S. enterica* were available. The assumption of this work was that gene transfer between strains occuring by recombination should lead to statistically significant incongruence between data sets and, more precisely, between an individual gene DNA data set and a whole-genome data set. The whole-genome data set includes MLEE (Herzer et al. 1990), RAPD, and *rrn* RFLP (Desjardins et al. 1995) data and is representative of the bulk of *E. coli* genome polymorphism. Incongruence between molecular trees has long been recognized to result not only from tree reconstruction artifacts, but also from truly different gene histories. This interpretation is particularly relevant to organisms such as viruses or bacteria, in which horizontal transfer and recombination events are often cited as the causes of incongruence (Dykhuizen and Green 1991; de Queiroz 1993; de Queiroz, Donoghue, and Kim 1995; hypothesis III of Gogarten, Hilario, and Olendzenski 1996; Maddison 1997). Specific methods for detecting such transfers have been developed (Lawrence and Hartl 1992). General methods (Templeton 1983; Farris et al. 1995; Huelsenbeck, Bull, and Cunningham

Key words: *Escherichia coli,* ECOR collection, incongruence length difference test, extra intestinal virulence, molecular phylogeny.

Address for correspondence and reprints: Erick Denamur, INSERM U 458, Hôpital Robert Debré, 48, boulevard Sérurier, 75935 Paris cedex 19, France. E-mail: denamur@infobiogen.fr.

Mol. Biol. Evol. 15(12):1685–1695. 1998

1996; Huelsenbeck and Bull 1996) for detecting incongruence between ‘‘process partitions’’ (Bull et al. 1993) can also be used to detect significant incongruence between bacterial molecular phylogenies (Dykhuizen and Green 1991) and to identify the participants in recombination events. For each of the 13 DNA data sets studied here, incongruence with the whole-genome data set was measured using the incongruence length difference (ILD) test of Farris et al. (1995). This test was used because it was developed in the parsimony framework and it is simple and efficient (Cunningham 1997). Then, for each data set significantly incongruent with the whole-genome data set, iterative removal of strains followed in each case by a new ILD test was used to identify the strain(s) responsible for incongruence, possibly due to a recombination event. The donor group was identified by visual inspection of aligned sequences. In some cases, there were multiple incongruences affecting numerous strains. The removal of strains responsible for incongruence from the data set led to a ‘‘prior-agreement’’ approach, i.e., a phylogeny for *E. coli* based on several genes, excluding genes having multiple incongruences with the whole-genome data, strains responsible for punctual incongruences, and genes incongruent with each other. This gives a gene phylogeny as close as possible to a strain phylogeny of *E. coli* (clonal framework). This approach identified the B2 group rather than the A group (Herzer et al. 1990) as the most basal group of *E. coli* strains.

Materials and Methods

Selection of Nucleotide Sequence Data

Nucleotide sequences were extracted from GenBank. Only genes sequenced for a sufficient number of ECOR strains and for *S. enterica* were selected (table 1). Eleven genes were of chromosomal origin and two were plasmid genes. The 11 chromosomal genes were structural genes scattered throughout the genome, 510–1,722 nt in size. They were: *putP* (1,467 bp at 23.25 min) (Nelson and Selander 1992), *icd* (1,164 bp at 25.74 min) (Wang, Whittam, and Selander 1997), *trpA* (807 bp at 28.33 min), *trpB* (1,194 bp at 28.35 min), *trpC* (1,359 bp at 28.37 min) (Milkman and McKane 1995), *gapA* (882 bp at 40.10 min) (Nelson, Whittam, and Selander 1991), *pabB* (1,009 bp at 40.79 min) (Guttman and Dykhuizen 1994*a*), *gnd* (1,335 bp at 45.21 min) (Bisercic, Feutrier, and Reeves 1991, Nelson and Selander 1994), *crr* (510 bp at 54.61 min) (Hall and Sharp 1992), *mdh* (864 bp at 72.86 min) (Boyd et al. 1994), and *aceK* (1,722 bp at 90.86) (Nelson et al. 1997). Two fertility factor F-related plasmid genes known to be frequently exchanged between strains (Boyd et al. 1996) were also studied to check the response of the ILD test. These two genes were the *finO* gene (441 bp), the fertility inhibition gene, and *traD* (523 bp), a transfer gene.

Sequence Analyses

For each of the 13 data sets (table 1), sequences were aligned by eye using the ED program of the MUST package (Philippe 1993). Sequence saturation in a given data set was tested using MUST and PAUP 3.1.1. (Swofford 1993) by plotting the pairwise number of observed differences (y axis) against the pairwise number of inferred substitutions met in the pathway joining the two taxa in the most parsimonious (MP) tree (x axis), as recovered by MUST from the phylogram saved from PAUP (Vidal and Lecointre 1998). From each data set, the MP tree was obtained from heuristic, branch-and-bound, or exhaustive search of PAUP 3.1.1, depending on the number of taxa, and boostrapping (Felsenstein 1985) was performed with PAUP using 1,000 iterations.

To investigate incongruence between each DNA data set and the whole-genome data set, the ILD test was performed using both the XARN software (Farris et al. 1995) and the test version 4.0.0d64 of PAUP* written by David L. Swofford. The nonnucleotide whole-genome data set is large, containing data for 72 reference ECOR strains (Ochman and Selander 1984) and 320 characters, including MLEE, RAPD, and *rrn* RFLP binary coded characters (Herzer et al. 1990; Desjardins et al. 1995). Phylogenetic reconstructions using parsimony generate for the whole-genome data set a tree with a topology identical to that for MLEE (Herzer et al. 1990) (data not shown). For each of the 13 incongruence tests, the taxonomic sample of the whole-genome data was reduced and made identical to each of the 13 DNA data sets to make combination possible. DNA and whole-genome data were combined using MERGE software (unpublished data). The ILD test for incongruence tests the null hypothesis of congruence between data sets (Mickevich and Farris 1981; Farris et al. 1995). Parsimony analyses are carried out for data set x and data set y separately. Then, the lengths of each MP tree obtained are added ($L_x + L_y$), and this length is compared to the sum of the lengths ($L_p + L_q$) of the MP trees obtained from two data sets, p and q, of the same size as the original data sets and generated by random partitioning of the original couple of data sets. For W random partitions, S is the number of times when ($L_x + L_y$) < ($L_p + L_q$), and the null hypothesis of congruence is rejected when the error rate ‘‘alpha,’’ also called P-value in PAUP*, ($1 - S/(W + 1)$), is below a particular threshold, i.e., 5%. This indicates that there is more incongruence between the data sets than would be expected from chance alone (for a more detailed description of the test, including the complete expression of the ILD, see Farris et al. 1995). In our analysis, we generated 1,000 random partitions for each test and the value considered in table 2 is alpha = X/1,000, as given in the XARN manual. In this case, there is incongruence when alpha < 50/1,000. In the course of this work, the alpha value appeared to be sensitive to the order of taxa in the matrix. Consequently, for each gene, 100 tests were performed from 100 different taxon orders, and the mean alpha value was considered (table 2). The same ILD test was also performed through the ‘‘partition homogeneity test’’ of PAUP*. P-values of PAUP* were not sensitive to the order of taxa and were very close to mean alpha values from XARN (table 2). When a given DNA data set appeared to be incongruent with the whole-genome data (alpha value or P-value < 50/1,000),

Table 1
Sequences Available from GenBank of the Strains of the ECOR Collection and *Salmonella enterica* for 11 Chromosomal Genes (*putP* to *aceK*) and 2 Plasmid Genes (*finO* and *traD*)

Strain	*putP*	*icd*	*trpA*	*trpB*	*trpC*	*gapA*	*pabB*	*gnd*	*crr*	*mdh*	*aceK*	*finO*	*traD*
A													
ECOR 1 . . .			+	+	+				+				
ECOR 5 . . .								+					
ECOR 8 . . .			+	+	+	+	+						
ECOR 10 . .	+	+				+	+	+		+	+		
ECOR 11 . .								+				+	+
ECOR 25 . .								+				+	
ECOR 9 . . .												+	+
ECOR 4 . . .			+	+	+	+	+	+					
ECOR 6 . . .								+	+				
ECOR 16 . .			+	+	+	+	+	+					
ECOR 14 . .	+	+				+		+		+	+		
ECOR 18 . .								+					
ECOR 19 . .			+	+	+								
ECOR 20 . .								+					
ECOR 21 . .			+	+	+			+					
ECOR 17 . .		+	+	+	+						+		
ECOR 24 . .			+	+	+								
ECOR 15 . .		+	+	+	+			+					
ECOR 23 . .								+					
B1													
ECOR 58 . .	+	+				+		+	+	+	+	+	+
ECOR 27 . .			+	+	+							+	+
ECOR 69 . .		+	+	+	+			+	+		+		
ECOR 28 . .			+	+	+				+				
ECOR 29 . .			+	+	+								
ECOR 32 . .	+	+				+		+		+	+		
ECOR 30 . .												+	+
ECOR 68 . .			+	+	+	+	+	+					
ECOR 70 . .	+	+	+	+	+	+		+		+	+		+
ECOR 71 . .			+	+	+			+	+			+	+
ECOR 72 . .			+	+	+								
B2													
ECOR 51 . .			+	+	+				+				
ECOR 52 . .	+	+	+	+	+	+		+		+	+		
ECOR 54 . .			+	+	+								
ECOR 56 . .			+	+	+			+					
ECOR 57 . .								+					
ECOR 65 . .						+	+	+					
ECOR 61 . .									+				
ECOR 62 . .												+	+
ECOR 63 . .								+					
ECOR 64 . .	+	+				+		+		+	+		
ECOR 59 . .													+
ECOR 60 . .			+	+	+								
ECOR 66 . .									+				
D													
ECOR 35 . .								+	+	+		+	+
ECOR 38 . .						+	+	+		+			
ECOR 39 . .						+	+					+	
ECOR 40 . .	+	+				+	+	+		+	+		
ECOR 46 . .			+	+	+			+		+			
ECOR 49 . .						+	+	+		+		+	+
ECOR 50 . .			+	+	+	+	+	+	+	+			
ECOR 44 . .								+		+			
ECOR 47 . .								+		+		+	+
ECOR 48 . .												+	
UG													
ECOR 31 . .			+	+	+								
ECOR 43 . .								+					
ECOR 37 . .		+	+	+	+				+		+	+	
ECOR 42 . .												+	+
S. enterica . . .	+	+	+	+	+	+	+	+	+	+	+	+	+

NOTE.—The order of the ECOR strains is as in the phenogram of Herzer et al. (1990). The order of the chromosomal genes reflects their localization in the genetic map. The four groups A, B1, B2, and D and the ungrouped strains (UG) are as described by Herzer et al. (1990).

Table 2
Characteristics of the Phylogenetic Data from 11 Chromosomal and 2 Plasmid Genes

Gene	No. of Nucleo-tides	No. of In-form-ative Sites	No. of Trees	CI	RI	Length of the MP Tree	Most External Group	Mean Alpha Value	P-value	Incongruent Taxa	Identification of the Donor Group	P-value Without the Identified Incongruent Taxa
putP . . .	1,467	62	1/1	0.796/0.876	0.744/0.841	113/339	B2	33/1,000	42/1,000	ECOR 70	A	1,000/1,000
icd . . .	1,164	39	1/5	0.680/0.816	0.677/0.615	100/244	B2	7/1,000	2/1,000	ECOR 52	D	196/1,000
										or		
										ECOR 40	B2	117/1,000
trpA . . .	807	50	16/3	0.804/0.891	0.932/0.888	92/285	B2 + D + UG	711/1,000	470/1,000			
trpB . . .	1,194	71	2/2	0.732/0.880	0.878/0.839	142/317	B2	898/1,000	880/1,000			
trpC . . .	1,359	71	5/5	0.720/0.857	0.844/0.809	164/434	B2	7/1,000	10/1,000	ECOR 4, 17, 19, 21, 24	B1	190/1,000
gapA . . .	882	8	1/3	1/0.985	1/0.947	11/65	D	590/1,000	980/1,000			
pabB . . .	1,009	42	1/5	0.927/0.966	0.962/0.898	55/358	UR	5/1,000	6/1,000	ECOR 68	A	364/1,000
										or		
										ECOR 4	B1	347/1,000
gnd . . .	1,335	327	1/2	0.544/0.519	0.661/0.638	1,051/1,180	—	1/1,000	1/1,000	Multiple		
crr . . .	510	11	2/6	0.947/0.907	0.957/0.828	19/54	B2	410/1,000	767/1,000			
mdh . . .	864	21	1/56	0.811/0.930	0.885/0.803	37/158	UR	300/1,000	230/1,000			
aceK . . .	1,722	97	4/4	0.810/0.840	0.789/0.678	200/482	UG	52/1,000	2/1,000	ECOR 10	B2 or D or UG	961/1,000
finO . . .	441	25	3/3	0.698/0.697	0.787/0.800	63/66	—	1/1,000	1/1,000	Multiple		
traD . . .	523	29	3/3	0.740/0.675	0.642/0.597	73/83	—	1/1,000	1/1,000	Multiple		

NOTE.—For number of trees, consistency index (CI), retention index (RI), and length of the MP tree, the first number corresponds to the number without the outgroup, whereas the second number corresponds to the number with the outgroup. The alpha value (Farris et al. 1995) and the P-value (test version 4.0.0d64 of PAUP* written by David L. Swofford) given by the ILD test is calculated between the gene and the whole-genome data sets without outgroup. The mean alpha value corresponds to 100 values obtained by taxon shuffling (100 data sets differing in the order of taxa). UR = unresolved. The four groups A, B1, B2, and D and the ungrouped strains (UG) are as described by Herzer et al. (1990).

the MP tree obtained from DNA sequences was examined by eye to determine the nature of the incongruence. Two situations were encountered in this work: (1) multiple incongruences, with the DNA tree being completely scrambled with respect to the monophyly of the strain groups (A, B1, D, and B2), and (2) incongruence in the DNA tree being punctual with respect to the monophyly of these main strain groups. In the last case, the removal of each strain separately followed by a new PAUP* run made it possible to identify the strain, the removal of which resulted in an increase in the P-value to >50/1,000. The sequences responsible for statistically significant incongruence (presumably resulting from a recombination) were then identified. Visual inspections of the MP trees and aligned sequences allowed identification of the donor group. Thus, genes can be combined for phylogenetic analysis excluding those strains identified as being responsible for incongruences. The number of possible combinations was very small because there were very few common strains for which the sequence was known for each gene involved. Once the combinations were defined, ILD tests were performed with all the DNA data sets to be combined, in order to check their mutual congruence. A ''prior-agreement'' approach was then used. This term was proposed by Chippindale and Wiens (1994) to describe the approach of Bull et al. (1993) for the same purpose but in a different technical context. This approach combined, for a single phylogenetic analysis, genes for which sequences were known for a sufficient number of common strains and that were mutually congruent, and it excluded sequences that may have been produced by recombination. Three gene combinations satisfied these conditions: *trpA* + *trpB* + *trpC; trpB* + *trpC* + *crr;* and *gapA* + *mdh* + *putP*.

Results and Discussion

Combination of the ILD Test and Visual Inspection of Trees Discriminates Punctual Incongruences from Multiple Incongruences and Identifies the Donor Group

Sequence alignments were easy and did not imply indels. For each data set, the variability of DNA sequences was low and no saturation was detected. Overall, the 13 data sets exhibited low homoplasy contents (high consistency indexes [CIs] and high retention indexes [RIs]) (table 2). Both plasmid genes had completely scrambled trees (fig. 1*A* and data not shown) with respect to the monophyly of each group (multiple incongruence) and very low alpha and P-values (table 2). No removal of single strains or groups of strains increased P-values. Thus, the ILD test responded adequately to data sets of sequences known to have been intensively exchanged, resulting in completely scrambled trees (Boyd et al., 1996).

Among the 11 metabolic genes studied, 6 showed significant incongruence with the whole-genome data set: *putP*, *icd*, *trpC*, *pabB*, *aceK*, and *gnd* (table 2). Considering each MP tree and bootstrappings, the first five data sets exhibited punctual incongruences with regard to the monophyly of the traditional groups A, B1, D, and B2, whereas *gnd* showed, as previously reported (Bisercic, Feutrier, and Reeves 1991; Nelson and Selander 1994), multiple incongruences completely mixing these groups with high bootstrap supports (fig. 1*B*). For *putP, icd, trpC, pabB*, and *aceK*, visual inspection of trees and sequence alignments completed by removals of individual strains and new ILD tests led to the identification of the strains responsible for the incongruences and the donor groups (table 2 and fig. 2). The donor group is identified as the group within which the strain is unexpectedly branched. This is based on the assumption that the unexpected branching point of the recipient is supported by synapomorphies shared exclusively by the donor group and the recipient. These synapomorphies must have been gained once from this donor. The extent of recombination—the length of the transferred stretch of DNA—is checked by visual inspection of aligned sequences.

For the *aceK* data set, only the removal of ECOR 10, which belongs to the A group, increased the P-value to 961/1,000 (fig. 2*A*). Although the position of the recipient strain, ECOR 10, within the MP tree does not allow us to discriminate the donor group among the D, B2, and UG groups, visual inspection of aligned sequences favors the UG origin (identification of a shared stretch of DNA between the recipient and the UG strain) (data not shown).

For *pabB* (fig. 2*B*), only the removal of each of the two strains incorrectly branched (ECOR 68 and 4) increased the P-value to clearly above 50/1,000. It is difficult to determine from our data which strain is the donor and which is the recipient, as ECOR 68 and 4 *pabB* sequences are identical, and no other B1 strain is available. However, Guttman and Dykhuizen (1994*a*, 1994*b*) argued on supplementary data from neighboring genes that the ECOR 4 strain probably gained a *pabB* gene for a B1 donor such as the ECOR 68 strain. Similar reasoning can be applied to the *icd* data as ECOR 40 D has the same *icd* sequence as ECOR 52 B2 (table 2 and data not shown).

The tree from *putP* (fig. 2*C*) was congruent with the whole-genome tree, except for the B1 strain, ECOR 70. Each of the three B1 strains, when individually removed, increased the P-value to 1,000/1,000, whereas removal of the other strains did not. These high P-values are due to the fact that removal of each B1 strain restores the monophyly of the two others. Indeed, ECOR 70 does share three uniquely derived characters with A strains and two with B1 strains. This explains why the removal of the ECOR 14 strain results in a P-value of 242/1,000. It relaxes the attraction between ECOR 70 and the A group, restoring the monophyly of the three B1 strains.

The tree based on the *trpC* gene showed unexpected paraphylies of groups A and B1 (fig. 2*D*). No removal of a single strain allowed recovery of a P-value above 50/1,000. The removal of various combinations of taxa chosen for restoring the monophyly of the A and B1 groups was tested. Among them, the smallest strain sample to be removed in order to obtain a P-value above

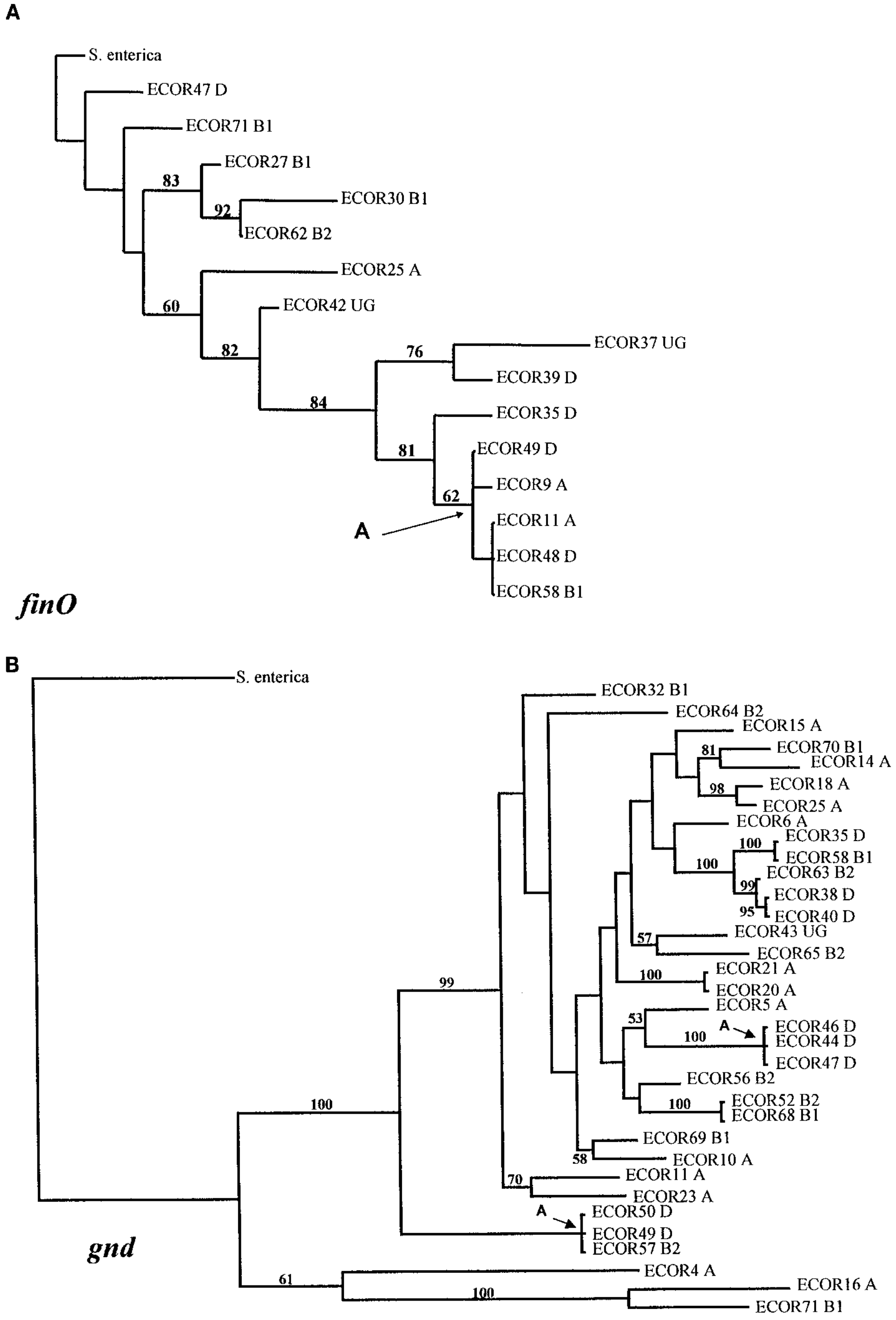

FIG. 1.—Examples of trees from plasmid (*A*) and chromosomal (*B*) DNA data sets showing multiple incongruences with the whole-genome data set. Numbers at branches are the bootstrap proportions obtained from 1,000 replicates. Only bootstrap proportions above 50% are given. Branch lengths are given under ACCTRAN optimization. The groups to which the ECOR strains belong, as defined by Herzer et al. (1990), are indicated after the number of the strain. *A,* One of the three most-parsimonious trees obtained from *finO* data. The two other most-parsimonious trees differ from this one only within clade A. *B,* One of the most-parsimonious trees obtained from *gnd* data. The other tree differs within the "A" nodes.

50/1,000 (0.190) was ECOR 4, 21, 19, 17, and 24. Visual inspection of aligned sequences allowed identification of a stretch of DNA in the 5′ end of *trpC* where ECOR 4, 21, 19, and 17 do share the same nucleotides as B1 strains at all variable positions (positions 357, 393, 432, 459, 477, 509, 516, and 531). It is clear that these ECOR strains gained a 5′ portion of *trpC* from a B1 donor, perhaps an ancestor shared by these four strains. Furthermore, ECOR 28 and 29 share the same nucleotides as the A strains in all the variable positions in the *trpC* 3′ end (position 1014, 1056, 1108, and 1167). This results in an unexpected branching in the MP tree (fig. 2*D*). Thus, ECOR 28 and 29 did gain a *trpC* 3′ end from A donors.

The tree for *gnd* was so scrambled (fig. 1*B*) that no removal of single strains or groups of strains gave a P-

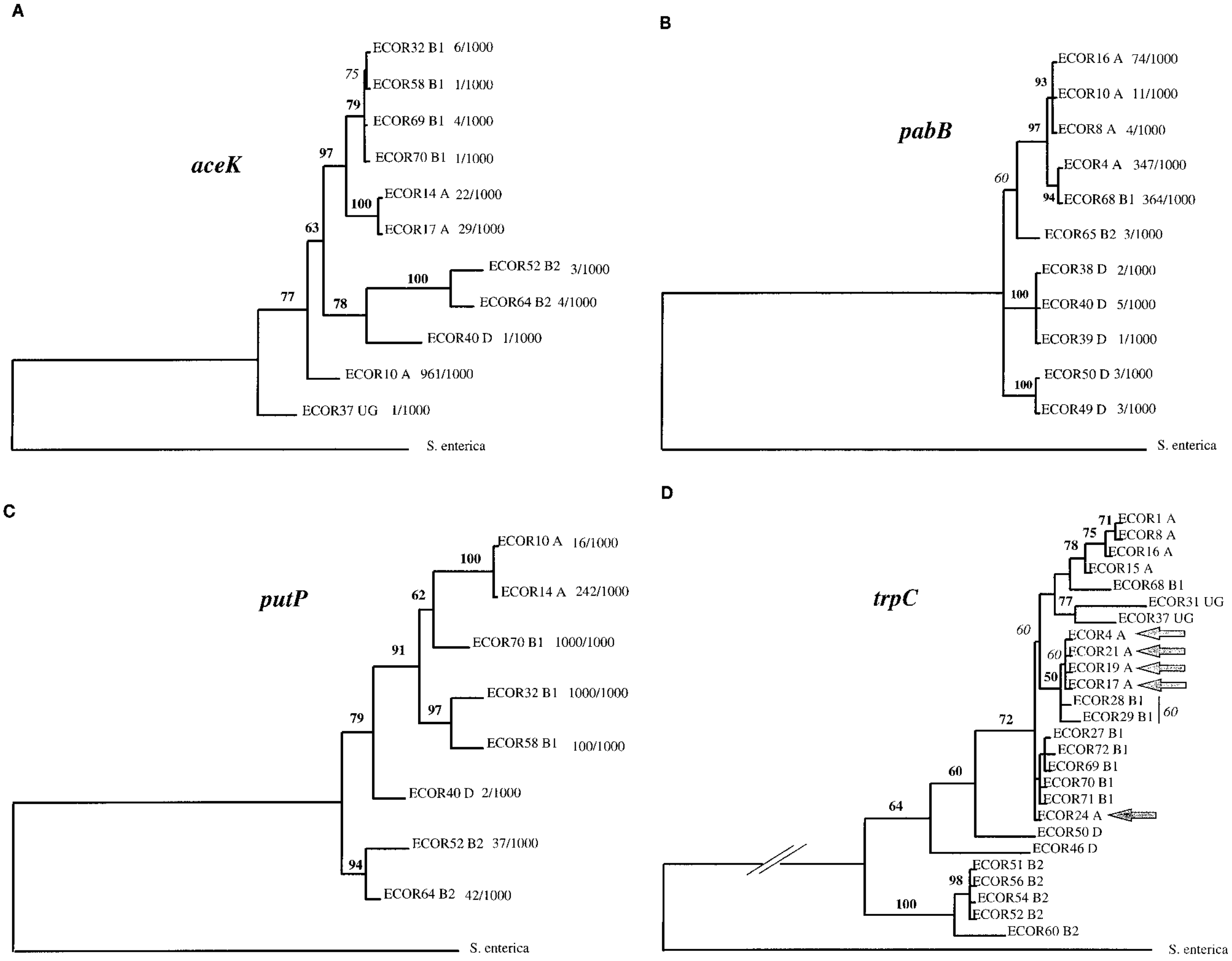

FIG. 2.—Examples of trees from chromosomal data sets showing punctual incongruences with the whole-genome data set. Numbers in bold are bootstrap proportions above 50%, whereas numbers in italics indicate the percentages of MP trees in which a node occurs, given only when this percentage does not reach 100%. Numbers on the right indicate the P-value from the ILD test obtained when the corresponding taxon is removed from the data set. Branch lengths are given under ACCTRAN optimization. *A,* Majority-rule consensus tree of four MP trees obtained from *aceK* sequences. *B,* Majority-rule consensus tree of five MP trees obtained from *pabB* sequences. *C,* The most-parsimonious tree obtained from the *putP* sequences. *D,* Majority-rule consensus tree of five MP trees obtained from *trpC* sequences. Arrows show the strains which are to be concomitantly removed to obtain a P-value above the 50/1,000 threshold when the ILD test is performed by PAUP* against the whole-genome data set.

value above 50/1,000. A minimum of 28 strains must be removed to recover a P-value above 50/1,000 (data not shown).

Our work is consistent with previous data showing that, with the exception of plasmid genes (Boyd et al. 1996), the *hsd* locus (Barcus, Titheradge, and Murray 1995), and the *gnd* locus close to the O antigen gene complex, which is subject to hitch-hiking (Bisercic, Feutrier, and Reeves 1991; Nelson and Selander 1994), recombination events do occur in *E. coli* but are rare (Milkman 1997).

Interpretation of Congruence

Low alpha or P-values (<50/1,000) were obtained when there was at least one node in the tree from the first data set contradicting another node in the tree from the second data set. Moreover, the two nodes must be well supported in their respective trees by synapomorphies. Each of the three data sets exhibiting multiple incongruences with the whole-genome data appears to be well structured in terms of branch length and bootstrap proportions (*gnd* and *finO,* fig. 1; *traD,* data not shown). Totally noisy data sets cannot yield significant incongruence with other data sets, but a single structured incongruence can be detected even when this contradictory signal is lost in noisy data. This was the case for *icd.* The *icd* MP tree was poorly resolved (data not shown), but the removal of ECOR 40 or 52 increased the alpha or P-value to above 50/1,000 (table 2). Thus, there is no relationship between alpha or P-values and global homoplasy measurements such as CIs and RIs.

A low alpha or P-value (<50/1,000) can be due to a single recombined strain (*putP, pabB*) or to multiple recombinations (*gnd,* plasmid genes). Therefore, there is no relationship between the alpha or P-value and the number of incongruences.

The interpretation of congruence is more ambiguous. The null hypothesis of congruence cannot be rejected if there is no signal in the data. Therefore, totally noisy data sets may appear to be as congruent with the whole-genome data as other more structured data sets. To determine whether a given data set is really congruent with the whole-genome data set, the alpha or P-value, resolution of the MP tree, and bootstrap proportions must all be considered. Very few informative sites for parsimony from *mdh* data (table 2) gave a tree that was so poorly resolved that this data set was not clearly incongruent with the whole genome data but was also not clearly congruent (data not shown). In contrast, congruent nodes are supported by high alpha or P-values and well-resolved trees for *trpA* (resolved tree except within the A group), *trpB* and *trpC* (resolved except within the A and B1 groups), *putP* (without ECOR 70), and *aceK* (without ECOR 10) genes.

Interpretation of congruence is also weakened by possible undetected recombinational events due to the insufficient lengths of transferred stretches of DNA. For example, the ILD test failed to detect the recombination in the 3′ end of ECOR 28 and 29 *trpC* (see above). When ECOR 28 and 29 were both removed, the P-value did not increase to above 50/1,000. In a similar way, aligned *trpC* sequences show a stretch of DNA (from position 350 to 550) in which all variable positions exhibit common nucleotides only shared by ECOR 68 B1 and A strains, possibly indicating that ECOR 68 is closed to A strains in the MP tree (fig. 1*D*). This shows the usefulness of additional tests that are able to detect recombinations of short stretches of DNA (Grassly and Holmes, 1997; Maynard Smith and Smith 1998).

When unique partial recombinational events occur, the recipient strain exhibits homoplastic sequences, i.e., positions that contradict each other; some group the recipient with its donor, others with its group of origin. These two homoplastic patterns are not randomly distributed along the sequence but are grouped in stretches. The interpretation of the ILD test can be ambiguous when the number of taxa is low and homoplastic patterns of the recipient are mixed and scattered along the sequences. For instance, in *putP* (fig. 2*C*), positions in which ECOR 70 shares a nucleotide with A strains are mixed with those in which ECOR 70 shares a nucleotide with B1 strains. These mixed homoplastic patterns could have been obtained through complex and ancient recombinational events, but they could be due to random homoplasy. Adding taxa would help to resolve this issue.

The B2 Group Strains Are the Most Basal Within the *E. coli* Tree

The MLEE tree has always been arbitrarily rooted on the A group (Herzer et al. 1990), leading to the conflicting evidence that the A group was the earliest to emerge in the *E. coli* tree. Here, the presence of an outgroup made it possible to find the following branching order from each metabolic gene once punctual incongruences identified above were removed: the first group to emerge was, with one exception, B2 (table 2), then came group D, the sister group of the A + B1 clade, in which A is the sister group of B1. The phylogenetic relationships and monophyly of the UG group are unclear. However, these phylogenies are gene, not strain, phylogenies.

To obtain strain phylogenies based on several genes, we used the prior-agreement principle. Without measuring incongruence, an *E. coli* tree based on several genes combined is unlikely to be reliable because of potential recombinations. The reason is that, in the case of horizontal gene transfers, evolutionary history of a particular recipient strain is found to be different from one gene to another. It does not make sense to combine genes with different histories to produce a single reconstruction, because, as stressed by Bull et al. (1993), combining the data not only obscures an important feature of history but runs the risk of producing a reconstruction that fails to represent either history. The removal of individual strains followed by ILD tests was used to identify the strains responsible for incongruence when these strains were not too numerous. This gave more reliable results than did previous studies in which authors suspected or evoked recombination events from different topologies. In such cases, the risk is that differences in trees, rather than being due to differences in gene histories, are due to different tree construction methods or different parameters being chosen for phylogenetic analyses performed by different authors. For example, the neighbor-joining tree published by Nelson et al. (1997) for *aceK* sequences differs in topology from the *aceK* MP tree described herein (fig. 2*A*). If such trees were interpreted in terms of recombination, the conclusions drawn would not be the same. Assuming that the congruence of characters is more important than the congruence of trees (Barret, Donoghue, and Sober 1991), our approach provides a valuable criterion for congruence, because it deals directly with the characters themselves rather than with trees, and trees are extracted from each data set using the same method, thus avoiding differences in trees due to the use of different tree reconstruction methods. It is unlikely that we detected all the recombination events, but we discarded the most obvious ones. Only three combinations were possible, mainly because there were too few strains common to all or part of the 10 metabolic gene data sets congruent with the whole-genome data set. We also checked that all members of each combination were congruent to each other using the ILD test (data not shown). In the trees obtained from the *trpA* + *trpB* + *trpC*, *trpB* + *trpC* + *crr*, and *gapA* + *mdh* + *putP* combinations (fig. 3 and data not shown), B2 strains are basal. Group D then emerges as the sister group of the rest. The A and B1 groups are sister groups. The A group is therefore not the most basal one.

Dating of *E. coli* Groups

Assuming a molecular clock for the gene combinations proposed above and calibrating molecular rates using a divergence time for *S. enterica* and *E. coli* of between 160 and 120 MYA as proposed by Ochman and Wilson (1987), the divergence time of the main *E. coli* strain groups can be estimated. Two combined data sets were used: *trpA* + *trpB* + *trpC* and *gapA* + *mdh* + *putP*. The *trpB* + *trpC* + *crr* combination was excluded

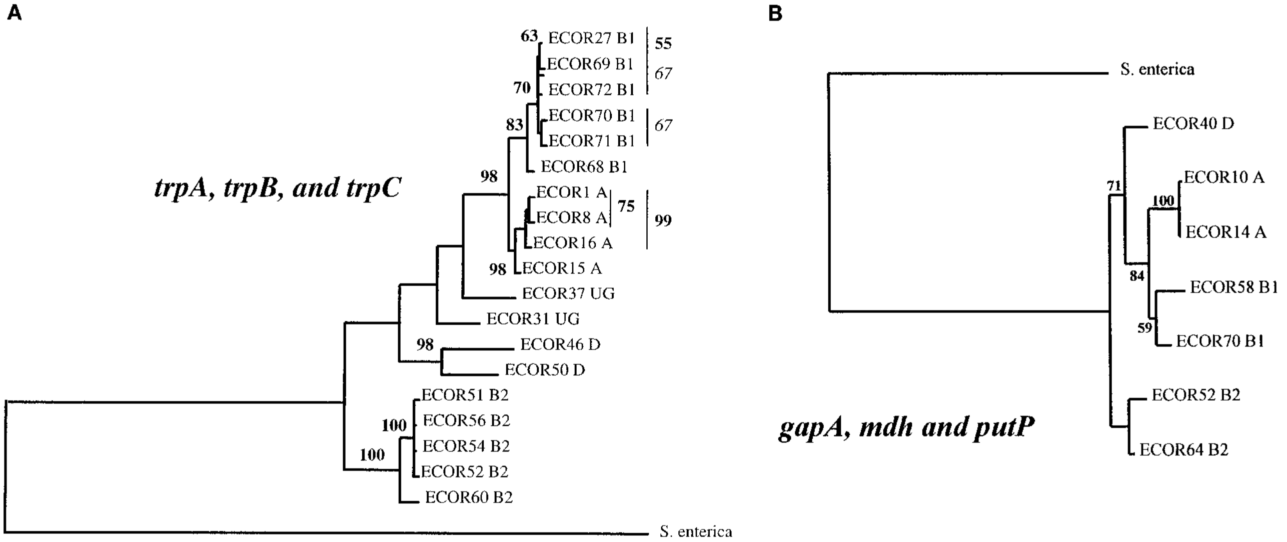

FIG. 3.—Trees obtained using the prior-agreement approach involving data from (*A*) *trpA, trpB,* and *trpC* genes, excluding strains in the *trpC* data set responsible for significant incongruence with the whole-genome data set, and from (*B*) *gapA, mdh,* and *putP* genes. Numbers in bold are bootstrap proportions above 50%, whereas numbers in italics refer to the proportions of MP trees in which the node occurs (only shown below 100%). Branch lengths are given under ACCTRAN optimization. *A,* Majority-rule consensus tree of three MP trees. The number of steps is 1,039, CI = 0.849, and RI = 0.817. *B,* The MP tree. The number of steps is 537, CI = 0.924, and RI = 0.672.

because of redundancy. Dating can be performed only if there is no mutational saturation in the data, especially between the sequences of *S. enterica* (the outgroup) and *E. coli* strains, and if there is no selective sweep (Guttman and Dykhuizen 1994*a*) within the species for one of these genes. No mutational saturation was detected between *S. enterica* and *E. coli* strains for any of the six genes. However, the ratio between interspecific and intraspecific molecular divergence for *gapA* was not the same as that for the other genes, suggesting that some relatively recent evolutionary event has purged most of the variability from the *E. coli gapA* gene (Guttman and Dykhuizen 1994*a*). Such a selective sweep leads to underestimation of divergence times between *E. coli* strains for the *gapA* + *mdh* + *putP* combination as compared with the *trpA* + *trpB* + *trpC* or *mdh* + *putP* combinations (data not shown). Assuming a divergence time of 120 MYA for *S. enterica* and *E. coli,* the divergence between the monophyletic group B2 and the rest was 22.4 MYA (calculated from *mdh* + *putP* data) to 23.2 MYA (*trpA* + *trpB* + *trpC* data). Assuming a divergence time of 160 MYA for *S. enterica* and *E. coli,* the divergence between the monophyletic group B2 and the rest was 25.8 MYA (*mdh* + *putP* data) to 30.8 MYA (*trpA* + *trpB* + *trpC* data).

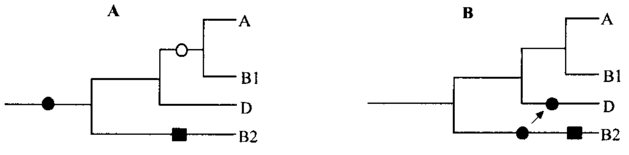

FIG. 4.—Evolution of virulence determinants within *E. coli* phylogenetic groups. The common topology linking *E. coli* groups is as shown by the gene data set studied in this work. Gains and losses of virulence determinants as described by Boyd and Hartl (1998) and Bingen et al. (1998) are mapped to show that there is a single most parsimonious option for *sfa* and *hly* virulence genes (a single gain in the B2 branch), whereas there are two equally parsimonious situations for the *kps* and *pap* virulence genes (*A* and *B*). ●, gain of *kps* and *pap* virulence genes; ○, loss of *kps* and *pap* virulence genes; ■, gain of *sfa* and *hly* virulence genes. The arrow indicates horizontal transfer.

Biological Significance

Escherichia coli is a commensal inhabitant of the gastrointestinal tract in humans and is one of the most frequently isolated pathogens (Berg 1996; Falkow 1996). Within the *E. coli* species, B2 group strains are highly pathogenic with numerous virulence determinants implicated in extraintestinal infections in nonimmunocompromised patients (Goullet and Picard 1986; Picard et al. 1991; Bingen et al. 1998; Boyd and Hartl 1998; unpublished data). Strains of phylogenetic group D seem to have fewer virulence determinants than B2 group strains, whereas the most closely related groups, A and B1, are most often devoid of extraintestinal virulence determinants (Bingen et al. 1998; Boyd and Hartl 1998; unpublished data). Thus, in the *E. coli* strain tree, the more basal the branching of a given group, the more virulent it is. This observation can be refined using the distribution of pathogenic determinants in the tree (Bingen et al. 1998; Boyd and Hartl 1998). Extraintestinal pathogenicity results from virulence determinants generally encoded by genes linked within pathogenicity islands (Pai's) (Hacker et al. 1997). These Pai's are mobile elements. Their G+C contents suggest that they have been acquired from another bacterial species. The *sfa* adhesin and *hly* hemolysin determinants are found almost exclusively in B2 strains (Bingen et al. 1998; Boyd and Hartl 1998). The most parsimonious scenario is that these two virulence determinants were acquired only in the B2 branch (one step), rather than being acquired via the common ancestor of all *E. coli* strains and lost later in the sister group of B2 (two steps) (fig. 4). The *kps* capsule and *pap* adhesin determinants are common to B2 and D strains (Bingen et al. 1998; Boyd and Hartl 1998), suggesting the parsimonious scenario that they were acquired once in the common ancestor of all strains and lost in the sister group of D (fig. 4*A*). Alternatively, the *kps* and *pap* determinants may have been acquired by D strains from B2 strains (fig. 4*B*) by horizontal transfer. This later scenario is somewhat favored by the widely held view of a commensal ancestor *E. coli* becoming pathogenic (Hacker et al. 1997), and by the trees published by Boyd and Hartl (1998) from *kpsD* and *papH* gene sequences. These trees are clearly not congruent with the MLEE tree (Herzer et al. 1990) or with our trees, suggesting multiple horizontal transfers (*kpsD*) or punctual horizontal transfers from B2 to D (*papH*). This may be an example of the earliest branch of a tree exhibiting derived characters (extraintestinal virulence determinants) occuring on its own branch. Phylogenetic patterns of living organisms sometimes show that ''primitive taxa,'' those branched off at the base of a tree because they exhibit many characters in the primitive state, can also have other characters in a very specialized state. For instance, the ornithorhynch is a member of the earliest branch of extant mammals, but is also one of the most specialized (derived) mammals, with, for example, its poisonous glands and beak. As the ornithorhynch, B2 strains are the most basal in the *E. coli* tree but are also very peculiar strains with respect to pathogenic determinants.

Acknowledgments

We thank Ivan Matic and François Taddei for useful discussions and Jacques Elion and Simon Tillier for constant encouragement. We are grateful to Kimberlyn Nelson for providing us with the *aceK* sequences.

LITERATURE CITED

BARCUS, V. A., A. J. B. TITHERADGE, and N. E. MURRAY. 1995. The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli.* Genetics **140**:1187–1197.

BARRETT, M., M. J. DONOGHUE, and E. SOBER. 1991. Against consensus. Syst. Zool. **40**:486–493.

BERG, R. D. 1996. The indigenous gastrointestinal microflora. Trends Microbiol. **4**:430–435.

BINGEN, E., B. PICARD, N. BRAHIMI, S. MATHY, P. DESJARDINS, J. ELION, and E. DENAMUR. 1998. Phylogenetic analysis of *Escherichia coli* strains causing neonatal meningitis suggests horizontal gene transfer from a predominant pool of highly virulent B2 group strains. J. Infect. Dis. **177**:642–650.

BISERCIC, M., J. Y. FEUTRIER, and P. R. REEVES. 1991. Nucleotide sequence of the *gnd* genes from nine natural isolates of *Escherichia coli*: evidence of intragenic recombination as a contributing factor in the evolution of the polymorphic *gnd* locus. J. Bacteriol. **173**:3894–3900.

BOYD, E. F., and D. L. HARTL. 1998. Chromosomal regions specific to pathogenic isolates of *Escherichia coli* have a phylogenetically clustered distribution. J. Bacteriol. **180**: 1159–1165.

BOYD, E. F., C. W. HILL, S. M. RICH, and D. L. HARTL. 1996. Mosaic structure of plasmids from natural populations of *Escherichia coli*. Genetics **143**:1091–1100.

BOYD, E. F., K. NELSON, F. S. WANG, T. S. WHITTAM, and R. K. SELANDER. 1994. Molecular genetic basis of allelic polymorphism in malate dehydrogenase (*mdh*) in natural populations of *Escherichia coli* and *Salmonella enterica*. Proc. Natl. Acad. Sci. USA **91**:1280–1284.

BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD, and P. J. WADELL. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. **42**:384–397.

CHIPPINDALE, P. T., and J. J. WIENS. 1994. Weighting, partitioning and combining characters in phylogenetic analysis. Syst. Biol. **43**:278–287.

CUNNINGHAM, C. W. 1997. Can three incongruence tests predict when data should be combined? Mol. Biol. Evol. **14**: 733–740.

DE QUEIROZ, A. D. 1993. For consensus (sometimes). Syst. Biol. **42**:368–372.

DE QUEIROZ, A. D., M. DONOGHUE, and J. KIM. 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. **26**:657–681.

DESJARDINS, P., B. PICARD, B. KALTENBÖCK, J. ELION, and E. DENAMUR. 1995. Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment length polymorphism. J. Mol. Evol. **41**:440–448.

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

FALKOW, S. 1996. The evolution of pathogenicity in *Escherichia coli, Shigella,* and *Salmonella.* Pp. 2723–2729 *in* F. C. NEIDHART, ed. *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.

FARRIS, J. S., M. KÄLLERSJÖ, A. G. KLUGE, and C. BULT. 1995. Testing significance of incongruence. Cladistics **10**:315–319.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution. **39**:783–791.

GOGARTEN, J. P., E. HILARIO, and L. OLENDZENSKI. 1996. Gene duplications and horizontal gene transfer during early evolution. Pp. 267–292 *in* R. D. MCROBERTS, P. SHARP, G. ALDERSON, and M. A. COLLING, eds. Evolution of microbial life. Cambridge University Press, Cambridge, England.

GOULLET, P., and B. PICARD. 1986. Highly pathogenic strains of *Escherichia coli* revealed by the electrophoretic patterns of carboxylesterase B. J. Gen. Microbiol. **132**:1853–1858.

GRASSLY, N. C., and E. C. HOLMES. 1997. A likelihood method for the detection of selection and recombination using nucleotide sequences. Mol. Biol. Evol. **14**:239–247.

GUTTMAN, D. S. 1997. Recombination and clonality in natural populations of *Escherichia coli.* Trends Ecol. Evol. **12**:16–22.

GUTTMAN, D. S., and D. E. DYKHUIZEN. 1994*a*. Detecting selective sweeps in naturally occuring *Escherichia coli.* Genetics **138**:993–1003.

———. 1994*b*. Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266**:1380–1383.

HACKER, J., G. BLUM-OEHLER, I. MÜHLDORFER, and H. TSCHÄPE. 1997. Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol. Microbiol. **23**:1089–1097.

HALL, B. G., and P. M. SHARP. 1992. Molecular population genetics of *Escherichia coli*: DNA sequence diversity at the *celC, crr,* and *gutB* loci of natural isolates. Mol. Biol. Evol. **9**:654–665.

HERZER, P. J., S. INOUYE, M. INOUYE, and T. S. WHITTMAN. 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli.* J. Bacteriol. **172**:6175–6181.

Huelsenbeck, J. P., and J. J. Bull. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. **45**: 92–98.

Huelsenbeck, J. P., J. J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. **11**:152–158.

Lawrence, J. G., and D. L. Hartl. 1992. Inference of horizontal genetic transfer from molecular data: an approach using the bootstrap. Genetics **131**:753–760.

Maddison, W. P. 1997. Gene trees in species trees. Syst. Biol. **46**:523–536.

Maynard Smith, J., and N. H. Smith. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. **15**:590–599.

Mickevich, M. F., and J. S. Farris. 1981. The implications of congruence in Menidia. Syst. Zool. **30**:351–370.

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

Milkman, R., and M. McKane. 1995. DNA sequence variation and recombination in *E. coli.* Pp. 127–142 *in* S. Baumberg, J. P. W. Young, E. M. H. Wellington, and J. R. Saunders, eds. Population genetics of bacteria. Cambridge University Press, Cambridge, England.

Nelson, K., T. S. Whittam, and R. K. Selander. 1991. Nucleotide polymorphism and evolution in the glyceraldehyde-3-phosphate dehydrogenase gene (*gapA*) in natural populations of *Salmonella* and *Escherichia coli.* Proc. Natl. Acad. Sci. USA **88**:6667–6671.

Nelson, K., and R. K. Selander. 1992. Evolutionary genetics of the proline permease gene (*pupP*) and the control region of the proline utilization operon in populations of *Salmonella* and *Escherichia coli*. J. Bacteriol. **174**:6886–6895.

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

Nelson, K., F. S. Wang, E. F. Boyd, and R. K. Selander. 1997. Size and sequence polymorphism in the isocitrate dehydrogenase kinase/phosphatase gene (*aceK*) and flanking regions in *Salmonella enterica* and *Escherichia coli.* Genetics **147**:1509–1520.

Ochman, H., and R. K. Selander. 1984. Standard reference strains of *Escherichia coli* from natural populations. J. Bacteriol. **157**:690–692.

Ochman, H., and A. C. Wilson. 1987. Evolution in bacteria: evidence for an universal substitution rate in cellular genomes. J. Mol. Evol. **26**:74–86.

Philippe, H. 1993. MUST: a computer package of management utilities for sequences and trees. Nucleic Acids Res. **21**: 5264–5272.

Picard, B., N. Picard-Pasquier, R. Krishnamoorthy, and P. Goullet. 1991. Characterization of highly virulent *Escherichia coli* strains by ribosomal DNA restriction fragment length polymorphism. FEMS Microbiol. Lett. **82**:183–188.

Swofford, D. L. 1993. Phylogenetic analysis using parsimony (PAUP). Version 3.1.1. Illinois Natural History Survey, Champaign.

Templeton, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the human and apes. Evolution **37**:221–244.

Vidal, N., and G. Lecointre. 1998. Weighting and congruence: a case study based on three mitochondrial genes in pitvipers. Mol. Phylogenet. Evol. **Special Volume**:1–9.

Wang, F. S., T. S. Whittam, and R. K. Selander. 1997. Evolutionary genetics of the isocitrate dehydrogenase gene (*icd*) in *Escherichia coli* and *Salmonella enterica.* J. Bacteriol. **179**:6551–6559.

Howard Ochman, reviewing editor

Accepted August 21, 1998