Inferring the Evolutionary History of IncP-1 Plasmids Despite Incongruence among Backbone Gene Trees

Research Article

Diya Sen[1,2], Celeste J. Brown[1,2,3], Eva M. Top[1,2,3] and Jack Sullivan[1,2,3]

[1]Institute for Bioinformatics and Evolutionary Studies (IBEST), [2]Bioinformatics and Computational Biology Graduate Program, [3]Department of Biological Sciences, University of Idaho, Moscow ID 83844-3051, USA

Corresponding author:

Jack Sullivan

Department of Biological Sciences

University of Idaho

257 Life Sciences South

Moscow ID 83844-3051

Phone: 1-208-885-4049

Fax: 1-208-885-7905

E-mail: jacks@uidaho.edu

Key words: plasmid, phylogeny, species tree, genomics, Horizontal Gene Transfer

Running Head: Plasmid phylogenomics.

Abstract

Plasmids of the incompatibility group IncP-1 can transfer and replicate in many genera of the Proteobacteria. They are composed of backbone genes that encode a variety of essential functions as well as accessory genes that have implications for human health and environmental remediation. While it is well understood that the accessory genes are transferred horizontally between plasmids, recent studies have also provided examples of recombination in the backbone genes of IncP-1 plasmids. As a consequence, phylogeny estimation based on backbone genes is expected to produce conflicting gene tree topologies. The main goal of this study was therefore to infer the evolutionary history of IncP-1 plasmids in the presence of both vertical and horizontal gene transfer. This was achieved by quantifying the incongruence among gene trees and attributing it to known causes such as, a) phylogenetic uncertainty, b) coalescent stochasticity, and c) horizontal inheritance. Topologies of gene trees exhibited more incongruence than could be attributed to phylogenetic uncertainty alone. Species-tree estimation using a Bayesian framework that takes coalescent stochasticity into account was well supported, but it differed slightly from the maximum likelihood tree estimated by concatenation of backbone genes. After removal of the gene that demonstrated a signal of intergroup recombination, the concatenated tree was congruent with the species-tree estimate, which itself was robust to inclusion/exclusion of the recombinant gene. Thus, in spite of horizontal gene exchange both within and among IncP-1 subgroups, the backbone genome of these IncP-1 plasmids retains a detectable vertical evolutionary history.

Introduction

Self-transferable broad-host-range (BHR) plasmids of Proteobacteria play a crucial role in bacterial adaptation because they can exchange genes among phylogenetically distant bacteria (Adamczyk and Jagura-Burdzy 2003). These extra-chromosomal DNA molecules provide bacteria with a variety of novel phenotypic traits and contribute to the alarmingly rapid spread of multi-drug resistance in human pathogens. The most promiscuous plasmids (i.e., those with the broadest host range) belong to the incompatibility groups IncP (Adamczyk and Jagura-Burdzy 2003), IncW (Fernandez-Lopez et al. 2006), IncU (Rhodes et al. 2004) and the recently defined group PromA (Gstalder et al. 2003, Van der Auwera et al. 2009). Plasmids belonging to the same group are said to be incompatible since they cannot co-exist in the same cell line. IncP, also called IncP-1, plasmids are considered to be among the most promiscuous and carry many kinds of accessory genes (Adamczy and Jagura-Burdz 2003, Schlüter et al. 2007, Thomas and Smith 1987). They replicate in different classes within Proteobacteria and can also mobilize non-self-transferable plasmids into Gram-positive bacteria (Mazodier, Petter, and Thompson 1989), cyanobacteria (Kreps et al. 1990), and even eukaryotes (Heinemann and Sprague 1990). Not only are they present in diverse environments such as manure (Binh et al. 2008), agricultural soils (Sen et al. 2011, Top et al. 1995), streams (Akiyama, Asfahl, and Savin 2010, Smalla et al. 2006), and wastewater treatment plants (Schlüter et al. 2007), they are also a cause for concern in the clinic because of the drug resistance they often encode (Ingram et al. 1973, Novais et al. 2006).

Plasmids typically consist of backbone genes that are involved in replication, stable inheritance and control, and conjugative transfer (Fig 1), in addition to accessory genes that confer variable host-beneficial traits. Adamczyk and Jagura-Burdz (2003) and Schlüter et al.

(2007) provide excellent reviews of backbone genes of IncP-1 plasmids. Of the nearly 45 backbone genes found on IncP-1 plasmids, approximately 33 are shared by all plasmids that have been completely sequenced to date. The conservation of this complement of backbone genes across the IncP-1 plasmids suggests that they may share a common phylogenetic history.

Phylogenies inferred from single or concatenated backbone genes have shown the IncP-1 group to be composed of five diverse subgroups: IncP-1α, -β, -γ, -δ, and -ε (Bahl et al. 2007, Haines et al. 2006, Pansegrau et al. 1994, Thorsted et al. 1998, Vedler, Wahter, and Heinaru 2004). Recently, three new potential IncP-1 subgroups have been identified: ζ (Norberg et al. 2011), η (Sen, Yano, Bauer, Rogers, Van der Auwera, Brown and Top, unpublished) and an unnamed subgroup (Pachulec and van der Does 2010), thus indicating the tremendous diversity that exists within this plasmid group. The biological significance of these subgroups is unknown and more work is required to distinguish between them phenotypically. Furthermore, the validity of classifying plasmids into subgroups based on the phylogeny of a single gene or few genes is tenuous because of evidence of horizontal gene transfer (HGT) between plasmid backbones in the group. HGT among plasmid backbones is mediated by conjugation and subsequent recombination between coexisting plasmids, resulting in chimeric plasmid genomes. In such cases, phylogenetic inference based on both recombinant and non-recombinant regions may lead to conflicting results. The first evidence of recombination between IncP-1 plasmids was clearly shown in the IncP-1β plasmid, pB10 (Schlüter et al. 2003). We recently showed another instance of recombination in the IncP-1δ plasmid pIJB1, which has two sets of the replication and transfer genes *trfA*-*trbE*, one of which was acquired from an IncP-1β plasmid and the other was native to its subgroup (Sen et al. 2010). Recently, more IncP-1 plasmids (namely, pB3, pBP136 and pAOVO02) were identified as recombinants through the analysis of concatenated alignments of

backbone genes (Norberg et al. 2011). It is thus becoming increasingly apparent that although IncP-1 plasmids are incompatible over long periods of time, they may co-exist long enough to allow recombination, which may also be true for other incompatibility groups. Thus, the contribution of recombination to the evolution of plasmids may be greater than previously thought, and a phylogenomic approach is required to elucidate the extent of HGT and how it impacts phylogenetic signal and phylogeny estimation.

It is known that phylogenies inferred from multiple loci often contradict each other (Pollard et al. 2006, Rokas et al. 2003) and that this incongruence among gene phylogenies can have three causes: phylogenetic uncertainty, coalescent stochasticity or random sorting of ancestral polymorphisms, (Maddison 1997), and HGT (Maddison 1997). Phylogenetic uncertainty can be attributed to random error, caused by the sample of characters chosen (Graybeal 1998), and/or systematic error (Yang, Goldman, and Friday 1994) caused by the introduction of analytical bias during phylogeny estimation. Coalescent stochasticity is caused by the random sorting of polymorphisms in an ancestral population and sometimes occurs in a way that is not in agreement with the species history, such that non-sister taxa or subgroups share the same states. This is often seen in large populations that have undergone recent speciation events (Maddison 1997). In the IncP-1 plasmid, recombination within subgroups (i.e., intra-group HGT) may lead to differential sorting of ancestral polymorphisms at multiple loci in the same way that recombination within sexually reproducing species does. HGT is the transfer of genetic material between different taxa, here plasmids of different subgroups (Maddison 1997). Classically, resolution of incongruence and estimation of species trees were accomplished either with consensus methods (Bryant 2003) or total evidence methods (Kluge 2004). Consensus methods rely on estimating a summary tree from a collection of gene trees, while total evidence involves

the analysis of a single gene tree from a collection of genes that are concatenated into a supermatrix. More recently, however several methods have been developed to estimate species trees using coalescent models (e.g., BEST, Liu 2008; STEM, Kubatko, Carstens, and Knowles 2009; *BEAST, Heled and Drummond 2010). These methods assume that incongruence is being generated exclusively by the stochastic sorting of ancestral polymorphisms (Liu et al. 2009), although they appear to be robust to HGT/hybridization to at least some degree (Chung and Anè 2011).

The goal of this study was to determine the evolutionary history of the backbone of IncP-1 plasmids by systematically examining congruence among gene trees estimated from the backbone genes. Results showed extensive incongruence among the trees, as expected. Therefore, we applied a series of phylogenetic analyses to estimate the evolutionary history of these plasmids in the face of this incongruence. This is the first gene-by-gene analysis of backbone genes of IncP-1 plasmids, and in spite of strong incongruence among genes, we derive a strongly supported estimate of the relationships among the five well-known IncP-1 subgroups by using species-tree estimation approaches.

Materials and Methods

Plasmids and genes.

Complete sequences of IncP-1 plasmids were extracted from GenBank or determined by us, a total of 65. Only backbone genes that were common to all plasmids were used for phylogeny estimation (dark grey ORFs in Fig. 1). For those plasmids that showed identical sets of backbone gene sequences, only one representative plasmid was retained. In addition, plasmids like pEST4011 that were missing a large section of backbone sequence shared by all

other plasmids were not included, nor was plasmid pIJB1 because of its duplicated *trfA* and *trb* genes of IncP-1δ and IncP-1β descent (Sen et al. 2010). The plasmids from the recently proposed ζ, η and an unnamed subgroups (see Introduction) were not included because their sequences only became publically available in the middle of this study, and because of their wide sequence divergence and lack of evidence that they are physically incompatible with the prototype IncP-1 plasmids. The final set of plasmids used in this study is shown in Table 1. All plasmids, including those that were previously published, are referenced only by their Genbank accession numbers or NCBI reference numbers.

Sequencing and annotation.

The following 15 plasmids were sequenced as part of this study (references refer to the studies that first isolated and described the plasmids): pB1 and pB12 (Dröge , Pühler, and Selbitschka 2000), pEMT3 (Top et al. 1995), pG527 (Götz et al. 1996), pC11 and pNB1, (Boon et al. 2001), pKV29 (Stolze et al. 2012), pTB30 (Dejonghe et al. 2002), pKS208 and pKS212 (Heuer et al. 2002), pRSB222, pRSB223 (Schlüter and Sczcepanowski, unpublished), pWEC911 (Smalla and Hill, unpublished), pYS1 (Sota, unpublished) and pDS3 and pMBUI1 (Sen, Yano, Bauer, Rogers, Brown, and Top, unpublished). All plasmid sequences were determined at the DOE Joint Genome Institute (Walnut Creek, CA) by either of two methods. Pyrosequencing of plasmids pDS3, pMBUI1, pRSB222, pRSB223 was performed on a GS FLX with the Titanium sequencing chemistry to approximately 90x coverage (Roche/454 Life Sciences, Branford CT). Sequence data were assembled using the Newbler software (Roche/454 Life Sciences, Branford CT). Plasmids pB1 and pB12, pC11, pEMT3, pG527, pKS208, pKS212, pNB1, pTB30, pWEC911, and pYS1 were sequenced using the Sanger method. Approximately 3-kb clone libraries were constructed for DNA sequencing of 384 clones and sequences were determined for

each plasmid in both directions. These sequences were assembled at JGI using PGA a platform for comparative genome assembly based on genetic algorithm optimization (Zhao et al. 2009). Any gap closure and polishing was done in house by primer walking. Automatic annotations were provided by the IGS Annotation Engine at the Institute for Genome Sciences, School of Medicine, University of Maryland (http://ae.igs.umaryland.edu/cgi/index.cgi) for plasmids pDS3, pMBUI1, pRSB222, and pRSB223 and by the J. Craig Venter Institute Annotation Service (http://www.jcvi.org/cms/research/ projects/annotation-service) for the rest of the plasmids. These were followed by manual annotation by the authors. GenBank accession numbers are provided in Table 1..

Nucleotide sequence alignments and model selection.

The amino acid sequences of each gene were aligned with ClustalX (Thompson, Gibson and Higgins 2002). Tranalign (Rice, Longden and Bleasby 2000) was used to align the nucleotide sequences of each gene guided by the aligned amino acid sequences. Nexus formatted files were created from aligned nucleotide sequences for analyses in PAUP* (Swofford 2003) and MrBayes v 3.1.2 (Ronquist and Huelsenbeck 2003) and PHYLIP formatted files were created for analyses in RAxML (Stamatakis 2006). For the concatenated tree, individual genes were aligned and concatenated and the concatenated alignment was partitioned by codon positions. Model selection for maximum likelihood (ML) and Bayesian estimation were done with the program DT-ModSel (Minin et al. 2003). A list of models selected for each analysis is shown in Table 2.

Maximum-likelihood analyses.

For gene-tree estimation, iterative heuristic searches were performed using PAUP* and the iterative approach described buy Sullivan et al. (2005). ML searches were carried out with

tree bisection and reconnection (TBR) branch swapping on 20 random starting trees generated by stepwise addition. For the concatenated data, an ML tree was inferred using the program RAxML (Stamatakis 2006) with the GTR+Γ model and parameters estimated separately for the three-codon partitions. Support values were estimated from 100 non-parametric bootstrap replicates (Felsenstein 1985). The tree with the highest likelihood was used as the ML estimate of the concatenated tree, referred to as ML_{concat} below.

Bayesian posterior probability distributions.

The program MrBayes v 3.1.2 (Ronquist and Huelsenbeck 2003) was used for estimating the posterior probability distributions of gene trees for each backbone gene and also for the concatenated data. A Markov Chain Monte Carlo algorithm was used to sample the posterior distribution of trees by running four chains for up to 8 million generations and sampling trees every 100 generations. Convergence of chains was assessed by plotting the standard deviation of split frequencies against the number of generations. A separate partitioned analysis was carried out for the concatenated data using GTR+I+Γ for the first and third codon positions and GTR+Γ for the second position. For both gene and concatenated data sets, trees sampled prior to convergence were discarded, and the remaining trees were used for Bayesian hypothesis testing.

Congruence tests.

Initial assessment of phylogenetic uncertainty as the cause of incongruence among gene trees was conducted using parametric bootstrap analyses (i.e., SOWH tests; Goldman et al. 2001). ML searches for each gene were conducted in PAUP* (Swofford 2003), as described above to provide an ML estimate of the gene tree, ML_{gene}. Model parameters and branch lengths were re-optimized after exclusion of missing and ambiguous characters. ML searches were

constrained to fit the topology of the concatenated tree to find the best fit of the individual gene data to the hypothesis that phylogenetic error is the only source of incongruence among the gene trees, ML_{hyp}. The test statistic was the difference in log likelihood scores of the two trees, [$\delta = \ln L(ML_{hyp}) - \ln L(ML_{gene})$], the significance of which was evaluated under a frequentist framework by simulation. The constrained tree (ML_{hyp}) was treated as the true tree on which 100 replicate data sets were simulated with SEQ-GEN (Rambaut and Grassly 1997) using the same model parameters that were optimised from the real data. The lengths of the simulated sequences were set to be identical to the length of each gene, and PAUP* was used to find the ML tree and the best tree constrained to fit the topology of the concatenated tree for every replicate. This provided the null distribution against which we compared the test statistic (δ) to evaluate the probability that phylogenetic uncertainty can explain the observed incongruence among the backbone gene trees.

Because the SOWH tests rely on point estimates of model parameters in simulation of the null distribution, we also assessed phylogenetic uncertainty with a Bayesian framework that marginalizes across uncertainty in model parameters. The tree filters option in PAUP* (Swofford 2003) was used to assess the proportion of trees in the posterior distribution of trees for each gene that was congruent with the topology of the concatenated ML tree, ML_{concat}. We also assessed the reciprocal congruence (i.e., proportion of trees in the posterior distribution of trees for the concatenated data consistent with the topology of each ML gene tree). This yielded the posterior probability that the incongruence is due to phylogenetic uncertainty.

Finally, we used the conservative non-parametric SH-test (Shimodaira and Hasegawa 1999) to assess the incongruence between each gene's ML tree and the concatenated tree. To render this test conservative, we included each of the 28 single-gene ML trees and the

concatenated tree in the centering step. We then calculated p-values for each gene using RELL bootstrap with 1000 replicates.

Species tree estimation with *BEAST.

To account for coalescent stochasticity, we applied the traditional notion of a species to IncP-1 subgroups identified by earlier studies. Thus, intra-group HGT can be treated as analogous to recombination within sexually reproducing species, and inter-group HGT treated as analogous to introgressive hybridization. Nexus formatted files of the 28 genes were used as input for *BEAST (Heled and Drummond 2010). Substitution models were chosen as above, and were unlinked across genes with parameters estimated separately for each gene. Plasmids were assigned to the five subgroups (a proxy for species) based on previous studies. In BEAST (v 1.6.0) a Markov Chain Monte Carlo algorithm was used to sample the posterior distribution of trees by conducting five independent runs of 100 million generations each using a Yule prior for the species tree, a piecewise linear and constant root prior for population size and uncorrelated, lognormal, relaxed-clocks. Post-burnin trees were combined with the program LogCombiner (BEAST v 1.6.0), and chains were assumed to converge when the average standard deviation of split frequencies was found to be < 0.011. The maximum clade credibility tree with posterior probability of each node was computed with the program TreeAnnotator (BEAST v 1.6.0).

Detection of recombinants.

To detect recombination among the plasmids in the data set, alignment files of the backbone genes were concatenated in the order and orientation in which they appear on IncP-1 plasmids (Fig. 1). The recombination detection programs RDP, GENECONV, BootScan, MaxChi, Chimaera, which are implemented in RDP3 (Martin et al. 2010), were run with default parameters. Only recombinants that were identified by at least two programs were considered.

Results

Plasmids and genes.

To infer the evolutionary history of plasmids from the incompatibility group IncP-1, a set of 65 completely sequenced IncP-1 plasmid genomes was retrieved from Genbank and our own plasmid sequence collection. They were selected based on previously published assignment to one of the five major IncP-1 plasmid subgroups (α-ε) or our own comparative sequence analysis. A total of 28 backbone genes were found to be common to all plasmids and therefore included in this study (Fig. 1, dark grey ORFs). After removing duplicates (plasmids with identical sets of backbone gene sequences), and including our 15 newly sequenced plasmids, a final set of 46 plasmids was obtained (Table 1). For example, the backbone gene sequences of pJP4 were identical to those of pB10, and therefore not included. Visual inspection suggested that all alignments were of good quality. Because the 5' ends of genes *trfA1* and *kfrC* did not appear to be homologous between different subgroups, they were excluded from the analysis.

Maximum-likelihood analyses.

Gene trees were produced by separate ML analyses of the 28 backbone genes using the nucleotide substitution models in Table 2 (Fig. S1). Analyses of 21 of those genes produced 4 topologies that were similar and differed only in the placement of the IncP-1δ and IncP-1α plasmids (Fig. 2). The remaining 7 gene trees were very different from each other and did not agree with any of the four common topologies (Fig. S1). Topology 1 (Fig 2A) was consistent with trees inferred from almost half of the genes (13 of 28; *trfA2*, *trbA*, *trbC*, *trbG*, *traG*, *traH*, *traI*, *kfrC*, *kfrB*, *kfrA*, *korB*, *korA* and *kleE*). Topology 2 (Fig. 2B) was found for three genes (*trbD*, *trbK* and *traJ*) and differed from topology 1 in that it swapped the positions of the IncP-1δ plasmid pAKD4 and the IncP-1α plasmids. Topology 3 (Fig. 2C) grouped pAKD4 with the

IncP-1α plasmids and was inferred from the three genes *trbF*, *trbI*, and *traE*. Topology 4 (Fig. 2D) grouped pAKD4 with the epsilon plasmids and was supported by the *klcA* and *korC* gene trees (Fig. 2D). The genes corresponding to topologies 1-4 and the unique topologies (U) are also indicated on Fig. 1 (inside circle). Multiple topologies indicate incongruence among gene phylogenies.

Examination of congruence.

To test the hypothesis that the 28 backbone genes of IncP-1 plasmids have a single evolutionary history, and that the incongruence described above is only due to phylogenetic uncertainty, each gene tree was compared statistically to the concatenated tree first using the parametric bootstrap. For the concatenated tree (shown in Fig. 3), individual genes were aligned and concatenated in the order in which they appear on IncP-1 plasmids; *trfA2*, *trbA*, *trbB*, *trbC*, *trbD*, *trbF*, *trbG*, *trbI*, *trbJ*, *trbK*, *traD*, *traE*, *traF*, *traG*, *traH*, *traI*, *traJ*, *traK*, *traL*, *kfrC*, *kfrB*, *kfrA*, *korB*, *incC*, *korA*, *kleE*, *korC* and *klcA* (Fig. 1). The concatenated tree had the same topology as topology 3 described (Fig. 2C). It represents the null hypothesis that all genes have a single history (i.e., all gene trees are estimates of a single gene tree) as would be the case in the absence of recombination, either within or among groups. The observed test statistic was evaluated against the distribution of test statistics generated under the null hypothesis. The observed test statistics were significantly larger (P-value < 0.01) than the null distributions generated for all 28 genes; in spite of being congruent with the concatenated tree at the deeper nodes, topology 3 still had significant differences at the terminal nodes. Therefore, the incongruence observed between each gene tree and the concatenated tree could not be attributed to phylogenetic uncertainty alone. Thus, the 28 backbone genes have multiple evolutionary histories (i.e., gene trees) because of horizontal transfer and/or coalescent stochasticity.

We use the *trbB* gene to illustrate incongruence between a gene tree and the concatenated tree because the difference between these two trees was the largest of all, with a test statistic of 651 log likelihood units (Fig. 4 A-C). After 14 plasmids were removed from the analysis (several IncP-1β plasmids, the IncP-1-δ plasmid, and all IncP-1ε plasmids, based on recombination detection – see below), the difference in the scores of the ML trees for *trbB* (ML_{trbB}) and the concatenated data (ML_{hyp}) decreased to 0 (Fig 4 D-F). The new test statistic fell within the null distribution, and the P-value was calculated to be 0.49 (Fig 4 D-F). Thus, the null hypothesis of discordance due to phylogenetic uncertainty could not be rejected for this analysis, which illustrates that the plasmids we removed were responsible for the incongruence between the concatenated tree and the gene tree observed for *trbB*.

In contrast to the frequentist approach used above, the Bayesian approach determines the conditional probability of the hypothesis that incongruence between a gene tree and the concatenated tree is due to phylogenetic uncertainty. MrBayes v 3.1.2. (Ronquist and Huelsenbeck 2003) was used to generate the posterior probability distribution of trees for each gene. The tree filter option in PAUP* (Swofford 2003) was then used to estimate the proportion of trees in the distribution that have the same topology as the concatenated tree, ML_{concat}. The fraction of trees retained in the filter represents the posterior probability of the hypothesis, and the probabilities of each gene tree being congruent with the concatenated tree can therefore be calculated. No trees were retained by the filtering procedure for any of the genes; therefore the probability that incongruence between each of the gene trees and the concatenated tree is due to phylogenetic uncertainty approaches zero. Again, elimination of the same 14 plasmids resulted in 6925 trees out of 7419 trees in the posterior distribution for the concatenated dataset that were consistent with the topology of the *trbB* tree (P-value = 0.93). This agrees with the results from

the parametric bootstrap analysis that incongruence was caused by the 14 plasmids. Overall, Bayesian hypothesis-testing and parametric bootstrap analyses show that incongruence among gene trees is not due to phylogenetic uncertainty alone.

Not surprisingly, the results of the non-parametric SH-tests are not as uniform (Table 3). This test suggested that 14 of the 28 gene trees are not significantly different than the concatenated tree, but the other 14 are different. Thus, even our conservative implementation of this relatively low power test detected significant incongruence between the ML estimate of the gene tree and the concatenated tree for half of the backbone genes.

Species tree estimation with *BEAST.

*BEAST (Heled and Drummond 2010) was used for estimating the phylogenetic history represented by a species tree for IncP-1 plasmids. Figure 5A shows the maximum clade credibility trees estimated by *BEAST. The species-trees are consistent with topology 1, which was found for almost half of the ML gene trees (13 genes out of 28, Fig. 2A). This suggests that the incongruence detected in the parametric bootstrap and Bayesian tests above may largely be attributable to coalescent stochasticity (i.e., recombination within subgroups).

Effect of recombination on tree estimation.

To detect recombination between plasmids, individual genes were aligned and concatenated in the order and orientation in which they appear in IncP-1 plasmids (Fig. 1, all dark grey ORFS, clockwise starting with *trfA2).* The concatenated alignment was analyzed using RDP, GENECONV, BootScan, MaxChi and Chimaera, commonly used algorithms for detecting recombination among nucleotide sequences. Extensive recombination was detected within the IncP-1β subgroup, supporting the conclusion that incongruence is attributable to coalescent stochasticity. In contrast, only one instance of recombination between subgroups was detected

from position 7293 to 9902 of the concatenated alignment of the IncP-1α plasmids. This region corresponds to most of the *traE* and *traD* genes, specifically from nucleotide 26 of *traE* to 36 nucleotides from the end of *traD* (Fig. 1; the genes are approximately 2,000 bp and 400 bp long, respectively). The recombination appears to have been between the ancestor of the IncP-1α plasmids and an IncP-1δ plasmid similar to pAKD4. Visual inspection of the aligned *trbB* genes also clearly showed a recombination event that included pAKD4, several of the IncP-1β plasmids and the ancestor of the IncP-1ε plasmids, which would help explain the highly incongruent *trbB* gene tree described above. The amount of recombination among the IncP-1β plasmids may have masked this recombination event from the detection programs.

The low nodal support (posterior probability) in the species tree at the node uniting IncP-1α, -β, -δ, -ε plasmids in Figure 5A prompted us to exclude the long putative recombinant gene *traE* in species-tree estimation. After exclusion of *traE* the same topology was obtained, but with higher posterior probability at the relevant node (Fig. 5B). Similarly, when we excluded *traE* and kept all other ML parameters constant, an ML_{concat} tree was obtained that was now congruent with topology 1 and no longer with topology 3 (data not shown). Topology 1 was the topology that was congruent with almost half of the gene trees as well as the species tree. These results support the conclusion that the *traE* gene has been involved in inter-group recombination.

Discussion

Our goal was to infer the evolutionary history of IncP-1 plasmids from their backbone genes in the presence of HGT both within and among subgroups. Recent studies have shown that these genes have evolved not only by acquiring mutations during vertical gene transfer, but also by recombining with homologs on other IncP-1 plasmids (Schlüter et al. 2003, Sen et al. 2010,

Norberg et al. 2011). Our challenge was therefore to infer the phylogeny of these plasmids in the presence of both vertical and horizontal inheritance. Our approach was to examine congruence among the inferred phylogenies of the backbone genes and determine the causes of incongruence so that they could be accommodated in phylogeny estimation. In order to rule out phylogenetic uncertainty as one of the possibilities, we compared each gene tree to a tree obtained from concatenating alignments of all 28 backbone genes. Concatenation ignores multiple histories of the underlying data and represents a single history for all 28 backbone genes. The null hypothesis that each gene tree is consistent with that single history was rejected for all 28 genes; the observed incongruence could not be attributed to phylogenetic uncertainty alone, indicating the presence of coalescent stochasticity and/or inter-group HGT.

Assessing the impact of coalescent stochasticity in plasmid phylogeny is somewhat less straightforward, but we have applied species-tree estimation procedures that attempt to model stochastic lineage sorting. In these analyses, we have used existing plasmid 'taxonomies' based on usually a single gene to group plasmid backbone genomes into putative taxa (subgroups). Within these subgroups, we have assumed that recombination behaves in a manner analogous to independent assortment in sexually reproducing species. We thus used *BEAST (Heled and Drummond 2010) to estimate the backbone "species"-tree and accommodate the stochastic process of sorting ancestral polymorphisms. The output from *BEAST included a maximum clade credibility tree (Fig. 5A) with the same topology as topology 1 (Fig. 2A) and not topology 3 (Fig. 2C) as consistent with the concatenated data set. However, the posterior probability of the node uniting the IncP-1α, IncP-1β and IncP-1ε plasmids was moderate, 0.89 (Fig. 5A). To address indirectly if this relatively low nodal probability was due to intergroup recombination, and therefore violation of the assumption that all incongruence is due to coalescent stochasticity,

we identified the long *traE* gene of IncP-1α plasmids as a putative recombinant with an IncP-1δ pAKD4-like plasmid, and excluded it from the species tree estimation. This is analogous to eliminating putative hybrids from phylogenetic analysis. The topology of the species tree estimated in the absence of *traE* (Fig. 5B) was identical to that produced before removing it, but support for the node in question increased (Fig. 5A). This suggests that inclusion of this putatively chimeric gene generated by inter-group HGT is the cause of reduced support. Recombination in this gene was also detected previously by Norberg et al. (2011) and Sen et al. (2010).

Plasmids of the IncP-1β subgroup, which can be further divided into IncP-1β1 and IncP-1β2 plasmids based on reciprocal monophyly (Fig. 3), have undergone extensive recombination. It is important to note that these plasmids or the corresponding recombinant genes were not excluded from species tree estimation because they all occurred within a subgroup (treated here as species) and were not expected to interfere with the analysis. Recombinants can be grouped into a) recombination events within the IncP-1β2 subgroup (pB4, pNB1, pA1, pB1, pRSB222, pRSB223 and pYS1), and b) recombination events between IncP-1β1 and IncP-1β2 plasmids (pAOVO02, pAKD18, pDS3 and pB10). Recombination in plasmid pAOVO02 and between pB10 and a pB4-like plasmid had been suggested earlier (Schlüter et al. 2003, Heuer et al. 2004, Norberg et al. 2011). Interestingly there were fewer observations of recombination between members of the different subgroups. Except for recombination between the IncP-1α and IncP-1δ plasmid, few other instances were observed: those in the *trbB* gene and another between members of the IncP-1β1 subgroup and the IncP-1ε plasmid pEMT3 (data not shown). Recombination crossover points previously detected just upstream and downstream from *trbB* by Norberg et al. (2011) support our finding that the *trbB* region is prone to recombination. There

are two possible explanations for the limited recombination between members of different subgroups; one is that as similarity between members of different subgroups decreases so does the possibility of recombination. The second is that because the other subgroups don't have as many sequenced plasmids as the IncP-1β subgroup, the genomes of putative recombinants have not yet been sequenced.

Our study provides yet another example of how concatenation may fail to produce accurate estimates of the species tree in complex datasets. Although almost half of the gene trees (13 out of 28) supported topology 1, the concatenated tree supported topology 3. Excluding *traE* from the concatenated dataset and keeping all other ML tree estimation parameters constant, yielded a tree that was congruent with topology 1. Thus, the number or configuration of variable sites in the long *traE* gene may have been enough to dominate the phylogenetic signal in the other genes, a phenomenon called data swamping such that one or a few partitions provide all the signal in a concatenated analysis (e.g., Edwards 2009).

While the relationships among the subgroups are largely congruent among gene trees, caution must be exercised in choosing genes for inferring phylogenies. Genes like *trfA2*, responsible for plasmid replication and *traI*, responsible for conjugative transfer are often used for inferring plasmid phylogenies. Their gene trees (Fig. S1) were largely in agreement with topology 1 (Fig. 2A) and the species trees estimated by *BEAST (Fig. 5), and therefore these genes are recommended to infer the phylogenetics of IncP-1 plasmids. Other genes that would be suitable because they showed the same topology as the species tree are *traG*, *traH*, *trbA*, *trbC*, *trbG*, *kfrA*, *kfrB*, *kfrC*, *korB*, *korA* and *kleE*. Previous studies have generally used these same genes, establishing in many cases the species tree defined here. Vedler et al. (2004) used individual *trfA2, traG,* and *korA* gene trees to clearly establish that the then newly defined IncP-

1δ subgroup was separate from the IncP-1α subgroup. Similarly, three of the four genes chosen to infer the phylogeny of the then novel IncP-1ε group, *trfA2, korB and trbA,* are part of this set of genes (Bahl et al. 2007). To define the InP-1γ subgroup, Haines et al. (2006) built a tree using five concatenated genes, *korA, incC2, korB, korC* and *kfrC;* three of these (*korA, korB, kfrC*) generated a tree with topology 1 in our study, while the *korC* tree had topology 4 and the *incC* tree a unique topology (Fig. 2 and Fig. S1). Interestingly, their tree based on the five concatenated sequences is more similar to topology 4 than to topology 1, with the IncP-1δ and –β plasmids sharing a common ancestor rather than the IncP-1α and –β plasmids. Unfortunately there were no IncP-1ε plasmids available at that time, so the effect of the five-gene concatenation on the topology with respect to that subgroup could not be evaluated. To infer the phylogeny of two novel IncP-1β2 catabolic plasmids, a tree was recently generated based on 24 concatenated backbone protein sequences, including TraE (Król et al. 2012); interestingly its topology corresponded to topology 3 of the concatenated tree in this study. Based on our findings, topology 1 most closely represents the true evolutionary history of IncP-1 plasmids. Therefore we recommend using the genes that generated trees of topology 1 (Fig. 2).

To summarise, the backbones of IncP-1 plasmids have evolved by a combination of vertical and horizontal gene transfer, with the majority of recombination events being restricted to within the IncP-1β subgroup. Why recombination is seen so frequently in IncP-1β plasmids and why the *traE* gene of IncP-1α and IncP-1δ plasmids underwent recombination remains unknown. These recombination events may either be neutral, or selectively advantageous to their hosts. In fact, recombination may be a tool that adds further flexibility to plasmids by allowing them to rapidly adapt to changing bacterial hosts and environmental conditions.

Acknowledgements

New plasmid genome sequence data were generated with financial support from the National Science Foundation Grant EF-0627988. We are grateful to Kerrie Barry, Brian Foster, and Alla Lapidus at the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) for providing draft genome sequences, supported by the DOE Office of Science under Contract No. DE-AC02-05CH11231. We also acknowledge the National Institutes of Health NCRR COBRE grant P20RR16448 for an IBEST fellowship to D. S. and for support for the IBEST Computational Resources Core at the University of Idaho.

We thank the following people for providing us with some of the plasmids that we sequenced and used in this study: N. Boon (pNB1, pC11), W. Dejonghe (pTB30, pWDL7), A. Schlüter and R. Szczepanowski (pB1, pB12, pKV29, pRSB222 and pRSB223), K. Smalla and H. Heuer (pG527, pKS208, pKS212, pWEC911), and M. Sota (pYS1). We also appreciate the help of J. Król and H. Yano for finalizing, annotating, and submitting a few plasmid genome sequences (pNB1, pC11, pTB30 and pMBUI1, pKS208).

Literature Cited

Adamczyk M, Jagura-Burdzy G. 2003. Spread and survival of promiscuous IncP-1 plasmids. Acta Biochim. Pol. 50:425-453.

Akiyama T, Asfahl, KL, Savin MC. 2010. Broad-host-range plasmids in treated wastewater effluent and receiving streams. J. Environ. Qual. 39:2211-2215.

Bahl MI, Hansen LH, Goesmann A, Sørensen SJ. 2007. The multiple antibiotic resistance IncP-1 plasmid pKJK5 isolated from a soil environment is phylogenetically divergent from members of the previously established α, β and δ sub-groups. Plasmid 58:31-43.

Binh CT, Heuer H, Kaupenjohann M, Smalla K. 2008. Piggery manure used for soil fertilization is a reservoir for transferable antibiotic resistance plasmids. FEMS Microbiol. Ecol. 66:25-37.

Bryant D. 2003. A classification of consensus methods for phylogenetics. BioConsensus 163-183.

Boon N, Goris J, De Vos P, Verstraete W, Top EM. 2001. Genetic diversity among 3-chloroaniline- and aniline-degrading strains of the *Comamonadaceae*. Appl. Environ. Microbiol. 67: 1107-1115.

Chung Y, Ané C. 2011. Comparing two Bayesian methods of gene tree/species tree reconstructions: Simulations with incomplete lineage sorting and horizontal gene transfer. Syst. Biol. 60:261-275.

Dejonghe W, Goris J, Dierickx A, De Dobbeleer V, Crul K, De Vos P, Verstraete W, Top E. 2002. Diversity of 3-chloroaniline and 3,4-dichloroaniline degrading bacteria isolated from three different soils and involvement of their plasmids in chloroaniline degradation. FEMS Microbiol. Ecol. 42: 315-325.

Dröge M, Pühler A, Selbitschka W. 2000. Phenotypic and molecular characterization of conjugative antibiotic resistance plasmids isolated from bacterial communities of activated sludge. Mol. Gen. Genet. 263:471-482.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1-19.

Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783-791.

Fernandez-Lopez R, Garcillan-Barcia MP, Revilla C, Lazaro M, Vielva L, de la Cruz F. 2006. Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution. FEMS Microbiol. Rev. 30:942-966.

Götz A, Pukall R, Smit E, Tietze E, Prager R, Tschäpe H, van Elsasa JD, and Smalla K. 1996. Detection and characterization of broad-host-range plasmids in environmental bacteria by PCR. Appl. Environ. Microbiol. 62: 2621-2628.

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47:9-17.

Gstalder ME, Faelen M, Mine N, Top EM, Mergeay M, Couturier M. 2003. Replication functions of new broad host range plasmids isolated from polluted soils. Res. Microbiol. 154:499-509.

Guiney, D. G. 1993. Broad host range conjugative and mobilizable plasmids in Gram-negative bacteria.. In: Clewell DB, editor. Bacterial Conjugation, Plenum Publishing Corp., New York. p. 75-104

Haines AS, Akhtar P, Stephens ER, Jones K, Thomas CM, Perkins CD, Williams JR, Day MJ, Fry JC. 2006. Plasmids from freshwater environments capable of IncQ retrotransfer are diverse and include pQKH54, a new IncP-1 subgroup archetype. Microbiology 152:2689-2701.

Heinemann JS, Sprague GFJ. 1989. Bacterial conjugative plasmids mobilize DNA transfer between bacteria and yeast. Nature 340:205-209.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570-580.

Heuer H, Krögerrecklenfort E, Wellington EM, Egan S, van Elsas JD, van Overbeek L, Collard JM, Guillaume G, Karagouni AD, Nikolakopoulou TL, Smalla K. 2002. Gentamicin resistance genes in environmental bacteria: prevalence and transfer. FEMS Microbiol. Ecol. 42:289-302.

Heuer, H, Szczepanowski R, Schneiker S, Pühler A, Top EM, Schlüter A. 2004. The complete sequences of plasmids pB2 and pB3 provide evidence for a recent ancestor of the IncP-1ß group without any accessory genes. Microbiology 150: 3591-3599.

Ingram, LC, Richmond MH, Sykes RB. 1973. Molecular characterization of the R factors implicated in the carbenicillin resistance of a sequence of *Pseudomonas aeruginosa* strains isolated from burns. Antimicrob. Agents Chemother. 3:279-288.

Kluge AG. 2004. On total evidence: for the record. Cladistics 20:205-207.

Kreps S, Ferino F, Mosrin C, Gerits J, Mergeay M, Thuriaux P. 1990. Conjugative transfer and autonomous replication of a promiscuous IncQ plasmid in the cyanobacterium *Synechocystis* PCC 6803. Mol. Biol. Genet. 221:129-133.

Król, JE, Penrod JT, McCaslin H, Rogers LM, Yano H, Dejonghe W, Brown CJ, Parales RE, Wuertz S, Top EM. 2012. Genomic and functional analysis of the IncP-1β plasmids pNB8c and pWDL7::*rfp* explains their role in 3-chloroaniline catabolism. Appl. Environ. Microbiol. 78: 828-838.

Kubatko LS, Carstens BC, Knowles LL. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971-973.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542-2543.

Liu L, Yu L, Kubatko L, Pearl DK, Edwards SV. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320-328.

Maddison W. 1997. Gene trees in species trees. Syst. Biol. 46:523-536.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 26:2462-2463.

Mazodier P, Petter R, Thompson C. 1989. Intergeneric conjugation between *Escherichia coli* and *Streptomyces* species. J. Bacteriol. 171:3583-3585.

Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674-683.

Norberg P, Bergstrom M, Jethava V, Dubhashi D, Hermansson M. 2011. The IncP-1 plasmid backbone adapts to different host bacterial species and evolves through homologous recombination. Nat. Commun. 2:268.

Novais A, Canton R, Valverde A, Machado E, Galan JC, Peixe L, Carattoli A, Baquero F, Coque TM. 2006. Dissemination and persistence of *bla*CTX-M-9 are linked to class 1 integrons containing CR1 associated with defective transposon derivatives from Tn*402* located in early antibiotic resistance plasmids of IncHI2, IncP1-α, and IncFI groups. Antimicrob. Agents Chemother. 50:2741-2750

Pachulec E, van der Does C. 2010. Conjugative plasmids of *Neisseria gonorrhoeae*. PLoS One 5:e9962.

Pansegrau W, Lanka E, Barth PT, Figurski DH, Guiney DG, Haas D, Helinski DR, Schwab H, Stanisich VA, Thomas CM. 1994. Complete nucleotide sequence of Birmingham IncPα plasmids: Compilation and comparative analysis. J. Mol. Biol. 239:623-663.

Pollard DA, Iyer VN, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235-238.

Rhodes G, Parkhill J, Bird C, Ambrose K, Jones MC, Huys G, Swings J., Pickup RW. 2004. Complete nucleotide sequence of the conjugative tetracycline resistance plasmid pFBAOT6, a member of a group of IncU plasmids with global ubiquity. Appl. Environ. Microbiol. 70:7497-7510.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276-277.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Schlüter A, Heuer H, Szczepanowski R, Forney LJ, Thomas CM, Pühler A, Top EM. 2003. The 64,508 bp IncP-1β antibiotic multiresistance plasmid pB10 isolated from a waste-water treatment plant provides evidence for recombination between members of different branches of the IncP-1β group. Microbiology 149:3139-3153.

Schlüter A, Szczepanowski R, Pühler A, Top EM. 2007. Genomics of IncP-1 antibiotic resistance plasmids isolated from wastewater treatment plants provides evidence for a widely accessible drug resistance gene pool. FEMS Microbiol. Rev. 31:449-477.

Sen D, Yano H, Suzuki H, Krol JE, Rogers L, Brown CJ, Top EM. 2010. Comparative genomics of pAKD4, the prototype IncP-1δ plasmid with a complete backbone. Plasmid 63:98-107.

Sen, D, Van Der Auwera G, Rogers L, Thomas CM, Brown CJ, Top EM. 2011. Broad-host-range plasmids from agricultural soils have IncP-1 backbones with diverse accessory genes. Appl. Environ. Microbiol. 77: 7975-7983.

Shimodaira, H, Hasegawa, M. 1999. Multiple comparisons of log-likelihoods with application to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

Smalla, K, Haines AS, Jones K, Krögerrecklenfort E, Heuer H, Schloter M, Thomas CM. 2006. Increased abundance of IncP-1β plasmids and mercury resistance genes in mercury-polluted river sediments: first discovery of IncP-1β plasmids with a complex *mer* transposon as the sole accessory element. Appl. Environ. Microbiol. 72: 7253-7259.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688-2690.

Stolze, Y., F. Eikmeyer, D. Wibberg, G. Brandis, C. Karsten, I. Krahn, S. Schneiker-Bekel, P. Viehöver, A. Barsch, M. Keck, E.Top, K. Niehaus, and A. Schlüter. 2012. IncP-1β plasmids of *Comamonas* sp. and *Delftia* sp. strains isolated from a wastewater treatment plant mediate resistance to and decolorization of the triphenylmethane dye crystal violet. Microbiology 158: 2060-2072.

Sullivan J, Abdo Z, Joyce P, Swofford, DL. 2005. Evaluating the performance of a successive-approximations approach to parameter optimization in maximum-likelihood phylogeny estimation. Mol. Biol. Evol. 22:1386-1392.

Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (* and other methods). Version 4, Sinauer Associates, Sunderland (MA).

Thomas CM, Smith CA. 1987. Incompatibility group P plasmids: Genetics, evolution, and use in genetic manipulation. Ann. Rev. Microbiol. 41:77-101.

Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX.Curr. Protoc. Bioinformatics Chapter 2: Unit 2.3.

Thorsted PB, Macartney DP, Akhtar P et al. (12 coauthors). 1998. Complete sequence of the IncPβ plasmid R751: Implications for evolution and organisation of the IncP backbone. J. Mol. Biol. 282:969-990.

Top EM, Holben WE, Forney LJ. 1995. Characterization of diverse 2,4-dichlorophenoxyacetic acid-degradative plasmids isolated from soil by complementation. Appl. Environ. Microbiol. 61:1691-1698.

Van der Auwera GA, Krol JE, Suzuki H, Foster B, Van Houdt R, Brown CJ, Mergeay M, Top EM. 2009. Plasmids captured in *C. metallidurans* CH34: defining the PromA family of broad-host-range plasmids. Antonie van Leeuwenhoek 96:193-204.

Vedler E, Vahter M, Heinaru A. 2004. The completely sequenced plasmid pEST4011 contains a novel IncP1 backbone and a catabolic transposon harboring *tfd* genes for 2,4-dichlorophenoxyacetic acid degradation. J. Bacteriol. 186:7161-7174.

Yang Z, Goldman N, Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316-324.

Zhao F, Hou H, Bao Q, Wu J. 2009. PGA4genomics for comparative genome assembly based on genetic algorithm optimization. Genomics 94: 284–286.

Tables

Table 1. General features of the plasmids included in this study listed by subgroup.

Plasmids	IncP-1 subgroup	Origin/ Isolation method / Host[a]	Accession number[b]
pB5	IncP-1α	Municipal WWTP[c] in Germany, exogenous	CP002151
pBS228	IncP-1α	Wastewater of antibiotic factory in Russia, host unknown	NC_008357
pG527	IncP-1α	Pig manure, Germany, exogenous	JX469830
pSP21	IncP-1α	Municipal WWTP in Germany, exogenous	CP002153
pTB11	IncP-1α	WWTP in Germany, exogenous	NC_006352
pWEC911	IncP-1α	Sugar beet rhizosphere in U.K., exogenous	JX469833
RK2	IncP-1α	Hospital in U.K., *Pseudomonas aeruginosa*, *Klebsiella aerogenes*	NC_001621
pA1	IncP-1β	Soil in Japan, *Sphingomonas sp.* A1	NC_007353
pA81	IncP-1β	PCB contaminated soil in Czech Republic, *Achromobacter xylosoxidans* A8	AJ515144
pADP-1	IncP-1β	Soil in U.S.A, *Pseudomonas sp.* ADP	NC_004956
pAKD1	IncP-1β	Agricultural soil in Norway, exogenous	JN106164
pAKD18	IncP-1β	Agricultural soil in Norway, exogenous	JN106169
pAKD26	IncP-1β	Agricultural soil in Norway, exogenous	JN106171
pAMMD1	IncP-1β	Pea rhizosphere in U.S.A., *Burkholderia ambifaria* AMMD	NC_008385

pAOVO02	IncP-1β	Polluted soil in U.S.A., *Acidovorax sp*. JS42	NC_008766
pB1	IncP-1β	Municipal WWTP in Germany, exogenous	JX469829
pB3	IncP-1β	Municipal WWTP in Germany, exogenous	NC_006388
pB4	IncP-1β	Municipal WWTP in Germany, exogenous	AJ431260
pB8	IncP-1β	Municipal WWTP in Germany, exogenous	NC_007502
pB10	IncP-1β	Municipal WWTP in Germany, exogenous	NC_004840
pB12	IncP-1β	Municipal WWTP in Germany, exogenous	JX469826
pBP136	IncP-1β	Diseased whooping cough patient in Japan, *Bordetella pertussis* BP136	NC_008459
pC11	IncP-1β	Municipal WWTP in Germany, *Delftia acidovorans* C1	HQ891317
pCNB1	IncP-1β	Industrial WWTP in China, *Comamonas sp.* CNB-1	NC_010935
pDS3	IncP-1β	Creek in U.S.A., exogenous	JX469834
pKS212	IncP-1β	Hospital WWTP in Belgium, exogenous	JX469831
pKV29	IncP-1β	Municipal WWTP in Germany, *Delftia sp.* KV29	JN648090
pNB1	IncP-1β	Orchard soil in Belgium *Delftia acidovorans* LME1	JF274988
pRSB222	IncP-1β	Municipal WWTP in Germany, exogenous	JX469824 JX469825
pRSB223	IncP-1β	Municipal WWTP in Germany, exogenous	JX469828
pTB30	IncP-1β	Agricultural soil in Belgium, *Comamonas testosteroni* TB30	JF274987

pTP6	IncP-1β	Contaminated river sediments in Kazakhstan, exogenous	NC_007680
pUO1	IncP-1β	Industrial WWTP in Japan, *Delftia acidovorans* B	NC_005088
pWDL7	IncP-1β	Orchard soil in Belgium, *Comamonas testosteroni* WDL7	GQ495894
pYS1	IncP-1β	Polluted soil in Japan, *Burkholderia cepacia*	JX469832
R751	IncP-1β	Hospital in UK, *Klebsiella aerogenes*	NC_001735
pKS208	IncP-1γ	Hospital WWTP in Belgium, exogenous	JQ432564
pMBUI1	IncP-1γ	University of Idaho Arboretum Pond, exogenous	JQ432563
pQKH54	IncP-1γ	River in U.K., exogenous	NC_008055
pAKD4	IncP-1δ	Agricultural soil in Norway, exogenous	GQ983559
pAKD16	IncP-1ε	Agricultural soil in Norway, exogenous	JN106167
pAKD25	IncP-1ε	Agricultural soil in Norway, exogenous	JN106170
pEMT3	IncP-1ε	Agricultural soil in U.S.A, exogenous	JX469827
pHH128	IncP-1ε	Manured soil in Germany, exogenous	JQ004406
pHH3414	IncP-1ε	Manured soil in Germany, exogenous	JQ004408
pKJK5	IncP-1ε	Manured soil in Denmark, exogenous	NC_008272

[a] The original hosts of plasmids captured by exogenous isolation are not known.
[b] See Materials and Methods for references to studies that described the plasmids whose sequences were not previously published. All previously published plasmids are only referred to here by their RefSeq or Genbank/EMBL/DDBJ accession numbers.
[c] WWTP, wastewater treatment plant.

Table 2. Nucleotide substitution models[a] chosen for the 28 genes.

Genes	Models selected for ML analyses	Models selected for Bayesian analyses
Concatenated Data	1st codon: GTR+Γ 2nd codon: GTR+Γ 3rd codon: GTR+Γ	1st codon: GTR I+ Γ 2nd codon: GTR+Γ 3rd codon: GTR I+ Γ
trfA2	HKY+Γ	HKY+Γ
trbA	K81uf+Γ	GTR+ Γ
trbB	TrN+Γ	GTR+ Γ
trbC	HKY+Γ	HKY+Γ
trbD	HKY+I	HKY+I
trbF	TrN+Γ	GTR+ Γ
trbG	TrN+I	GTR+I
trbI	TrN+I+Γ	GTR+I+ Γ
trbJ	TrN+I+Γ	GTR+I+ Γ
trbK	TrN+Γ	GTR+ Γ
traD	TrN+I	GTR+I
traE	GTR+ Γ	GTR+ Γ
traF	HKY+Γ	HKY+Γ
traG	TIM+ I+Γ	GTR+I+ Γ
traH	GTR+ Γ	GTR+ Γ
traI	TrN+I+Γ	GTR+I+ Γ
traJ	HKY+Γ	HKY+Γ
traK	TVM+Γ	GTR+ Γ
traL	TVM+Γ	GTR+ Γ
kfrC	HKY+Γ	HKY+Γ
kfrB	HKY+Γ	HKY+Γ
kfrA	TrN+Γ	GTR+ Γ
korB	TrN+I+Γ	GTR+I+ Γ
incC	HKY+Γ	HKY+Γ

korA	HKY+Γ	HKY+Γ
kleE	TVM+Γ	GTR+ Γ
korC	TVM+Γ	GTR+ Γ
klcA	HKY+I+Γ	HKY+I+Γ

[a] HKY: variable base frequencies, different transition and transversion rates; K81uf: variable base frequencies, three substitution rates; TrN: variable base frequencies, equal transversion rates, variable transition rates; TVM: variable base frequencies, equal transition rates, variable transversion rates; TIM: variable base frequencies, variable transition rates, two transversion rates; GTR: variable base frequencies, six substitution rates; Γ: gamma distributed rate variation among sites; I: proportion of unchanging sites.

Table 3. Results of SH-tests. Each gene tree was compared to the concatenated tree.

Gene	Diff -ln L	P
klcA	409.04842	0.000*
kfrC	40.94226	0.404
kfrB	63.47913	0.154
kfrA	127.54953	0.179
incc1	237.69133	0.003*
trfA2	163.54237	0.023*
trbK	39.36975	0.346
trbJ	535.70115	0.000*
trbI	408.5485	0.002*
trbG	230.63018	0.030*
trbF	351.9579	0.000*
trbD	37.1931	0.318
trbC	78.0097	0.141
trbB	700.09528	0.000*
trbA	133.64327	0.014*
traL	132.82945	0.051
traK	210.90414	0.006*

traJ	31.30619	0.417
traI	162.64551	0.158
traH	52.51454	0.226
traG	159.0419	0.105
traF	87.27873	0.116
traE	254.17049	0.054
traD	196.59165	0.009*
korC	340.93166	0.000*
korB	113.80035	0.196
kleE	187.10069	0.000*
korA	110.91598	0.002*

Figure Legends

Figure 1. Genetic map of a typical IncP-1 plasmid showing the different functional modules: region involved in initiating replication, composed of origin of replication (*oriV*) and replication initiation gene (*trfA*); *trb*, involved in mating bridge formation during conjugation; *tra,* involved in DNA processing for transfer during conjugation; *ctl,* also called central control region, is composed of regulatory genes and involved in maintaining plasmid stability; accessory regions are composed of host-beneficial genes. Genes that were included in this study are colored dark grey. Numerals inside the circle indicate tree topologies that were shared by several genes (topologies 1-4) or unique to one gene (U); topologies were inferred in this study by maximum likelihood (see Fig. 2).

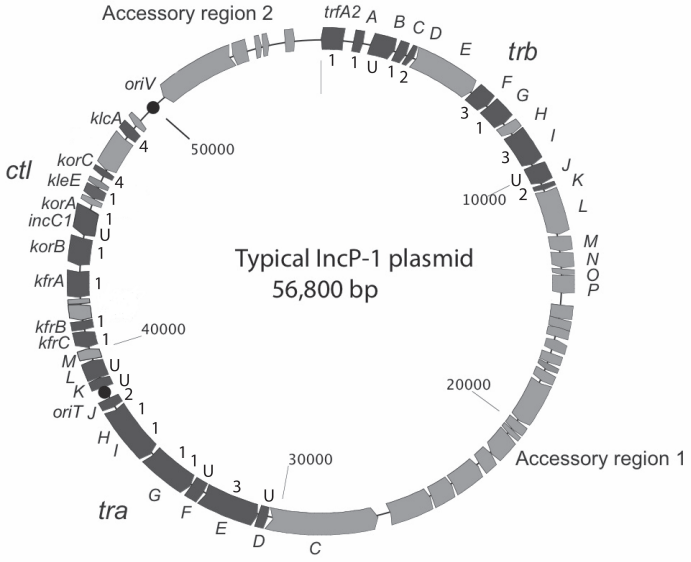
Figure 2. Cladograms showing four topologies produced by 21 gene trees. A) topology 1: supported by 46% of gene trees, namely, those of *trfA2*, *trbA*, *trbC*, *trbG*, *traG*, *traH*, *traI*, *kfrA*, *kfrB*, *kfrC*, *korB*, *korA* and *kleE*. B) topology 2: supported by gene trees of *trbD*, *trbK* and *traJ*. C) topology 3: supported by gene trees of *trbF*, *trbI* and *traE*. C) topology 4: supported by gene trees of *korC* and *klcA*. Trees were rooted using IncP-1γ as outgroup.

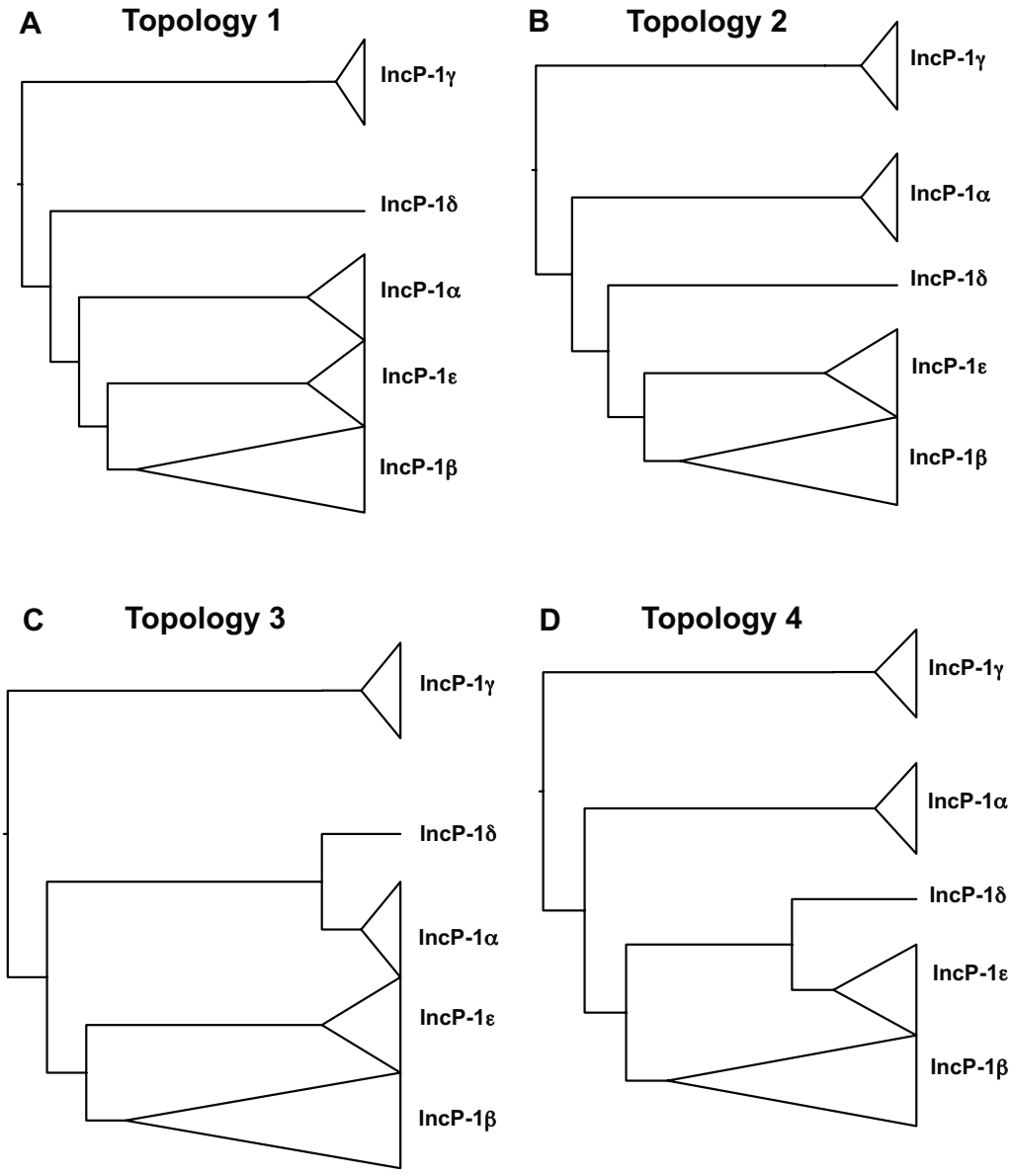
Figure 3. ML tree of concatenated data estimated from a partitioned analysis based on codon position. Nodal support is shown as non-parametric ML bootstrap values. The tree was rooted using IncP-1γ as outgroup.

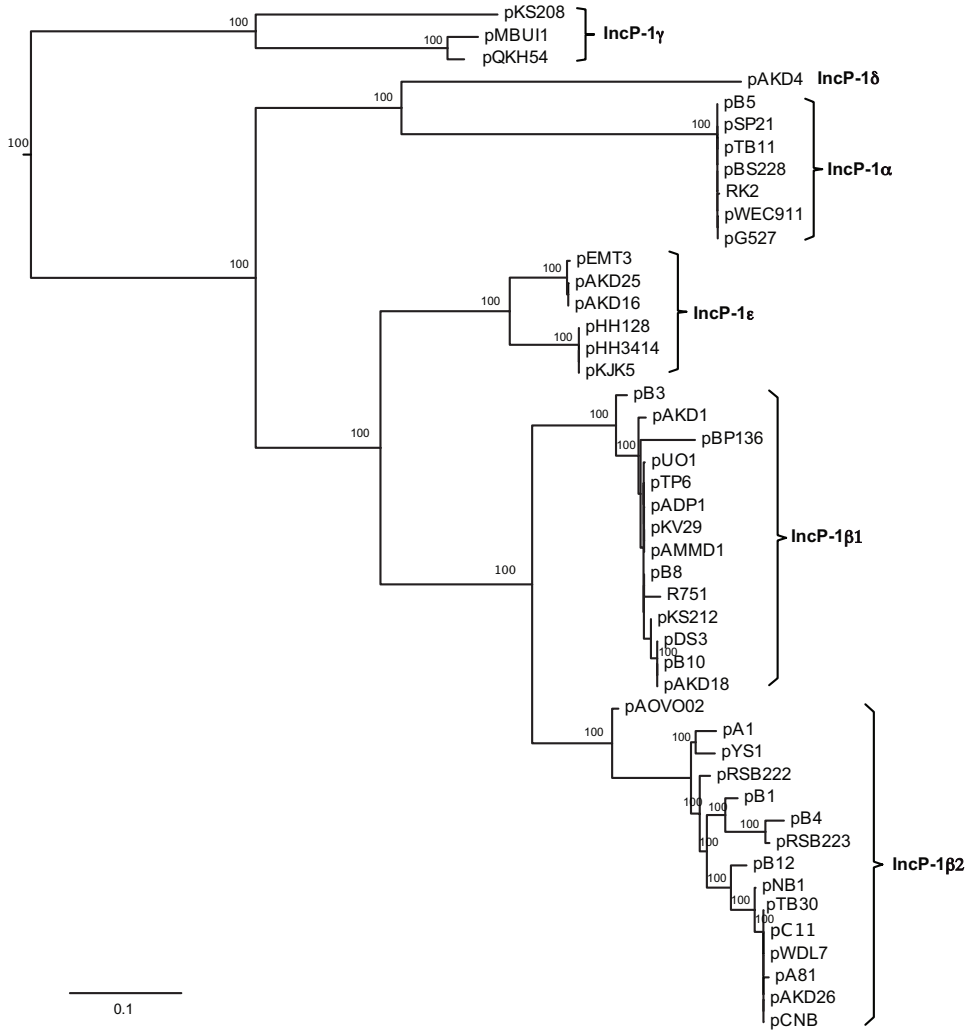
Figure 4. Congruence test for *trbB*. A) ML tree for *trbB*, ML_{trbB}, B) ML tree constrained to fit concatenated tree, ML_{Hyp}, C) Null distribution and test statistic δ [lnL (ML_{hyp}) - lnL (ML_{trbB})] = 651.6. P-value of obtaining a test statistic higher than 651.6 is < 0.01. Panels D-F) Re-evaluated difference between ML_{trbB} and ML_{hyp} after removing 14 plasmids (from top to bottom in panel A: two IncP-1β plasmids pB3 and pAOVO02, all six IncP-1ε plasmids pAKD16, pAKD25, pEMT3, pHH128, pHH3414, and pKJK5, the IncP-1δ plasmid pAKD4, IncP-1β plasmids pB4, pB12, pA1, pRSB222 and pYS1).

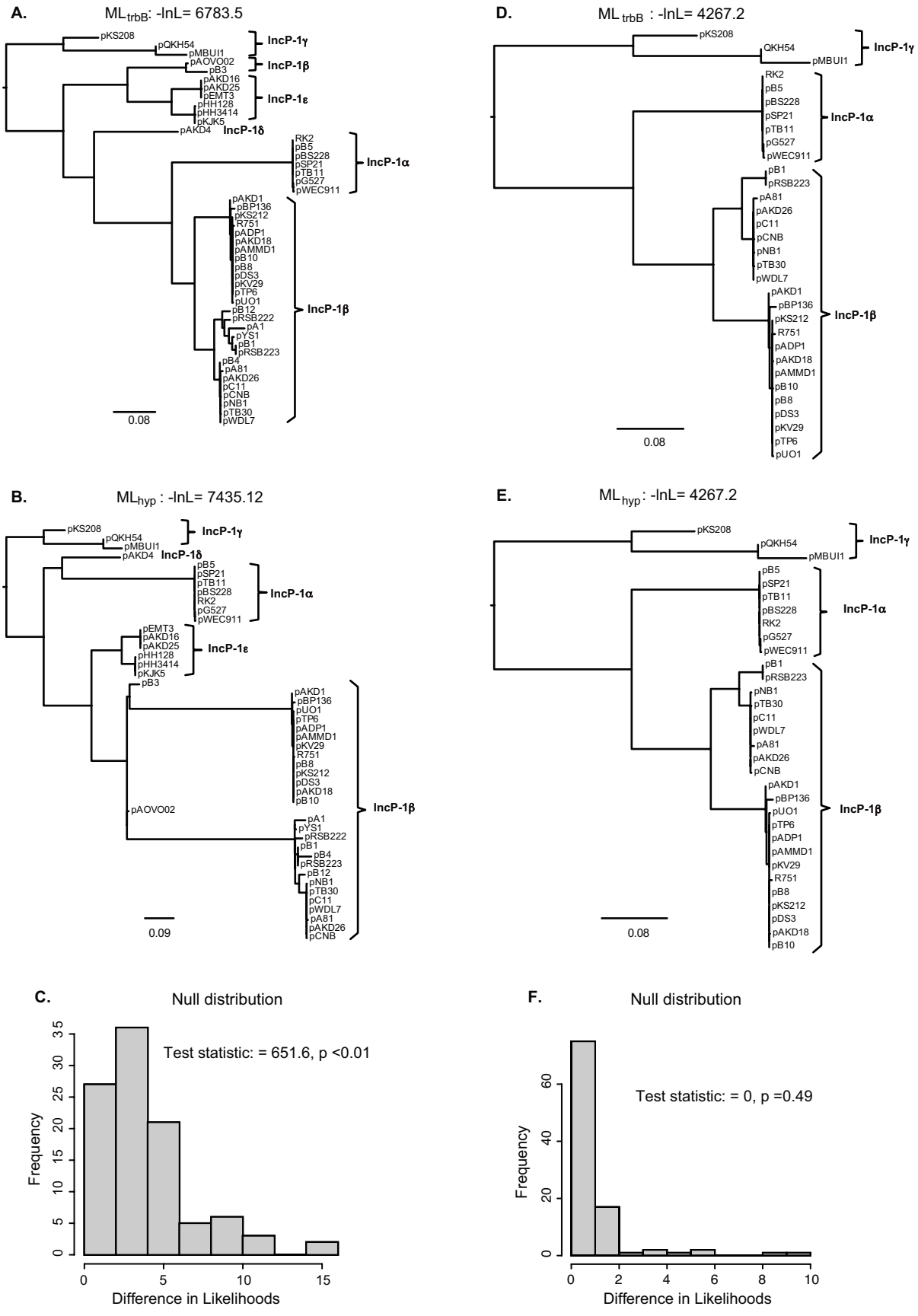
Figure 5. A) Species tree estimated as maximum clade credibility tree by *BEAST. B) Species tree estimated as maximum clade credibility tree by *BEAST after removal of recombinant gene *traE*. Nodal support is shown as posterior probabilities.

Figure S1. ML trees of 28 backbone genes.

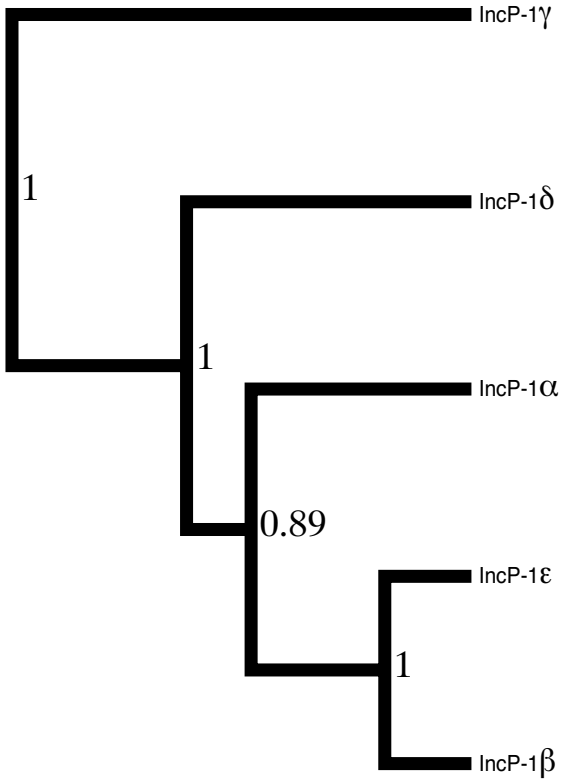

Accessory region 2
trfA2
A
B
C
D
E
trb
F
G
H
I
J
K
L
M
N
O
P
Accessory region 1
C
D
E
F
G
tra
H
I
J
K
L
M
oriT
incC1
korA
kleE
korC
ctl
klcA
oriV
kfrA
korB
kfrB
kfrC
50000
10000
20000
30000
40000
Typical IncP-1 plasmid
56,800 bp


A
Topology 1
IncP-1γ
IncP-1δ
IncP-1α
IncP-1ε
IncP-1β
B
Topology 2
IncP-1γ
IncP-1α
IncP-1δ
IncP-1ε
IncP-1β
C
Topology 3
IncP-1γ
IncP-1δ
IncP-1α
IncP-1ε
IncP-1β
D
Topology 4
IncP-1γ
IncP-1α
IncP-1δ
IncP-1ε
IncP-1β

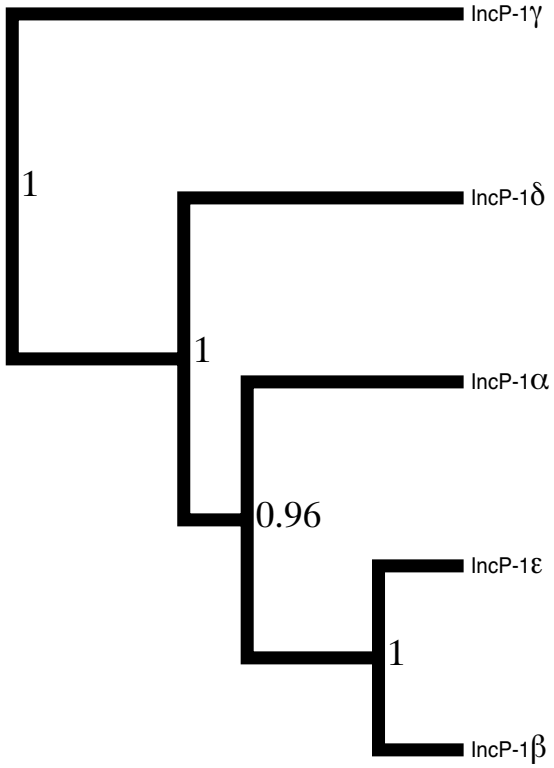

pKS208
pMBUI1
pQKH54
IncP-1γ
pAKD4
IncP-1δ
pB5
pSP21
pTB11
pBS228
RK2
pWEC911
pG527
IncP-1α
pEMT3
pAKD25
pAKD16
pHH128
pHH3414
pKJK5
IncP-1ε
pB3
pAKD1
pBP136
pUO1
pTP6
pADP1
pKV29
pAMMD1
pB8
R751
pKS212
pDS3
pB10
pAKD18
IncP-1β1
pAOVO02
pA1
pYS1
pRSB222
pB1
pB4
pRSB223
pB12
pNB1
pTB30
pC11
pWDL7
pA81
pAKD26
pCNB
IncP-1β2
0.1


A.
ML trbB: -lnL= 6783.5
pKS208
pQKH54
pMBUI1
pAOVO02
pB3
pAKD16
pAKD25
pEMT3
pHH128
pHH3414
pKJK5
pAKD4
IncP-1γ
IncP-1β
IncP-1ε
IncP-1δ
RK2
pB5
pBS228
pSP21
pTB11
pG527
pWEC911
IncP-1α
pAKD1
pBP136
pKS212
R751
pADP1
pAKD18
pAMMD1
pB10
pB8
pDS3
pKV29
pTP6
pUO1
pB12
pRSB222
pA1
pYS1
pB1
pRSB223
pB4
pA81
pAKD26
pC11
pCNB
pNB1
pTB30
pWDL7
IncP-1β
0.08
D.
ML trbB : -lnL= 4267.2
pKS208
QKH54
pMBUI1
IncP-1γ
RK2
pB5
pBS228
pSP21
pTB11
pG527
pWEC911
IncP-1α
pB1
pRSB223
pA81
pAKD26
pC11
pCNB
pNB1
pTB30
pWDL7
pAKD1
pBP136
pKS212
R751
pADP1
pAKD18
pAMMD1
pB10
pB8
pDS3
pKV29
pTP6
pUO1
IncP-1β
0.08
B.
ML hyp : -lnL= 7435.12
pKS208
pQKH54
pMBUI1
pAKD4
IncP-1γ
IncP-1δ
pB5
pSP21
pTB11
pBS228
RK2
pG527
pWEC911
IncP-1α
pEMT3
pAKD16
pAKD25
pHH128
pHH3414
pKJK5
IncP-1ε
pB3
pAKD1
pBP136
pUO1
pTP6
pADP1
pAMMD1
pKV29
R751
pB8
pKS212
pDS3
pAKD18
pB10
pAOVO02
IncP-1β
pA1
pYS1
pRSB222
pB1
pB4
pRSB223
pB12
pNB1
pTB30
pC11
pWDL7
pA81
pAKD26
pCNB
0.09
E.
ML hyp: -lnL= 4267.2
pKS208
pQKH54
pMBUI1
IncP-1γ
pB5
pSP21
pTB11
pBS228
RK2
pG527
pWEC911
IncP-1α
pB1
pRSB223
pNB1
pTB30
pC11
pWDL7
pA81
pAKD26
pCNB
pAKD1
pBP136
pUO1
pTP6
pADP1
pAMMD1
pKV29
R751
pB8
pKS212
pDS3
pAKD18
pB10
IncP-1β
0.08
C.
Null distribution
Test statistic: = 651.6, p <0.01
Frequency
0
5
10
15
20
25
30
35
0
5
10
15
Difference in Likelihoods
F.
Null distribution
Test statistic: = 0, p =0.49
Frequency
0
20
40
60
0
2
4
6
8
10
Difference in Likelihoods

A


IncP-1γ
1
IncP-1δ
1
IncP-1α
0.89
IncP-1ε
1
IncP-1β

B


IncP-1γ
1
IncP-1δ
1
IncP-1α
0.96
IncP-1ε
1
IncP-1β