Available online at www.sciencedirect.com
SCIENCE DIRECT®
Molecular Phylogenetics and Evolution 30 (2004) 243–250
MOLECULAR PHYLOGENETICS AND EVOLUTION
www.elsevier.com/locate/ympev

Short communication

Decreasing the effects of horizontal gene transfer on bacterial phylogeny: the *Escherichia coli* case study

Patricia Escobar-Páramo,[a] Audrey Sabbagh,[b] Pierre Darlu,[b] Olivier Pradillon,[a] Christelle Vaury,[c] Erick Denamur,[a,*] and Guillaume Lecointre[d]

[a] *INSERM E0339, IFR 02, Faculté de Médecine Xavier Bichat, 16 rue Henri Huchard, Paris 75018, France*
[b] *INSERM U 535, Hôpital Paul Brousse, Villejuif, France*
[c] *Centre d'Etude du Polymorphisme Humain, Hôpital Saint Louis, Paris, France*
[d] *IFR CNRS 1541, Muséum National d'Histoire Naturelle, Paris, France*

Received 31 October 2002; revised 2 May 2003

Abstract

Phylogenetic reconstructions of bacterial species from DNA sequences are hampered by the existence of horizontal gene transfer. One possible way to overcome the confounding influence of such movement of genes is to identify and remove sequences which are responsible for significant character incongruence when compared to a reference dataset free of horizontal transfer (e.g., multilocus enzyme electrophoresis, restriction fragment length polymorphism, or random amplified polymorphic DNA) using the incongruence length difference (ILD) test of Farris et al. [Cladistics 10 (1995) 315]. As obtaining this "whole genome dataset" prior to the reconstruction of a phylogeny is clearly troublesome, we have tested alternative approaches allowing the release from such reference dataset, designed for a species with modest level of horizontal gene transfer, i.e., *Escherichia coli*. Eleven different genes available or sequenced in this work were studied in a set of 30 *E. coli* reference (ECOR) strains. Either using ILD to test incongruence between each gene against the all remaining (in this case 10) genes in order to remove sequences responsible for significant incongruence, or using just a simultaneous analysis without removals, gave robust phylogenies with slight topological differences. The use of the ILD test remains a suitable method for estimating the level of horizontal gene transfer in bacterial species. Supertrees also had suitable properties to extract the phylogeny of strains, because the way they summarize taxonomic congruence clearly limits the impact of individual gene transfers on the global topology. Furthermore, this work allowed a significant improvement of the accuracy of the phylogeny within *E. coli*.

Keywords: *Escherichia coli*; Horizontal gene transfer; Incongruence length difference test; Congruence; ECOR collection

1. Introduction

Horizontal gene transfer is a well-known phenomenon in bacterial evolution (Dykhuizen and Green, 1991; Ochman, 2001). This movement of genes among strains with different evolutionary histories produces taxonomic incongruence between phylogenies reconstructed from different genes, making difficult the reconstruction of the species phylogeny. One possible way to overcome this problem is to identify and remove from the analysis those gene sequences acquired through horizontal transfer. We have previously demonstrated the usefulness of this approach by using the incongruence length difference (ILD) test of Farris et al. (1995) in the reconstruction of the *Escherichia coli* phylogeny (Lecointre et al., 1998). Transfers were identified by testing character incongruence between the dataset from individual genes and a reference dataset from which horizontal transfers are assumed to have no effects on the strain phylogeny. This reference dataset was obtained from whole genome analysis encompassing multilocus enzyme electrophoresis (MLEE), restriction fragment length polymorphism, and random amplified polymorphic DNA data.

However, obtaining this "whole genome dataset" prior to the reconstruction of a phylogeny is clearly

* Corresponding author. Fax: +33-1-44-85-61-49.
E-mail address: denamur@bichat.inserm.fr (E. Denamur).

1055-7903/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S1055-7903(03)00181-7

troublesome. Therefore, the ILD test as performed in Lecointre et al., 1998 cannot be generalized as a method for phylogenetic reconstruction for any bacteria species and alternative methods that do not require the use of such a reference dataset are needed.

2. How to identify sequence data responsible for character incongruence without the whole genome dataset?

A possible strategy consists on applying the ILD test between each individual gene and the simultaneous analysis of the remaining genes $(n-1)$ of the dataset. Such strategy has been used to trace the source of incongruence in various molecular datasets as sequences from Hawaiian drosophilids (Baker and DeSalle, 1997), mammalian mitochondria (Schevchuk and Allard, 2001), and salmonellae (Brown et al., 2002). This protocol implicitly assumes that the set of the $n-1$ genes must globally reflect the phylogeny of strains, that the effects of single horizontal transfers are swamped into the signal for strain phylogeny and that the transfers affect different strains in different genes. Note that the set of $n-1$ genes changes in each ILD test as the "1" gene used for the analysis is different each time. Thus the possibility that the phylogenetic signal of a particular gene is imposed over the overall genes is avoided.

We studied the phylogenetic relationship of 30 ECOR strains representative of the *E. coli* genetic diversity (Ochman and Selander, 1984) using the DNA sequences of 11 housekeeping genes encompassing 7 metabolic (*trpA*, *trpB*, *thrB*, *putP*, *pabB*, *icd*, and *purM*) and 4 polymerase (*polB*, *dnaE*, *dinB*, and *umuC*) (Bjedov et al., 2003) genes. These genes have been chosen as they are distributed all over the *E. coli* chromosome and have a relatively high level of sequence polymorphism giving phylogenetic signals within *E. coli* (Table 1). Furthermore, these genes do not show any evidence for a selective advantage of diversity as it has been demonstrated for genes encoding (or near to genes encoding) membrane antigens, antibiotic resistance, restriction-modification, or mismatch repair systems (Denamur et al., 2000). *E. fergusonii*, the closest relative to *E. coli*, (Lawrence et al., 1991) was used as the outgroup to limit possible long branch attraction effects (Felsenstein, 1978). Some of the sequences were extracted from GenBank (compiled in Bjedov et al., 2003; Lecointre et al., 1998) and others were generated de novo for this study. PCR and sequencing conditions are as in Denamur et al. (2000) and primer sequences were taken from the literature (Bjedov et al., 2003; Escobar-Páramo et al., in press; Guttman and Dykhuizen, 1994; Pupo et al., 2000) except for *icd* (icd-F: 5′-GAA AgT AAA gTA gTT gTT CCg g-3′; icd-R: 5′-gAT gAT CgC gTC ACC AAA C/TTC-3′). The GenBank database accession numbers from the previously available sequences were for *trpA* and *trpB*: U23490, U23491, U23495–U23499, U25419, U25422, U25423, U25426, U25428, U25884, *thrB*: AF293252, AF293264, AF293266, *putP*: L01150, L01153, L01154, L01155, L01156, and L01158, *pabB*: U07749, U07755–U07757, U07760, and U07764, *icd*: AF017589, AF017592, AF017595, AF017596, AF017598, AF017601, AF017601–AF017603, *purM*: AF293161, AF293164, AF293166, *polB*: AF483912–AF483939, *dnaE*: AF483940–AF483969, *dinB*: AF483080–AF483106, and *umuC*: AF483970–AF483998. The new sequences have been deposited in the GenBank database under Accession Nos. AY132827–AY132995. Sequences were aligned using CLUSTAL (Higgins et al., 1992) analysis from the Sequence navigator package. A phylogenetic analysis from each individual gene was performed using parsimony with the heuristic search of PAUP*4.0 (Swofford, 2002) to ensure a consistent methodological framework with our initial study (Lecointre et al., 1998) involving the use of the ILD test for detecting character incongruence. The whole genome dataset is a large non-nucleotide dataset containing 320 binary coded characters corresponding to the presence/absence of bands in electrophoretic patterns obtained from MLEE, restriction fragment length polymorphism, and random amplified polymorphic DNA data as in Lecointre et al. (1998). ILD tests were performed each individual gene against that whole genome data, and also each individual gene against the remaining $n-1$ genes, using the partition homogeneity test function of PAUP*4.0 (Lecointre et al., 1998). To ensure the best accuracy of *P* values, for each ILD test, only positions informative for parsimony were kept (Darlu and Lecointre, 2002; Lee, 2001) (Table 1). Moreover, the low level of homoplasy in each dataset as indicated by high retention indexes (Table 1), shows that there is not enough noise to burden the severity of the ILD test (Dolphin et al., 2000). Sequence data responsible for character incongruence were detected by removing single strains and/or combinations of strains based on their "incongruent" position in visual inspection of gene trees, followed by new ILD tests. Those strain sequences which removal increases the *P* value above the threshold of 5% were considered as transfers. For each test, non-informative positions are redefined and removed. Gene sequences responsible for character incongruence were replaced by question marks in the whole combined matrix made of 11 genes and the phylogenetic analysis of that matrix was done using parsimony.

The semistrict consensus trees of each of the 7 metabolic genes are presented in Fig. 1 illustrating the level of taxonomic incongruence among the different phylogenetic trees. The level of taxonomic incongruence for the 4 polymerase genes varies in the range of what is observed for the 7 metabolic genes (Bjedov et al., 2003). The semistrict consensus tree of the simultaneous analysis of the 11 genes after removing transfers detected by

Table 1
Phylogenetic data of the 11 genes

Gene*	Position in mn	No. of Nt	No. of IS	No. of trees	CI	RI	Length of the MP trees	Each individual gene against the "Whole Genome Data"			Each individual gene against the 10 remaining genes		
								P value with all taxa	Taxa responsible for character incongruence	*P* value without incongruent taxa	*P* value with all taxa	Taxa responsible for character incongruence	*P* value without incongruent taxa
thrB	0.06	1045	51	48	0.63	0.71	172	0.01	4,13,70,60,50	0.05	0.02	4,13,57,60,50	0.19
polB	1.37	977	82	27	0.63	0.83	223	0.01	24,70	0.33	0.01	70,24,35,37	0.20
dnaE	4.42	985	45	48	0.71	0.86	115	0.01	24,17,47	0.15	0.01	24,17,70,58,35,40,60,59,51, 52,68	0.17
dinB	5.41	1010	85	12	0.65	0.83	247	0.01	52,51,58,70,4,23,20, 17,24,45	0.14	0.01	58,51,52,35,70,34,45,1,10,13	0.76
putP	23.25	892	85	9	0.62	0.79	287	0.01	4,23,24,34,58,70,46,47	0.06	0.01	4,17,23,24,34,45,58,35,47,64	0.20
icd	25.74	1166	73	30	0.59	0.69	280	0.01	52,68,23	0.54	0.01	23,24,4,68,52,62	0.29
umuC	26.52	1230	48	28	0.73	0.86	115	0.06	None		0.55	None	
trpA	28.33	727	61	96	0.67	0.89	153	0.30	None		0.02	35	0.57
trpB	28.35	1138	88	6	0.69	0.89	220	0.01	24,37,58,13,45	0.05	0.01	13,45,46,51,52,35,40,41,37,58	0.08
pabB	40.80	1003	54	2	0.78	0.86	178	0.01	41,46	0.17	0.01	41,46,51,52,57	0.21
purM	56.46	1025	43	11	0.63	0.79	112	0.01	1,4,10,13,17,23,24,51, 52,57,35,40,41	0.08	0.01	4,13,17,20,35,26,70,23,52,58	0.23

No, number; Nt, nucleotides; IS, informative sites; CI, consistency index; RI, retention index; and MP, most parsimonious.
* The gene order corresponds to the positions of the genes on the genetic map starting from 0.06′ (*thrB*) to 56.46′ (*purM*).

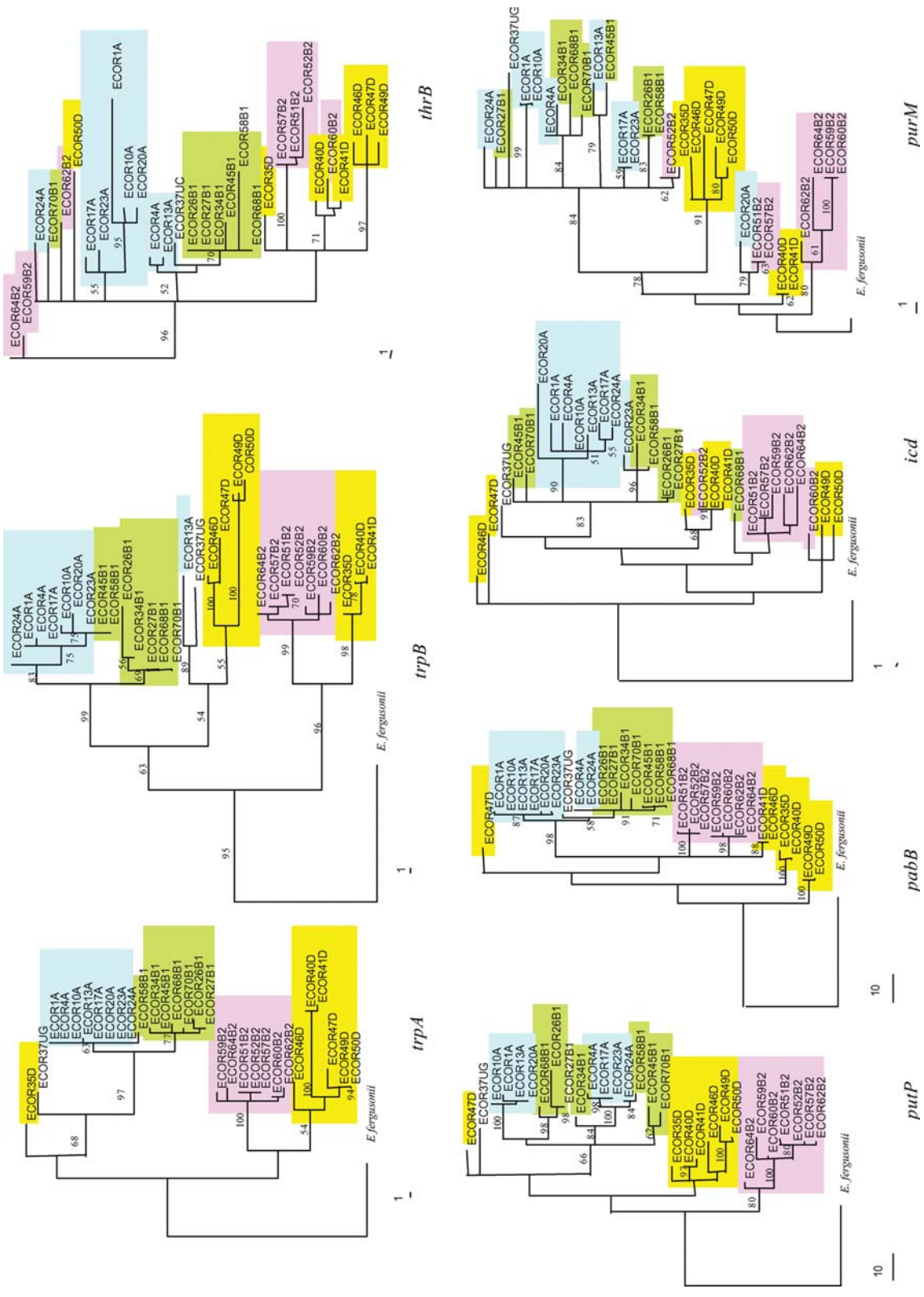

Fig. 1. Semistrict consensus trees calculated from each of the 7 metabolic genes (see Table 1 for additional phylogenetic information). All the trees were rooted on *E. fergusonii*, except *thrB* tree which has been rooted on B2 strains. Letter and color codes identifying strains in groups A (blue), B1 (green), D (yellow), and B2 (red) are as defined by the MLEE profiles in Herzer et al. (1990).

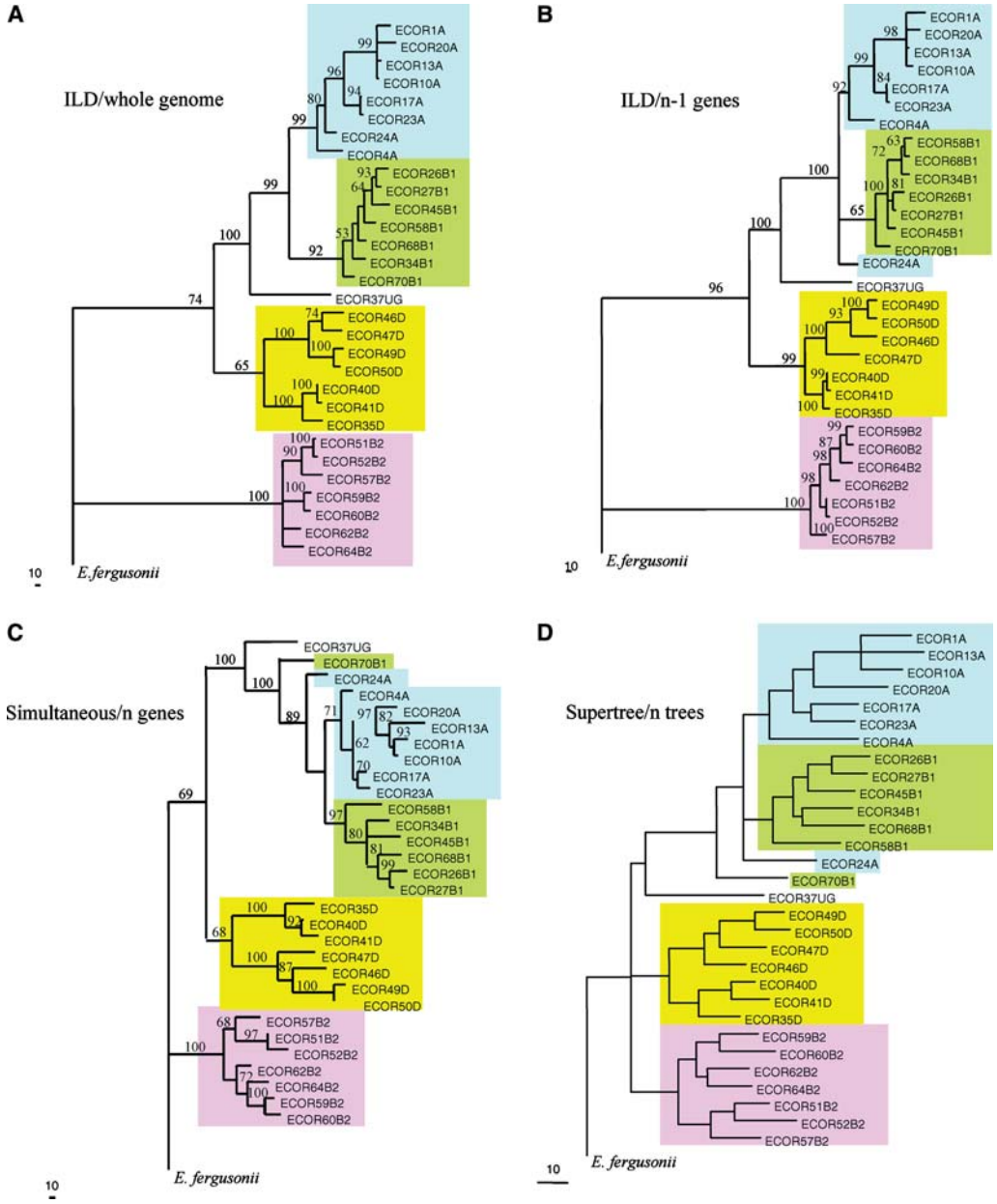

Fig. 2. (A) Semistrict consensus of 10 most parsimonious (MP) trees obtained from the simultaneous analysis of all the 11 gene sequences after removing transfers identified by the ILD test applied for each individual gene against the whole genome dataset. Sequences interpreted as transfers have been replaced by question marks. The number of informative positions is 679. Numbers above nodes are bootstrap proportions calculated from 1000 iterations. Only bootstrap above 50% are indicated. Length of MP trees is 2071, CI = 0.64, RI = 0.79. (B) Semistrict consensus of 15 MP trees obtained from the simultaneous analysis of the 11 genes after removal of the transfers identified by the ILD test applied for each individual gene against the 10 remaining ones. Sequences interpreted as transfers have been replaced by question marks. The number of informative positions is 650. Length of MP trees is 1800, CI = 0.70, RI = 0.82. (C) Semistrict consensus of 2 MP trees obtained from the simultaneous analysis of the 11 genes and all strains. The number of informative positions is 715. Tree length is 2646, CI = 0.53, RI = 0.70. (D) Supertree constructed from each individual gene MP tree, including incongruent taxa. The size of the matrix is 515 nodes (the matrix is available in Nexus format at the journal website). This supertree is a semistrict consensus of 8 trees of 701 steps. CI = 0.73, RI = 0.76. All trees have been obtained using PAUP4 with branch length given under ACCTRAN (Swofford, 2002). Heuristic searches were conducted using 100 random addition sequences. Color code is as in Fig. 1.

the use of the ILD test between each gene and the whole genome data and the corresponding tree obtained when the transfers removed were detected through ILD tests performed between each gene and the $n-1$ genes are shown in Figs. 2A and B, respectively. The topologies of these two trees are very similar. However, the bootstrap values supporting the monophyly and the position of the D group as well as the monophyly of B1 strains excluding ECOR70 are higher when the transfers removed are those detected by the ILD test performed against the

$n-1$ genes, than when they were detected against the whole genome dataset. This may be due to the fact that the number of transfers that are removed by the ILD test, when it is performed against the $n-1$ genes, is larger than when performed against the whole genome dataset (Table 1). This, in turn, is a consequence of the better structure of the $n-1$ gene dataset than the whole genome dataset. Another discrepancy between these two trees is the position of strain ECOR24 which is inside (Fig. 2A) or outside (Fig. 2B) the A group. Interestingly, the ribotype profile of this strain is typical of strains of the B1 group (Clermont et al., 2001) while both MLEE analysis (Herzer et al., 1990) and a method which uses the presence/absence of three DNA fragments to classify the strains in the 4 phylogenetic groups (Clermont et al., 2000) put ECOR24 in group A. In addition, it possesses *kps*, *pap*, and *hly* virulence genes specific of D and B2 group strains (Boyd and Hartl, 1998). Lastly, the monophyly of the B1 group including ECOR70 strain is not well supported when the ILD test is performed against the $n-1$ genes (Fig. 2B). Similar than strain ECOR24 described above, ECOR70 exhibits discrepancies between molecular typing methods. It has MLEE (Herzer et al., 1990) and ribotype (Clermont et al., 2001) profiles typical of that of strains of group B1 but the Clermont et al. (2000) method classifies it as belonging to group A. Thus, these two strains obviously have an atypical genome.

These results not only suggest that the combined dataset of the $n-1$ genes could potentially replace the whole genome dataset for the ILD test but also that the simultaneous analysis of the $n-1$ genes is roughly equivalent to a reference dataset non-sensitive to horizontal gene transfer, at least at that level of horizontal gene transfer. This implies that the simultaneous analysis of several genes swamps the effects of horizontal gene transfer when they are well distributed among the different genes. Thus, a second alternative for phylogenetic reconstruction of bacteria of similar level of horizontal gene transfer as *E. coli* could be the simultaneous analysis of a sufficient number of genes, excluding genes under diversifying selection. To test this alternative, we compared the previous trees to a tree obtained by the simultaneous analysis of the 11 genes without removing transfers using parsimony (Fig. 2C). Both simultaneous analysis of the 11 genes with and without removal of transfers lead to similar phylogenetic trees, except for the position of strains ECOR24 and ECOR70 discussed above (Figs. 2A and C). These two strains may correspond to highly mosaic genomes which have undergo horizontal gene transfers at a higher rate than the bulk of the *E. coli* species and/or new groups of *E. coli* underrepresented in our collection.

Consequently, it seems that the simultaneous analysis of the 11 genes (without removing transfers) (Fig. 2C), in which the transfers are widespread among the different genes, is effective in reducing the effects of horizontal gene transfer in strains with low level of horizontal gene transfer as *E. coli* (Maynard Smith et al., 1993). Because its simplicity in use, this method seems to be a good alternative for phylogenetic reconstruction for bacteria with low level of horizontal gene transfer as *E. coli* specially when dealing with a large dataset. It is important to note that a sufficient number of sequences from independent genes will be necessary in order to reduce the effects of horizontal transfer.

3. How to pool the data: simultaneous analysis versus supertree

Because there is poor chance that non-contiguous genes can be transferred together at once, supertrees seemed promising in summarizing the phylogeny of strains without the effects of isolated transfers on the final topology. To test this statement, we use supertrees to combine the different gene phylogenies using the Matrix Representation with Parsimony analysis (Baum, 1992; Ragan, 1992). Supertrees record only nodes from individual gene phylogenies and not their robustness, so that a single strong and false signal due to horizontal transfers in a particular gene, cannot swamp a weak but repeated signal over several genes. A single transfer, affecting a single gene, and supported by high bootstrap proportion, cannot annihilate the phylogeny of strains even if less supported. If a transfer affects a taxon in one gene and not in others, the node corresponding to the misplaced taxon in one tree should be swamped in the supertree by other nodes placing the taxon at the correct place (under the condition that at least two other genes support the correct position). Moreover, supertrees have the interesting property to allow a study of taxonomic congruence of trees with different sets of terminals, i.e., when some taxa are lacking in some trees, which inevitably occur when taxa are removed because of suspicious transfers.

We compare the trees obtained from the simultaneous analysis of the 11 genes before and after removal of transfers (Figs. 2A–C) with the supertrees obtained from the corresponding separate analyses (using PAUP*4.0). The supertree generated from the individual 11 gene trees, after removing strains responsible for character incongruence as identified with the ILD test against the whole genome dataset, yielded an almost identical tree (data not shown) to that obtained from the simultaneous analysis of the same data (Fig. 1A). However, relationships among the major groups are poorly resolved in the supertree. The supertree generated after removing strains responsible for character incongruence as identified with the ILD test against the $n-1$ genes (data not shown) is also less resolved than the tree generated by simultaneous analysis of the same

data (Fig. 1B), and it places both ECOR 24 and 70 strains outside groups A and B1. The supertree obtained from separate analyses of the 11 genes without removals (including incongruent taxa) also places these two strains outside groups A and B1 (Fig. 2D) but the general topology of this supertree is once again, less resolved than when using the simultaneous analysis method (Fig. 2C). Thus, at a low level of horizontal gene transfer, supertrees give similar topologies than simultaneous analyses but the level of resolution is always higher with this last method.

We also compared the results of the simultaneous analysis of the 11genes using parsimony with the results using maximum likelihood [HKY model (Hasegawa et al., 1985) with Gamma distribution (6 categories): estimated values of $ti/tv = 2.36$; $\alpha = 3.98$, using PAUP*4.0]. These two trees have identical topologies, with strains ECOR24 and ECOR70 outside groups B1 and A, indicating that the dataset is structured enough to provide congruent trees whatever the models and methods used (data not shown).

4. A more accurate *E. coli* phylogeny

The amount of data presented here allows a significant improvement of the accuracy of the phylogeny of *E. coli*. The relationships between the 4 main strain groups (A, B1, D, and B2) found previously, with B2 group being the more basal (Lecointre et al., 1998), is retrieved by all the used approaches, especially with *E. fergusonii* as a closer outgroup than *Salmonella*, which confirms that the basal position of the group B2 in our previous study was not due to an artefactual attraction of long branches (Felsenstein, 1978). ECOR24 and ECOR70 appear outside group A and B1 suggesting the possibility of additional groups in the *E. coli* phylogeny. ECOR37, an ungrouped strain (Herzer et al., 1990), emerges clearly between D and A/B1 groups (Fig. 2). Interestingly, this strain isolated from a marmoset, has a locus of enterocyte effacement pathogenicity island (Bergthorsson and Ochman, 1998) and is very closed to the enterohemoragic O157:H7 isolates (Escobar-Páramo and Denamur, personal data). Clades within the 4 main phylogenetic groups can now be delineated with high bootstrap values. Two clades composed by ECOR51, 52, and 57 and ECOR59, 60, 62, and 64 strains form the B2 group. ECOR51 and 52 are always sister-groups. ECOR59 and 60 are always sister-groups. The D group is also clearly composed of two main subgroups (ECOR35, 40, and 41 and ECOR46, 47, 49, and 50 strains). ECOR40 and 41 are always sister-groups. ECOR49 and 50 are always sister-groups. Within the A group, 2 clades can be delineated (ECOR1, 10, 13, and 20 and ECOR17, 23 strains). The positions of the other A strains are less clear. No clear delineation emerges within the B1 group strains (Fig. 2). The delineation of such clades is of epidemiologic interest, especially as a link between virulence and phylogeny has been reported in *E. coli* (Picard et al., 1999).

We would like to conclude that in order to avoid the problems of horizontal transfer in the phylogenetic reconstruction of bacteria with a level of horizontal gene transfer of the same order as *E. coli*, the simultaneous analysis is an accurate approach for a sufficient number of genes in which horizontal transfer (homoplasy) is widespread among the genes. The use of the ILD test to detect character incongruence between each gene and a reference dataset free of transfer remains a suitable method for estimating the level of horizontal gene transfer in bacterial species. However, in the absence of such a reference dataset, performing the ILD test against the $n - 1$ remaining genes set is a valid solution. The ILD test may also be useful in helping to identify highly mosaic strains.

Acknowledgments

This work was partially supported by grants from La Fondation pour la Recherche Médicale (PE-P) and from the Programme de Recherche Fondamentale en Microbiologie et Maladies Infectieuses et Parasitaires—MENRT.

References

Baker, R.H., DeSalle, R., 1997. Multiple sources of character information and the phylogeny of Hawaiian drosophilids. Syst. Biol. 46, 654–673.

Baum, B.R., 1992. Combining trees as a way of combining datasets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41, 3–10.

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

Bjedov, I., Lecointre, G., Tenaillon, O., Vaury, C., Radman, M., Taddei, F., Denamur, E., Matic, I., 2003. Polymorphism of gene encoding SOS polymerases in natural populations of *Escherichia coli*. DNA Repair 2, 417–426.

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

Brown, E.W., Kotewicz, M.L., Cebula, T.A., 2002. Detection of recombination among *Salmonella enterica* strains using the incongruence length difference test. Mol. Phylogenet. Evol. 24, 102–120.

Clermont, O., Bonacorsi, S., Bingen, E., 2000. Rapid and simple determination of the *Escherichia coli* phylogenetic groups. Appl. Environ. Microbiol. 66, 4555–4558.

Clermont, O., Cordevant, C., Bonacorsi, S., Marecat, A., Lange, M., Bingen, E., 2001. Automated ribotyping provides rapid phylogenetic subgroup affiliation on clinical extraintestinal pathogenic *Escherichia coli* strains. J. Clin. Microbiol. 39, 4549–4553.

Darlu, P., Lecointre, G., 2002. When does the incongruent length difference test fail? Mol. Biol. Evol. 19, 432–437.

Dolphin, K., Belshaw, R., Orme, C.D.L., Quicke, D.L.J., 2000. Noise and incongruence: interpreting results of the incongruence length difference test. Mol. Phylogenet. Evol. 17, 401–406.

Denamur, E., Lecointre, G., Darlu, P., Tenaillon, O., Acquaviva, C., Sayada, C., Sunjevaric, I., Rothstein, R., Elion, J., Taddei, F., Radman, M., Matic, I., 2000. Evolutionary implications of the frequent horizontal transfer of mistmatch repair genes. Cell 103, 711–721.

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

Escobar-Páramo, P., Giudicelli, C., Parsot, C., Denamur, E., in press. The evolutionary history of *Shigella* and enteroinvasive *Escherichia coli* revised. J. Mol. Evol.

Farris, J.S., Kallersjo, M., Kluge, A.G., Bult, C., 1995. Testing significance of incongruence. Cladistics 10, 315–319.

Felsenstein, J., 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.

Guttman, D.S., Dykhuizen, D.E., 1994. Detecting selective sweeps in naturally occurring *Escherichia coli*. Genetics 138, 993–1003.

Hasegawa, M., Kishino, H., Yano, T., 1985. Dating the human–ape spliting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 29, 170–179.

Herzer, P.J., Inouye, S., Inouye, M., Whittman, T.S., 1990. Phylogenetic distribution of branched RNA-linked multicopy single-stranded DNA among natural isolates of *Escherichia coli*. J. Bacteriol. 172, 6175–6181.

Higgins, D.G., Bleasby, A.J., Fuchs, R., 1992. CLUSTALV: improved software for multiple sequence alignment. CABIOS 8, 189–191.

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

Lecointre, G., Rachdi, L., Darlu, P., Denamur, E., 1998. *Escherichia coli* molecular phylogeny using the incongruence length difference test. Mol. Biol. Evol. 15, 1685–1695.

Lee, M.S.Y., 2001. Uninformative characters and apparent conflict between molecules and morphology. Mol. Biol. Evol. 18, 676–680.

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

Ochman, H., 2001. Lateral and oblique gene transfer. Curr. Opin. Gen. Dev. 11, 616–619.

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

Picard, B., Sevali-Garcia, J., Gouriou, S., Duriez, P., Brahimi, N., Bingen, E., Elion, J., Denamur, E., 1999. The link between phylogeny and virulence in *Escherichia coli* extraintestinal infection. Infect. Immun. 67, 546–553.

Pupo, G.M., Lan, R., Reeves, P.R., 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characters. Proc. Natl. Acad. Sci. USA 97, 10567–10572.

Ragan, M.A., 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1, 53–58.

Schevchuk, N.A., Allard, M.W., 2001. Sources of incongruence among mammalian mitochondrial sequences: COII, COIII, and ND6 genes are main contributors. Mol. Phylogenet. Evol. 21, 43–54.

Swofford, D.L., 2002. PAUP* Phylogenetic Analysis Using Parsimony (* and other methods). Sinauer Associates, Sunderland, MA.