INVESTIGATION OF EVOLUTIONARY RELATIONSHIPS OF PLASMIDS BY COMBINGING SEQUENCE ALIGNMENTS WITH GENOME SIGNATURE

by

MEGAN PATCH

(Under the Direction of Jan Mrázek)

ABSTRACT

Current DNA comparisons focus on single gene alignments to reconstruct a phylogenetic tree. Standard phylogenetic methods based on sequence alignments are not always applicable to plasmids because many plasmids do not share homologous genes. To overcome this limitation, sequence alignment methods are combined with genome signature, facilitating the comparison of any DNA sequences regardless of the presence of homologous genes. Nucleotide alignment, amino acid alignment, and genome signature are integrated to combine vertical evolutionary history, which reflects shared ancestry with horizontal evolutionary, which reflects shared hosts.

INDEX WORDS: Genome Signature, Phylogeny, Plasmid, DNA Sequence Comparison, Evolutionary Tree

INVESTIGATION OF EVOLUTIONARY RELATIONSHIPS OF PLASMIDS BY COMBINGING SEQUENCE ALIGNMENTS WITH GENOME SIGNATURE

by

MEGAN PATCH

B.A. Luther College, 2008

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2010

INVESTIGATION OF EVOLUTIONARY RELATIONSHIPS OF PLASMIDS BY COMBINGING SEQUENCE ALIGNMENTS WITH GENOME SIGNATURE

by

MEGAN PATCH

Major Professor:	Jan Mrázek
Committee:	William Whitman
	Timothy Hoover

Electronic Version Approved:

Marueen Grasso
Dean of the Graduate School
The University of Georgia
July 2010

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION

Genome comparison methods are expanding to keep up with the growing volume of sequenced DNA. Faster, more reliable methods are needed to understand the wealth of data and interpret the sequences. Evolutionary studies have focused on evaluating genomes by comparing homologous sequences within and between taxonomical groups. The focus of these comparisons fell on rRNA sequences, which are universally present in Eukaryotes, Bacteria, and Archaea (Woese 1987). rRNA genes are relatively stable over generations and have been used to reflect the evolutionary history of the entire chromosome by using the representative rRNA phylogenetic tree as a representative organismal phylogenetic tree. Unfortunately, rRNA comparisons show the evolution of that particular gene and can only suggest evolutionary changes for the rest of the genome.

Single gene comparisons gave way to comparisons involving collection of genes. Relative gene content similarity has been used to provide resolution lost at the deep-rooted branches (Fitz-Gibbon and House 1999; Snel et al. 1999). This gene-centered model uses families of protein-encoding genes present in a group of genomes, requiring all genomes in the set to share some homologous sequences. This method has revealed evolutionary relationships that are not apparent from 16S rRNA gene trees, specifically the relationship of *Archaeoglobus* to the methanogens (Fitz-Gibbon and House 1999).

Average nucleotide identity in alignable segments of the genome has been used to compare complete genomes (Konstantinidis and Tiedje 2005). Again, this method requires parts

of the genome to be alignable, thus necessitating homologous segments. Another accepted comparison method evaluates the order of homologous genes (Sankoff et al. 1992), in which genomes are compared by calculating the necessary changes to get from one genome to the other, in terms of deletions, insertions, transpositions, and inversions.

Supertrees have been gaining focus as a tool to overcome some drawbacks of single-gene phylogeny. Supertrees are derived from individual phylogenetic trees. Each tree is constructed from various single gene sequence alignments using the same set of reference DNA sequences. Genes representing each individual phylogenetic tree do not need to be found in every genome, plasmid, bacteriophage, or virus that is being compared. The final supertree is constructed from overlapping individual trees. Supertrees overcome some problems of standard phylogeny methods because they are not dependent on each sequence containing every gene used to construct the tree.

With the increase in fully sequenced genomes, whole genome comparisons became more readily available. Nucleotide and amino acid identity are accurate methods for genome comparison, however, sequence identity is not always applicable when homologous sequences are not present, for instance, plasmid comparisons. NUCmer (Delcher et al. 2002) is a fast comparison method used to find matching nucleotide segments in a pair of DNA sequences. PROmer translates all six reading frames to hypothetical protein sequences and compares the translated amino acid sequences using a similar algorithm as NUCmer. PROmer compares amino acid sequences rendering it insensitive to synonymous mutations and therefore reflects more distant sequence similarities.

Genome signatures have been utilized as a means to compare DNA sequences without the reliance of homologous genes (Campbell, A. et al. 1999). Genome signature assesses the

dinucleotide relative abundances in a DNA sequence. Given that closely related sequences tend to have a similar dinucleotide relative abundance, as shown by Campbell et al. (1999), the genome signature difference between closely related sequences should be small. Genome signature can be used to compare any DNA sequences, regardless of the presence of homologous genes. Genome signature is less accurate than sequence alignment comparisons due to possible noise and artifacts, however, and these sequence comparisons have a limited available use. One major problem is the occurrence of a small difference in genome signature between two distantly related sequences (Mrázek 2009). Also, measurement of dinucleotide frequency may be too general to distinguish between distantly related sequences.

Genome signature utilizes oligonucleotide relative abundance differences when comparing DNA sequences. Oligonucleotide relative abundance is a measure of DNA composition which tends to be consistent throughout a genome (Campbell et al. 1999). Multiple studies suggest potential for organism specific preferential nucleotide replacement during replication and repair (Echols and Goodman 1991). It is therefore likely that integrations into a genome will gradually develop similar oligonucleotide relative abundance to the native genome due to biases in the replication and repair machinery of the cell. Because mobile genetic elements do not necessarily contain specific replication and repair machinery, acquired plasmids may be modified to reflect the genome signature of the host.

Mahalanobis distance has been suggested to improve the accuracy of linking a plasmid with its natural host by genome signature comparison between the plasmid and alternative prokaryotic chromosomes. Mahalanobis distance is a statistical technique of multivariate analysis which evaluates the variance of genome signature along the chromosome and assigns more weight to the most informative dinucleotides (Suzuki et al. 2008).

Single gene, whole genome, and genome signature comparisons have limitations. Comparisons involving mobile genetic elements are not as informative with single or multiple gene comparisons. While genome signature comparisons can be utilized by any DNA sequence set, they are less accurate than the traditional single or multiple gene comparisons. Combining the advantages of sequence alignments with those of genome signature allows for the elucidation of closely related sequences with the ability to distinguish between distantly related sequences.

Mobile genetic elements are important contributors to the "universal tree of life." Plasmids are non-essential, extrachromosomal DNA which can benefit the host by providing antibiotic resistance (Cohen et al. 1972; Foster 1983), metal resistance (Novick 1968), symbiotic nitrogen fixation (Frost et al. 2005) or supply other genes to improve the fitness of the host (Novick 1969). Plasmids are a major player in lateral gene transfer, facilitating the movement of genes between hosts. Lateral gene transfer is a key component in the evolution of both plasmids and genomic hosts, and genome comparisons are not complete without it. Understanding the effects of lateral gene transfer will aid in the understanding of the evolutionary path of prokaryotes as well phenotypic variation. Lateral gene transfer adds to the diversity and adaptability of prokaryotes.

Viruses and bacteriophages, like plasmids, are key players in lateral gene transfer. These genetic elements are ubiquitous in nature and rely on the host cell for replication. Lateral gene transfer accounts for a substantial amount of the bacterial genomic DNA (Canchaya et al. 2003). In fact, many areas of the genome are labeled as "phage associated genes" when no homologous genes in bacterial sequences are found and this genomic sequence is distinctive of phage DNA. This is especially true in lysogenic bacteriophages that have integrated their genome into the host chromosome (Canchaya et al. 2003). The phage genome can also include genes acquired from

previous bacterial host genomes and phage genome integration into a new bacterial chromosome effectively transfers bacterial DNA between bacterial hosts.

Genome sequence comparisons have allowed us to study evolutionary, ecological and other relationships between organisms. A major goal of genome comparison is to create a "universal tree of life," elucidating the evolutionary relationships of all organisms on Earth (Woese 1987). These comparisons have focused on the chromosome and less on other genetic elements including plasmids, bacteriophages, and viruses, which have been left off universal trees (Brussow 2009). These mobile genetic elements contribute to lateral gene transfer and are therefore important in the evolution of genomes.

Development of a method to compare any DNA sequence regardless of similarity is the main goal of this work. The resulting comparison method needs to be useful for all DNA sequences, especially those with no sequence homology. Subsequently, the construction of a "universal tree of life" using this new comparison method will help to understand the connections of the tree of life at the more deep-rooted branches, allowing for analysis of more distantly related sequences. Understanding the microbial evolutionary past may give insight into predicting the fate of genomes and other genetic elements, as well as the function of genomic features.

Multiple sequence comparison methods can be combined to give resolution to the output. In the case of the little or no sequence similarity between DNA sequences, genome signature is added to piece together sequences that do not share sequence identity. Distances derived from NUCmer, PROmer, and genome signature comparisons use weights that emphasize NUCmer results for closely related sequences, PROmer for less closely related sequences, and genome signature for sequences with little or no recognizable homology and is visualized by a

dendrogram. The versatility of this integrated method, provided by the layered comparisons, allows it to be used for sequences other than genomes, including plasmids, viruses and phages.

CHAPTER 2

INTEGRATED METHOD FOR COMPARING DNA SEQUENCES

BACKGROUND

Standard techniques for phylogenetic reconstruction involve building phylogenetic trees from aligned homologous protein or DNA sequences. Such techniques are appropriate for analyses of individual gene or protein families but face limitations in elucidating relationships among whole genomes or species. Phylogenetic relationships among species are often derived from rRNA gene sequences, which are universally present in Eukaryotes, Bacteria and Archaea (Olsen and Woese 1993). New techniques for assessing relationships among different organisms emerged with availability of complete genomes, including measuring similarity of gene content (Fitz-Gibbon and House 1999; Snel et al. 1999), conservation of gene order (Belda et al. 2005; Korbel et al. 2002), or average nucleotide identity in alignable segments of the genome (Konstantinidis and Tiedje 2005). Still, all these methods require identification of orthologous genes in the compared genomes. Moreover, interpretations of the resulting trees with respect to phylogeny can be less straightforward because, for example, similarity of gene content can relate to metabolic and environmental similarities between organisms rather than directly to their phylogenetic relationship.

Eukaryotes, Bacteria and Archaea all share at least some homologous genes, which, although not without uncertainties and caveats, allows construction of a "universal tree of life" using some of the alignment-based methods described above (Ciccarelli et al. 2006; Doolittle 1999). However, the need to identify orthologous sequences makes the same techniques

inapplicable for similar universal comparisons among plasmids or viruses, which often share little or no homologous sequences. The need to identify homologous sequences is bypassed in genome signature comparisons (Karlin and Burge 1995). The concept of genome signature refers to a relative invariance of some compositional characteristics of DNA sequences (typically arrays of oligonucleotide frequencies, sometimes normalized to factor out differences in frequencies of the embedded shorter oligonucleotides) within the same genome or between closely related genomes. The genome signature comparisons are universally applicable to DNA sequences because they do not require that the sequences are homologous. On the other hand, noise and possibility of convergent evolution of genome signature make genome signature-based methods far less accurate as tools for phylogenetic reconstruction than alignment-based methods (Mrázek 2009).

The goal of this work was to design a method to build a universal tree for prokaryotic plasmids. The unique challenge of this task is in the need to combine the alignment-based methods for accurate classification of closely related plasmids and a genome signature-based comparison for clustering branches beyond the range of recognizable sequence homology. The resulting dendrogram is not a phylogenetic tree in the strict sense, that is, it does not reflect the levels of evolutionary divergence from a single common ancestor. This only applies for the parts of the dendrogram where direct sequence homology determines the clustering of the nodes. Similarity of genome signature does not necessarily indicate that the plasmids belong to the same lineage but genome signatures of plasmids were shown to be at least moderately similar to those of their host chromosomes (Campbell et al. 1999; Suzuki et al. 2008). Consequently, the clustering at the lower levels of the tree, indicating sequence similarity, is more likely to reflect

the evolutionary history of the plasmids on the same branch with respect to residing in the same host or related hosts.

Implementation of this integrated method produces a dendrogram which generally follows the standard taxonomical classifications of the hosts at the level of phylum. The combination of homology-based and composition-based methods facilitates comparisons at a wide range of sequence similarity and allows examining relationships, including identification of possibly non-native plasmids and/or plasmids with a broad host range

METHODS

DNA Sequences

Plasmid and chromosome sequences used in this study are listed in Table 1. Sequences were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/) and IMG (http://img.jgi.doe.gov) databases. Only completely sequenced plasmids over 50 kb length were used.

Sequence Comparison

Pairwise sequence similarity was assessed in terms of distances derived from NUCmer, PROmer (Kurtz et al. 2004), and genome signature comparisons (δ*-differences) (Campbell et al. 1999). NUCmer and PROmer are extensions of MUMmer, a fast sequence comparison tool that finds exactly matching nucleotide segments between two DNA sequences. NUCmer joins consecutive exact matches into larger homologous segments allowing for nucleotide substitutions, insertions and deletions. PROmer is analogous to NUCmer but compares hypothetical amino acid sequences translated from all six reading frames, rendering the comparison insensitive to synonymous substitutions and therefore detects more distant sequence similarity.

NUCmer and PROmer outputs are used to assess pairwise distances among all analyzed sequences. The NUCmer distance between sequences a and b is defined as:

$$D_{NUC}(a,b) = \left[1 - \max(S_a / L_a, S_b / L_b)\right] \cdot 10000$$

where L_a and L_b are the lengths of the sequences a and b, respectively. S_a and S_b are the aggregate lengths of segments in a and b that have a NUCmer match somewhere in the other sequence. PROmer distance, $D_{PRO}(a,b)$, is defined analogously but with matching segments identified by PROmer. NUCmer and PROmer distances can be interpreted as a fraction of one sequence (generally the shorter one) that is homologous to the other sequence in the pairwise comparison. User-defined parameters and the choice of the program (NUCmer or PROmer) determine the level of sequence similarity required for sequences to be considered homologous.

A third pairwise distance is determined from comparisons of dinucleotide relative abundances between the two sequences (genome signature, or δ*-differences) (Campbell et al. 1999). A dinucleotide relative abundance of a dinucleotide XY is determined by:

$$\rho^*_{XY} = f^*_{XY} / f^*_X f^*_Y$$

where f^*_X is the frequency of the nucleotide *X* and f^*_{XY} is the frequency of the dinucleotide *XY* assessed from both DNA strands (Campbell et al. 1999). The genome signature difference (δ*-difference) between sequences *f* and *g* is defined as:

$$\delta^*(f,g) = 1/16 \sum \left| \rho^*_{XY}(f) - \rho^*_{XY}(g) \right|$$

where the sum extends over all 16 dinucleotides (Campbell et al. 1999). The genome signature distance between two chromosomes or plasmids *a* and *b* is determined as an average distance between all pairs of nonoverlapping segments of a given length l_0 from different DNA sequences. This approach makes the assessment of genome signature distance less sensitive to

intragenomic sequence heterogeneity and artifacts from comparing sequences of vastly different sizes (e.g., plasmids and chromosomes) than using dinucleotide relative abundances evaluated from the whole genomes.

The three distances, NUCmer, PROmer, and δ*-differences, allow comparisons among sequences at different levels of similarity: NUCmer for highly similar sequences at the nucleotide level, PROmer for less similar sequences insensitive to synonymous substitutions, and δ*-differences for sequences that do not share recognizable homologous segments. These three distances, labeled D_{NUC}, D_{PRO}, and D_{δ}, respectively, are combined into a single metric $D(a,b)$ using the following formula:

$$D(a,b) = \begin{cases} D_{NUC} & \text{if} \quad D_{NUC} \le C_1 \\ D_{NUC}\left(\dfrac{C_2 - D_{NUC}}{C_2 - C_1}\right) + D_{PRO}\left(\dfrac{D_{NUC} - C_1}{C_2 - C_1}\right) & \text{if} \quad D_{NUC} > C_1 \ \text{and} \ D_{PRO} < C_2 \\ D_{PRO} & \text{if} \quad D_{PRO} \ge C_2 \ \text{and} \ D_{PRO} \le C_3 \\ D_{PRO}\left(\dfrac{10000 - D_{PRO}}{10000 - C_3}\right) + D_{\delta}\left(\dfrac{D_{PRO} - C_3}{10000 - C_3}\right) & \text{if} \quad D_{PRO} > C_3 \end{cases}$$

where $D_{\delta}(a,b) = C_3 + \kappa\delta^*(a,b)$ and $\delta^*(a,b)$ is the average δ^*-difference between pairs of nonoverlapping segments from sequences a and b of length l_0. After extensive testing, the parameters were set as C_1=3000, C_2 = 5000, C_3 = 8000, κ = 40, and l_0 = 20kb. Pairwise distances among all compared sequences are organized in a distance matrix, which is subsequently converted to a dendrogram using the Unweighted Pair Group Method with Arithmatic mean (UPGMA).

Mahalanobis Distance

Malahanobis distance was used for confirmation of the correct placement of plasmids clustering outside of their expected taxonomical branch by comparing the genome signatures of the plasmids with all chromosomes to determine if those clustering near each other had the smallest Malahanobis distance (Suzuki et al. 2008).

We adopted the method proposed by Suzuki et al. (Suzuki et al. 2008) with minor modifications. When comparing a plasmid to a chromosome, the plasmid is first divided into nonoverlapping 5-kb segments and the genome signature for each segment is determined. Let $\vec{x}_i$ be the vector of the 10 nonredundant dinucleotide relative abundances assessed from the *i*-th 5 kb segment, and $\vec{\mu}$ the mean vector of dinucleotide relative abundances over all 5 kb segments from the chromosome. Note that the relative abundances are symmetrized for the two DNA strands, and consequently six pairs of complementary dinucleotides have identical relative abundance. The Mahalanobis distance D_i of the plasmid segment $\vec{x}_i$ from the chromosome is calculated as:

$$D_i = \sqrt{(\vec{x}_i - \vec{\mu})^T S^{-1} (\vec{x}_i - \vec{\mu})}$$

where S^{-1} is the inverse variance-covariance matrix of the dinucleotide relative abundances from all 5 kb segments from the chromosome. The final distance between the plasmid and the chromosome is defined as the average distance D_i over all the 5 kb segments from the plasmid. Plasmids and their native host chromosomes tend to have similar dinucleotide relative abundances (Campbell et al. 1999, Suzuki et al. 2008) and subsequently, a small Mahalanobis distance. Mahalanobis distance is a more accurate measure of relatedness between a plasmid and a host than δ*-differences (Suzuki et al. 2008), but its complex and time consuming computation makes it less practical for large sequence datasets. Consequently, we used δ*-differences fpr

initial construction of the dendrogram and subsequently applied Mahalanobis distances to verify unexpected placements of plasmids in the tree.

RESULTS AND DISCUSSION

Comparison among plasmids

The dendrogram featuring the complete set of plasmids ≥ 50 kb length and constructed by the integrated method, is presented in Figure 1. The dendrogram is divided into seven major branches, labeled A-G, and 26 sub-branches, which generally follow the standard taxonomical classifications of the plasmid hosts at the level of phylum and subphylum. α-proteobacterial plasmids cluster in branches C_5 and F_1, γ-proteobacterial plasmids cluster in branches B_2, B_4, and C_2, while branches C_3 and F_2 are dominated by plasmids from β-proteobacteria. δ-proteobacterial plasmids are distributed on branches with other proteobacteria. Firmicutes are found in two branches, A_1 and G_2, separating *Clostridia* plasmids from other Firmicutes. Cyanobacterial plasmids forms branches A_2 and A_4, Spirochaetes cluster in branch G_1, Actinobacteria make up branches D_2, E_2, and E_3, and the Deinococcus-Thermus group plasmids are found in branches C_6 and G_4, separating the *Deinococcus* and *Thermus* genera. The lone Bacteroidetes plasmid clusters in branch A_1 with the Firmicutes. Archaeal plasmids dominate branch D. Branches A_3, B_1, B_3, C_1, C_4, and G_3 combine plasmids from various proteobacterial clades, possibly reflecting broad host range plasmids that are easily transferred between distantly related species. The "mixed" branch E_1 contains plasmids from taxonomically diverse hosts including Actinobacteria, β-proteobacteria, and α-proteobacteria.

The dendrogram in Figure 1 reveals several groups of plasmids from different hosts clustered at low distances (< ~4000), which indicates significant sequence homology detected by NUCmer. These plasmids likely derive from a single lineage present in different hosts, or broad

host range plasmids. For example, further investigation of the homologous regions in γ-proteobacterial plasmids pP99-018 from *Photobacterium damselae* subsp. piscicida and pIP1202 from *Yersinia pestis* biovar Orientalis str. IP275, which cluster in branch B_2 in Figure 1, shows that these plasmids share a 98% nucleotide sequence identity when aligned by blast and the two plasmids contain multiple identical segments of ≥10 kb identified by MUMmer (data not shown). In other similar examples, the *Erwinia amylovora* plasmid pEL60 and *Citrobacter freundii* plasmid pCTX-M3 share 96% nucleotide identity and the uncultured plasmid pTP6 shares 96% and 99% nucleotide sequence identity with the *Enterobacter aerogenes* plasmid R751 and *Delftia acidovorans* plasmid pU01, respectively. These similarities are suggestive of recent transfers of particular plasmid lineages between different genera and, in the latter case, even different phyla. *Klebsiella pneumoniae* plasmids pK2044 and pLVPK, which reside in the same host and differ in length by less than 5 kb, are very close in the dendrogram and very similar in sequence. This suggests that these plasmids belong to a single lineage and diverged only recently.

Combined Plasmid and Chromosome Clustering

The dendrogram constructed from plasmids and their host chromosomes divides into 13 major branches, labeled A-N, and 30 sub-branches (Figure 2). The distribution of different taxa in the combined dendrogram is similar to that in the plasmid only dendrogram (Figure 1). Most α-proteobacteria are clustered in the branch H while some form separate smaller branches cluster in branches E_4, J_3, J_7, and M, γ-proteobacteria cluster in branches D and E_2, branches E_3 and I are dominated by β-proteobacteria, while δ-proteobacteria is distributed in different branches across the tree. Firmicutes divide into two branches, A_1 and L, separating *Clostridium* plasmids from the rest of the Firmicutes as seen in the plasmid tree (Figure 1). This is consistent with whole

genome clustering, in which *Clostridium* clusters near *Borrelia* and separate from *Bacillus* and *Staphylococcus* (Henz, Huson et al. 2005). Cyanobacteria form branches A_2 and B, Spirochaetes cluster in branch K, and Actinobacteria make up branches J_5 and J_6. Various proteobacteria make up branches A_3, E_1, and E_5. Branch J_1 involves a mixed taxonomy including plasmids and chromosomes from Actinobacteria, β-proteobacteria, α-proteobacteria, Deinococcus-Thermus, aand chromosomes from γ-proteobacteria and Archaea.

Notably, plasmids are generally located on the same main branches as their host chromosomes, which is consistent with earlier observations that plasmids tend to have similar genome signatures to their hosts (Campbell, Mrazek et al. 1999; Suzuki, Sota et al. 2008). There are some exceptions where plasmids cluster on different major branches than their host chromosomes or where both the plasmid and chromosome are separated from taxonomically related hosts (Table 2). These instances could indicate recent transfer of the plasmids between distant hosts. We used the more accurate Mahalanobis distance to verify the anomalous placement of plasmids in Figure 2 and some of such specific cases are discussed next.

Shewanella baltica OS155 plasmid pSbal01 and the host chromosome have a NUCmer distance zero (Figure 2), indicating that the entire plasmid has a copy integrated in the chromosome. In fact, the exact copy of the plasmid is located in the chromosome from position 2542737 to 2659500, and the first gene in this sequence encodes a phage integrase family protein. The plasmid has 100% identity to this segment of the chromosome according to BLAST nucleotide comparison with only one single nucleotide deletion in the plasmid sequence. It is possible that the plasmid has been integrated into the chromosome or its inclusion in the chromosome sequence could be an artifact from shotgun sequencing and subsequent assembly.

The plasmid 1 of the β-proteobacterium *Nitrosomonas eutropha* C91 as well as its host chromosome cluster within the γ-proteobacteria dominated branch D_1. Mahalanobis distance calculations are consistent with the close genome signature similarity of the β-proteobacterium *N. eutropha* C91 plasmid 1 to its host *N. eutropha* C91, whereas the next 19 closest chromosomes are γ-proteobacteria (the first 10 shown in Table 3), suggesting correct placement of *N. eutropha* C91 plasmid 1 on the γ-proteobacterial branch of the dendrogram. The next closest β-proteobacterial host is *Polaromonas naphthalenivorans* (D=6.16). We speculate that the genome signature similarity between *N. eutropha* and γ-proteobacteria could be a result of lateral transfer of γ-proteobacterial replication and repair genes into the *N. eutropha* chromosome. The phylogenetic distribution of blast hits (downloaded from the IMG databse at http://img.jgi.doe.gov) indicates that 44 *N. eutropha* proteins have closest hits of 90% sequence identity or better to γ-proteobacteria and many of these proteins are involved in replication and repair, listed in Table 4, (all 44 proteins are listed in Table 5). The acquisition of these proteins could over time lead to a shift in the genome signature of the *N. eutropha* chromosome. The plasmid 1 may have undergone a similar change of the genome signature as the host chromosome or it may have been laterally transferred from a γ-proteobacterium.

For some outliers in the dendrograms in Figures 1 and 2 (i.e., plasmids located on a "wrong" branch of the tree and not clustered with phylogenetically close hosts), a possibility of recent transfer between distinct hosts is further supported by the observation that hosts of such outliers are often environmentally related to species on the same branch. For example, the plasmids pLPL and pLPP from *Legionella pneumophila*, a γ-proteobacterium, cluster with Cyanobacteria and Firmicutes (branch A_3 in Figures 1 and 2). *L. pneumophila* often relies on other microorganisms to provide nutrients for growth in the same environment, particularly

green algae and cyanobacteria (Tison, Pope et al. 1980; Cotuk, Dogruoz et al. 2005). This supports the notion that the *L. pneumophila* plasmids pLPL and pLPP could have been obtained from a cyanobacterium. Mahalanobis distance of plasmid pLPL with 115 chromosomes (Table 6 and Table 1) supports its placement among the Firmicutes and Cyanobacteria. The closest chromosome is the native host *Legionella pneumophila* str. Corby, the second smallest Mahalanobis distance is *Acinetobacter* sp. ADP1, and the following seven closest chromosomes are Firmicutes, Cyanobacteria, and one Bacteroidetes, all found in the same region in the dendrogram (Figure 2 and Table 6).

Archaea present an unexpected picture, where archaeal plasmids cluster together on branch G (except the *M. jannaschii* extrachromosomal element on branch L) but their host chromosomes are distributed among different branches (Figure 2). These archaeal plasmids are of vastly different sizes, ranging from 50 kb to 400 kb. Mahalanobis distance places the plasmids close to their hosts, within the top three chromosomes with the shortest genome signature, however, *Halobacterium* sp. NRC-1 has the shortest Mahalanobis distance to *Rhodococcus* sp. RHA1 and three *Mycobacteria* chromosomes, which are on the same branch as *Halobacterium* sp NRC-1 on the dendrogram in Figure 2. Further, Halobacterium sp NRC-1 clusters within the range of sequence and amino acid identity to the Actinobacteria listed above.

The extrachromosomal DNA of *Methanocaldococcus jannaschii* clusters away from other archaeal plasmids on the branch L with Firmicutes. Mahalanobis distance comparisons confirm that apart of the native host chromosome *Clostridium* species are closest to the *M. jannaschii* extrachromosomal DNA in terms of genome signature (Table 7). Further, excluding the host chromosome, the next closest closest archaeon is *Halobacterium* sp. NRC-1 with D=47.65 (Table 1).

Rickettsia felis URRWXCal2 chromosome and plasmid pRF are distant from other α-proteobacteria. Mahalanobis distance places the plasmid closest to its host chromosome, followed by the δ-proteobacterium *Lawsonia intracellularis* PHE/MN1-00 and two Firmicutes, *Clostridium acetobutylicum* ATCC 824 and *Lactobacillus salivarius* subsp. salivarius UCC118 (Table 7). The closest α-proteobacterium, *Jannaschia* sp. CCSI, has the Mahalanobis distance D=79.51. Notably, *L. intracellularis* PHE/MN1-00 plasmid 3 clusters with *R. felis* URRWXCal2 plasmid pRF on the plasmid tree (Figure 1). Both *R. felis* and *L. intracellularis* are intracellular pathogens (Ogata et al., 2005; Smith, D. Lawson, G., 2001) and it is intriguing to speculate that the similarity of their genome signatures could be related to their similar lifestyles.

Four plasmids from uncultured hosts were included in the dataset and all cluster within the branch E_1 in Figure 2, which includes β-proteobacteria, γ-proteobacteria, and one α-proteobacterium. The uncultured plasmid pB10 was reported most closely related to β-proteobacteria by Mahalanobis distance (Suzuki, Sota et al. 2008). Our tree places this plasmid near *Ralstonia eutropha* JMP134 plasmid 1 (a β-proteobacterium), at a distance less than 3000, which indicates significant nucleotide sequence similarity detected by NUCmer and suggests that the two plasmids derive from a single lineage. In a similar manner, plasmids pTP6 and pTB11, from uncultured hosts, are most closely related to *Enterobacter aerogenes* plasmid R751 and *Pseudomonas aeruginosa* plasmid pBS228, respectively, both γ- proteobacteria. This close clustering by nucleotide alignment suggests a shared evolutionary origin. Mahalanobis distances with the ten closest chromosomes are shown in Table 8.

The separation of Firmicutes into two distinct branches (Figures 1 and 2) is also present in other types of whole genome comparison trees (Henz, Huson et al. 2005). Mahalanobis distance results for the *Clostridium tetani* E88 plasmid pE88 support this separation, as the

nearest eight chromosomes are located on branches K-M on Figure 2 (D range 4.00 - 10.67; Figure 2 and Table 1), while the tenth chromosome, *Lactobacillus salivarius* subsp. salivarius UCC118, is more distant (D=17.80; branch A_1 in Figure 2).

Dendrogram construction using prokaryotic genomes, as opposed to single genes or 16S RNA sequences was utilized by Henz et al. and obtained similar clustering results to our integrated method (Henz, Huson et al. 2005). As mentioned, on both trees, two branches of Firmicutes are present, separating the *Clostridia* from the rest of the Firmicutes, and the branch containing *Clostridia* also includes *Fusobacterium nucleatum*. Further, the outlying α-proteobacteria, *Rickettsia*, falls next to this branch of Firmicutes. Another similarity of note is the clustering of *Halobacterium* sp. near the Actinobacteria. Our integrated method places *Halobacterium* sp. chromosome next to chromosomes from Actinobacteria on a multi-host branch (branch J_1, Figure 1). Raw values of NUCmer, PROmer, and genome signature differences show a closer similarity of *Halobacterium* to Actinobacteria species *Nocardia farcinica*, *Mycobacterium* sp., and *Rhodococcus* sp. than the Archaea *Natronomonas pharaonis* and *Methanocaldococcus jannaschii* (Table 6). The major multihost branch, J_1 in Figure 2, mainly includes β, γ and α-proteobacteria, Actinobacteria, and the Deinococcus-Thermus group. Mixed taxonomy branches are not necessarily unexpected as plasmids are transferred frequently between hosts and similarity of genome signature can reflect phylogenetic relationships but also metabolic similarities, environmental proximity or similarity of replication and repair machineries (Paz A, et al., 2006).

CONCLUSIONS

Current sequence comparison methods have limitations for the comparison of sequences that do not necessarily share homologous sequences, namely plasmids and viruses. Our integrated

method allows for accurate comparison of any DNA sequences and is useful when comparing different types and lengths of sequences such as plasmids and chromosomes.

Genome signature comparison offers a link for sequences that do not share homologous sequences and represents the backbone of the dendrogram. Many of the branch nodes are created with genome signature differences and the dendrogram shows that this method can identify related sequences as most taxonomical groups are placed together by genome signature data.

The integrated method is useful in studying the evolution of plasmids, especially in discovering possible instances of recent plasmid transfer between different hosts. Many instances of suggested plasmid transfer were found within our dataset (Table 2) and can be further investigated to determine if a transfer event occurred. Possibilities for investigation include Further, unknown sequences can be included in the comparison dataset and possible hosts and related sequences can be deduced.

Table 1. Table of plasmids and genomes used in this study.

Plasmids	Size (bp)	Abbreviation
α-proteobacteria		
Agrobacterium tumefaciens plasmid pTi-SAKURA	206479	A.tumef. pTi-SAKURA
Agrobacterium tumefaciens plasmid Ti	194140	A.timef. Ti
Agrobacterium tumefaciens strain C58 plasmid AT	542869	A.tumer. C58 AT
Agrobacterium tumefaciens strain C58 plasmid Ti	214331	A.tumef. C58 Ti
Gluconobacter oxydans 621H plasmid pGOX1	163186	G.oxyd. 621H pGOX1
Jannaschia sp. CCS1 plasmid1	86072	Janna. CCS1 1
Mesorhizobium sp. BNC1 plasmid 1	343931	Mesorh. BNC1 1
Mesorhizobium sp. BNC1 plasmid 2	131247	Mesorh. BNC1 2
Mesorhizobium loti plasmid pMLa	351911	M.loti pMLa
Mesorhizobium loti plasmid pMLb	208315	M.loti pMLb
Nitrobacter hamburgensis X14 plasmid 1	294829	N.ham. X14 1
Nitrobacter hamburgensis X14 plasmid 2	188318	N.hamb. X14 2
Nitrobacter hamburgensis X14 plasmid 3	121408	N.hamb. X14 3
Novosphingobium aromaticivorans plasmid pNL1	184457	N.aromat. pNL1
Oligotropha carboxidovorans plasmid pHCG3	133058	O.carb. pHCG3
Paracoccus denitrificans PD1222 plasmid 1	653815	P.denit. PD1222 1
Rhizobium sp. NGR234 plasmid pNGR234a	536165	Rhizo.NGR234 pNGR234a
Rhizobium etli CFN 42 plasmid p42a	194229	R.etli CFN 42 p42a
Rhizobium etli CFN 42 plasmid p42b	184338	R.etli CFN 42 p42b
Rhizobium etli CFN 42 plasmid p42c	250948	R.etli CFN 42 p42c
Rhizobium etli CFN 42 plasmid p42e	505334	R.etli CFN 42 p42e
Rhizobium etli CFN 42 plasmid p42f	642517	R.etli CFN 42 p42f
Rhizobium etli symbiotic plasmid p42d	371255	R.etli sym. p42d
Rhizobium leguminosarum bv. viciae 3841 plasmid pRL7	151564	R.legu.vic.3841 pRL7
Rhizobium leguminosarum bv. viciae 3841 plasmid pRL8	147463	R.legu.vic.3841 pRL8
Rhizobium leguminosarum bv. viciae 3841 plasmid pRL9	352782	R.legu.vic.3841 pRL9
Rhizobium leguminosarum bv. viciae 3841 plasmid pRL10	488135	R.legu.vic.3841 pRL10
Rhizobium leguminosarum bv. viciae 3841 plasmid pRL11	684202	R.legu.vic.3841 pRL11
Rhizobium rhizogenes plasmid pRi1724	217594	R.rhiz. pRi1724
Rhodobacter sphaeroides ATCC 17029 plasmid pRSPH01	122606	R.spha.A.17029pRSPH01
Rhodobacter sphaeroides 2.4.1 plasmid B	114178	R.sphae. 2.4.1 B

Table 1. Continued.

Rhodobacter sphaeroides 2.4.1 plasmid C	105284	R.sphae. 2.4.1 C
Rhodobacter sphaeroides 2.4.1 plasmid D	100828	R.sphae. 2.4.1 D
Rhodospirillum rubrum ATCC 11170 plasmid unnamed	53732	R.rubr. ATCC 11170 un
Rickettsia felis URRWXCal2 plasmid pRF	62829	R.felis URRWXCal2 pRF
Roseobacter denitrificans plasmid pTB1	106469	R.denit. pTB1
Roseobacter denitrificans plasmid pTB2	69269	R.denit. pTB2
Ruegeria sp. PR1b plasmid pSD20	76093	R. Pr1b pSD20
Ruegeria sp. PR1b plasmid pSD25	148650	R. PR1b pSD25
Silicibacter sp. TM1040 mega plasmid	821788	Silic. TM1040 mega
Silicibacter sp. TM1040 plasmid unnamed	130973	Silic. TM1040 unnamed
Silicibacter pomeroyi DSS-3 megaplasmid	491611	S.pom DSS-3 mega
Sinorhizobium meliloti FP2 plasmid pSB102	55578	S.melilo. FP2 pSB102
Sphingomonas sp. KA1 plasmid pCAR3	254797	Sphing. KA1 pCAR3
β-proteobacteria		
Achromobacter denitrificans plasmid pEST4011	76958	A.denit. pEST4011
Achromobacter xylosoxidans plasmid pA81	98192	A.xylos. pA81
Acidovorax sp. JS42 plasmid pAOVO01	72689	Acid. JS42 pAOVOO1
Acidovorax sp. JS42 plasmid pAOVO02	63609	Acid. JS42 pAOVOO2
Azoarcus sp. EbN1 plasmid 1	207355	Azoarcus EbN1 1
Azoarcus sp. EbN1 plasmid 2	223670	Azoarcus EbN1 2
Burkholderia cenocepacia HI2424 plasmid 1	164857	B.cenoc. HI2424 1
Burkholderia vietnamiensis G4 plasmid pBVIE01	397868	B.viet. G4 pBVIE01
Burkholderia vietnamiensis G4 plasmid pBVIE02	265616	B.viet. G4 pBVIE02
Burkholderia vietnamiensis G4 plasmid pBVIE03	226679	B.viet. G4 pBVIE03
Burkholderia vietnamiensis G4 plasmid pBVIE04	107231	B.viet. G4 pBVIE04
Burkholderia vietnamiensis G4 plasmid pBVIE05	88096	B.viet. G4 pBVIE05
Delftia acidovorans plasmid pUO1	67066	D.acidov. pU01
Methylibium petroleiphilum PM1 plasmid RPME01	599444	M.petro. PM1 RPME01
Nitrosomonas eutropha C91 plasmid1	65132	N.eutro. C91 1
Nitrosomonas eutropha C91 plasmid2	55635	N.eutro. C91 2
Polaromonas sp. JS666 plasmid 1	360405	Polar. JS666 1
Polaromonas sp. JS666 plasmid 2	338007	Polar. JS666 2
Polaromonas naphthalenivorans CJ2 plasmid pPNAP01	353291	P.napth. CJ2 pPNAP01

Table 1. Continued.

Polaromonas naphthalenivorans CJ2 plasmid pPNAP02	190172	P.napth. CJ2 pPNAP02
Polaromonas naphthalenivorans CJ2 plasmid pPNAP03	171866	P.napth. CJ2 pPNAP03
Polaromonas naphthalenivorans CJ2 plasmid pPNAP04	143747	P.napth. CJ2 pPNAP04
Polaromonas naphthalenivorans CJ2 plasmid pPNAP05	58808	P.napth. CJ2 pPNAP05
Ralstonia eutropha megaplasmid pHG1	452156	R.eutropha mega pHFG1
Ralstonia eutropha JMP134 plasmid 1	87688	R.eutro. JMP134 1
Ralstonia eutropha JMP134 megaplasmid	634917	R.eutr. JMP134 mega
Ralstonia metallidurans CH34 plasmid 1	233720	R.metal. CH34 1
Ralstonia metallidurans CH34 plasmid pMOL28	171461	R.met CH34 pMOL28
Ralstonia solanacearum strain GMI1000 megaplasmid pGMI1000MP	209450	R.sol. GMI1000 pGMI.
Rhodoferax ferrireducens T118 plasmid1	257447	R.ferr T118 1
δ-proteobacteria		
Desulfotalea psychrophila LSv54 plasmid large	121587	D.psyc. LSv54 lg
Desulfovibrio vulgaris subsp. vulgaris DP4 plasmid pDVUL01	198504	D.vul.vul.DP4 pDVUL01
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	202301	D.vul.vul.Hildenbor.
Lawsonia intracellularis PHE/MN1-00 plasmid 3	194553	L.intra PHE/MN1-00 3
Pelobacter propionicus DSM 2379 plasmid pPRO1	202397	P.prop DSM 2379 pPRO1
γ-proteobacteria		
Aeromonas punctata plasmid pFBAOT6	84749	A.punct. pFBAOT6
Aeromonas salmonicida subsp. salmonicida A449 plasmid 4	166749	A.sal.sal.A449 4
Aeromonas salmonicida subsp. salmonicida A449 plasmid 5	155098	A.sal.sal.A449 5
Citrobacter freundii plasmid pCTX-M3	89468	C.freundii pCTX-M3
Enterobacter aerogenes plasmid R751	53423	E.aerogenes R751
Erwinia amylovora plasmid pEL60	60145	E.amylovora pEL60
Escherichia coli plasmid NR1	94289	E.coli NR1
Escherichia coli plasmid p1658/97	125491	E.coli p1658/97
Escherichia coli plasmid pAPEC-O2-ColV	184501	E.coli pAPEC-01-ColV
Escherichia coli plasmid pAPEC-O2-R	101375	E.coli pAPEC-02-R
Escherichia coli plasmid pB171	68817	E.coli pB171
Escherichia coli plasmid pC15-1a	92353	E.coli pC15-1a
Escherichia coli plasmid pCoo	98396	E.coli pCoo
Escherichia coli plasmid pLEW517	65288	E.coli pLEW517
Escherichia coli plasmid pMUR050	56637	E.coli pMUR050

Table 1. Continued.

Escherichia coli plasmid pO113	165548	E.coli p0113
Escherichia coli plasmid pO86A1	120730	E.coli pO86A1
Escherichia coli plasmid R721	75582	E.coli R721
Escherichia coli K-12 CR63 plasmid F NC_002483	99159	E.coli K-12 CR63 F
Escherichia coli O157:H7 str. Sakai plasmid pO157	92721	E.coli O157:H7 pO157
Escherichia coli UTI89 plasmid pUTI89	114230	E.coli UTI89 pUTI89
Klebsiella pneumoniae plasmid pK2044	224152	K.pneumonia pK2044
Klebsiella pneumoniae plasmid pLVPK	219385	K.pneumoniae pLVPK
Legionella pneumophila str. Lens plasmid pLPL	59832	L.pneum. Lens pLPL
Legionella pneumophila str. Paris plasmid pLPP	131885	L.pneum. Paris pLPP
Listonella anguillarum plasmid pJM1	65009	L.anguillarum pJM1
Marinobacter aquaeolei VT8 plasmid pMAQU01	239623	M.aqua. VT8 pMAQU01
Marinobacter aquaeolei VT8 plasmid pMAQU02	213290	M.aqua. VT8 pMAQU02
Photobacterium damselae subsp. piscicida plasmid pP99-018	150157	P.dam pisci. pP99-018
Photobacterium profundum SS9 plasmid pPBPR1	80033	P.prof. SS9 pPBPR1
Proteus vulgaris plasmid Rts1	217182	P.vulg. Rts1
Pseudomonas sp. ADP plasmid pADP-1	108845	Pseudo. ADP pADP-1
Pseudomonas sp. ND6 plasmid pND6-1	101858	P. ND6 ND6-1
Pseudomonas aeruginosa plasmid pBS228	89147	P.aeruginosa pBS228
Pseudomonas aeruginosa plasmid Rms149	57121	P.aeruginosa Rms149
Pseudomonas putida plasmid NAH7	82232	P.putida NAH7
Pseudomonas putida plasmid pDTG1	83042	P.putida pDTG1
Pseudomonas putida plasmid pWW0	116580	P.putida pWW0
Pseudomonas putida plasmid pWW53	107929	P.putida pWW53
Pseudomonas resinovorans plasmid pCAR1	199035	P.resinov. pCAR1
Pseudomonas syringae pv. phaseolicola 1448A small plasmid	51711	P.syr.phase. 1448A sm
Pseudomonas syringae pv. phaseolicola 1448A large plasmid	131950	P.syr.phase. 1448A lg
Pseudomonas syringae pv. syringae plasmid pPSR1	72601	P.syr. syr. pPSR1
Pseudomonas syringae pv. tomato str. DC3000 plasmid pDC3000B	67473	P.syr. tom. pDC3000B
Pseudomonas syringae pv. tomato str. DC3000 plasmid pDC3000A	73661	P.syr. tom. pDC3000A
Salmonella enterica plasmid pOU1113	80156	S.enter pOU1113
Salmonella enterica subsp. enterica serovar Choleraesuis str.	138742	S.enter. ser. Choler.
Salmonella enterica subsp. enterica serovar Typhi plasmid pHCM1	218160	S.ent. ser. Ty. pHCM1

Table 1. Continued.

Salmonella enterica subsp. enterica serovar Choleraesuis strain RF-1 plasmid pKDSC50	49503	S.ent.Ch RF-1 pKDSC50
Salmonella enterica subsp. enterica serovar Newport str. SL254	176473	S.enter.ser.New.SL254
Salmonella enterica subsp. enterica serovar Typhi str. CT18 plasmid pHCM2	106516	S.ent.Typ. CT18 pHCM2
Salmonella typhi plasmid R27	180461	S.typhi R27
Salmonella typhimurium plasmid pU302L	84514	S.typh pU302L
Salmonella typhimurium plasmid R64	120826	S.typhim R64
Salmonella typhimurium IncN plasmid R46	50968	S.typhim IncN R46
Salmonella typhimurium LT2 plasmid pSLT	93939	S.typhim. LT2 pSLT
Serratia entomophila plasmid pADAP	106143	S.entomo. pADAP
Serratia marcescens plasmid R478	274762	S.marcescens R478
Shewanella sp. ANA-3 plasmid 1	278942	Shewan. ANA-3 1
Shewanella baltica OS155 plasmid pSbal01	116763	S.balt OS155 pSbal01
Shewanella baltica OS155 plasmid pSbal02	74000	S.balt OS155 pSbal02
Shewanella oneidensis MR-1 megaplasmid pMR-1	161613	S.onei.MR-1mega pMR-1
Shigella boydii Sb227 plasmid pSB4_227	126697	S.boy. Sb227 pSB4_227
Shigella dysenteriae Sd197 plasmid pSD1_197	182726	S.dys. Sd197 pSD1-197
Shigella flexneri virulence plasmid pWR501	221851	S.flex.vir. pWR501
Shigella sonnei P9 plasmid ColIb-P9	93399	S.sonnei P9 ColIb-P9
Shigella sonnei Ss046 plasmid pSS	214396	S.sonnei Ss046 pSS
Sodalis glossinidius str. 'morsitans' plasmid pSG1	83306	S.glos. mors. pSG1
Xanthomonas axonopodis pv. citri str. 306 plasmid pXAC64	64620	X.axon.cit.306 pXAC64
Xanthomonas campestris pv. vesicatoria str. 85-10 plasmid pXCV183	182572	X.ca.ve.85-10 pXCV183
Xylella fastidiosa plasmid pXF51	51158	X.fastidiosa pXF51
Yersinia enterocolitica plasmid pYVa127/90	66591	Y.enter pYVa127/90
Yersinia enterocolitica plasmid pYVe227	69673	Y.enter pYVe227
Yersinia enterocolitica plasmid pYVe8081	67720	Y.enter pYVe8081
Yersinia pestis biovar Medievalis str. 91001 plasmid pMT1	106642	Y.pe. Med91001 pMT1
Yersinia pestis biovar Orientalis str. IP275 plasmid pIP1202	182913	Y.pe.bi.IP275 pIP1202
Yersinia pestis KIM plasmid pCD1 (img:a)	70504	Y.pes. KIM pCD1
Yersinia pestis Pestoides F plasmid MT	137010	Y.pe.Pes. F MT
Yersinia pseudotuberculosis IP 32953 plasmid pYV	68526	Y.pseud. IP 32953 pYV
Yersinia ruckeri plasmid pYR1	158038	Y.ruck. pYR1

Table 1. Continued.

Actinobacteria		
Arthrobacter sp. FB24 plasmid 1	159538	Arthr. FB24 1
Arthrobacter sp. FB24 plasmid 2	115507	Arthr. FB24 2
Arthrobacter sp. FB24 plasmid 3	96488	Arthr. FB24 3
Arthrobacter aurescens TC1 plasmid TC1	328237	A.aures. TC1 TC1
Arthrobacter aurescens TC1 plasmid TC2	300725	A.aures. TC1 TC2
Corynebacterium striatum plasmid pTP10	51409	C.striatum pTP10
Gordonia westfalica plasmid pKB1	101016	G.westfalica pKB1
Micrococcus sp. 28 plasmid pSD10	50709	Micrococcus 28 pSD10
Mycobacterium sp. KMS plasmid pMKMS01	302089	Mycobac. KMS pMKMS01
Mycobacterium sp. KMS plasmid pMKMS02	216763	Mycobac. KMS pMKMS02
Mycobacterium gilvum PYR-GCK plasmid pMFLV01	321253	M.gil.PYR-GCK pMFLV01
Mycobacterium ulcerans plasmid pMUM001	174155	M.ulcer. pMUM001
Nocardioides sp. JS614 plasmid pNOCA01	307814	Nocard. JS614 pNOCA01
Nocardia farcinica IFM 10152 plasmid pNF1	184026	N.far IFM 10152 pNF1
Nocardia farcinica IFM 10152 plasmid pNF2	87093	N.far IFM 10152 pNF2
Rhodococcus sp. RHA1 plasmid pRHL1	112307	Rhodo. RHA1 pRHL1
Rhodococcus sp. RHA1 plasmid pRHL2	442536	Rhodo. RHA1 pRHL2
Rhodococcus sp. RHA1 plasmid pRHL3	332361	Rhodo. RHA1 pRHL3
Rhodococcus equi plasmid pREAT701	80610	R.equi pREAT701
Rhodococcus erythropolis plasmid pBD2	210205	R.eryth. pBD2
Rhodococcus erythropolis PR4 plasmid pREC1	104014	R.eryth. PR4 pREC1
Rhodococcus erythropolis PR4 plasmid pREL1	271577	R.eryth. PR4 pREL1
Streptomyces avermitilis MA-4680 plasmid SAP1	94287	S.aver. MA-4680 SAP1
Streptomyces coelicolor A3(2) plasmid SCP1	356023	S.coel A3(2) SCP1
Streptomyces lividans plasmid SLP2	50410	S.livid. SLP2
Streptomyces rochei plasmid pSLA2-L	210614	S.rochei pSLA2-L
Streptomyces violaceoruber plasmid pSV2	96742	S.violac. pSV2
Firmicutes		
Bacillus anthracis plasmid pX02	96231	B.anthrac. pX02
Bacillus anthracis str. 'Ames Ancestor' plasmid pXO1	181677	B.anthrac.Ames pX01
Bacillus cereus ATCC 10987 plasmid pBc10987	208369	B.cer A.10987pBc10987
Bacillus cereus E33L plasmid pE33L54	53501	B.cereusE33L pE33L54

Table 1. Continued.

Bacillus cereus E33L plasmid pE33L466	466370	B.cereusE33L pE33L466
Bacillus megaterium plasmid pBM400	53903	B.megat. pBM400
Bacillus thuringiensis serovar konkukian str. 97-27 plasmid	77112	B.thur.ser.konk.97-27
Bacillus thuringiensis str. Al Hakam plasmid pALH1	55939	B.thur. Al Hak. pALH1
Clostridium acetobutylicum ATCC 824 plasmid pSOL1	192000	C.acet.ATCC 824 pSOL1
Clostridium tetani plasmid pE88	74082	C. tetani pE88
Clostridium perfringens plasmid pCP13	54310	C.perfr. pCP13
Clostridium perfringens plasmid pCPF4969	70480	C.perf. pCPR4969
Clostridium perfringens plasmid pCPF5603	75268	C.perf. pCRP5603
Enterococcus faecalis plasmid pCF10	67673	E.faecalis pCF10
Enterococcus faecalis plasmid pRE25	50237	E.faecalis pRE25
Enterococcus faecalis V583 plasmid pTEF1	66320	E.faecalis V583 pTEF1
Enterococcus faecalis V583 plasmid pTEF2	57660	E.faecalis V583 pTEF2
Enterococcus faecium plasmid pHT beta	52890	E.faecium pHT beta
Geobacillus thermodenitrificans NG80-2 plasmid pLW1071	57693	G.ther.NG80-2 pLW1071
Lactobacillus salivarius subsp. salivarius UCC118 plasmid pMP118	242436	L.sali. UCC118 pMP118
Lactococcus lactis plasmid pIL105	536165	L.lactis pIL105
Lactococcus lactis plasmid pMRC01	60232	L.lactis pMRC01
Lactococcus lactis subsp. cremoris SK11 plasmid 3	74750	L.lactis crem. SK11 3
Listeria innocua plasmid pLI100	81905	L.inn. pLI100
Staphylococcus sp. 693-2 plasmid pLEW6932	51514	Staph. 693-2 pLEW6932
Staphylococcus aureus plasmid pLW043	57889	S.aureus pLW043
Cyanobacteria		
Anabaena variabilis ATCC 29413 plasmid A	366354	A.vari ATCC 29413 A
Anabaena variabilis ATCC 29413 plasmid C	300758	A.vari ATCC 29413 C
Nostoc sp. PCC 7120 plasmid pCC7120alpha	408101	Nos.PCC 7120 pCC7120a
Nostoc sp. PCC 7120 plasmid pCC7120beta	186614	Nos.PCC 7120 pCC7120b
Nostoc sp. PCC 7120 plasmid pCC7120delta	55414	Nos.PCC 7120 pCC7120d
Nostoc sp. PCC 7120 plasmid pCC7120gamma	101965	Nos.PCC 7120 pCC7120g
Synechocystis sp. PCC 6803 plasmid pSYSA	103307	Synec. PCC 6803 pSYSA
Synechocystis sp. PCC 6803 plasmid pSYSM	119895	Synec. PCC 6803 pSYSM
Synechocystis sp. PCC 6803 plasmid pSYSX	106004	Synec. PCC 6803 pSYSX

Table 1. Continued.

Spirochetes		
Borrelia afzelii PKo plasmid lp60	59958	B.afzel PKo lp60
Borrelia afzelii PKo plasmid lp60-2	59804	B.afzel PKo lp60-2
Borrelia burgdorferi plasmid lp54	53561	B.burg lp54
Borrelia burgdorferi B31 plasmid lp56	52971	B.burg. B31 lp56
Borrelia garinii PBi plasmid lp54	55560	B.gar. PBi lp54
Deinococcus-Thermus		
Deinococcus geothermalis DSM 11300 plasmid 1	574127	D.geo. DSM 11300 1
Deinococcus radiodurans R1 plasmid MP1	177466	D.rad. R1 MP1
Thermus thermophilus HB8 plasmid pTT27	256992	T.ther HB8 pTT27
Thermus thermophilus HB27 plasmid pTT27	232605	T.therm HB27 pTT27
Bacteroidetes		
Microscilla sp. PRE1 plasmid pSD15	101648	Microsc. PRE1 pSD15
Archaea		
Haloarcula marismortui ATCC 43049 plasmid pNG400	50060	H.marisA.43049 pNG400
Haloarcula marismortui ATCC 43049 plasmid pNG500	132678	H.mari.A.43049 pNG500
Haloarcula marismortui ATCC 43049 plasmid pNG600	155300	H.mari.A.43049 pNG600
Haloarcula marismortui ATCC 43049 plasmid pNG700	410554	H.marisA.43049 pNG700
Halobacterium sp. NRC-1 plasmid pNRC100	191346	Halo. NRC-1 pNRC100
Methanocaldococcus jannaschii extrachrom large extra-chromosomal	58407	M.jann extrachrom
Natronomonas pharaonis DSM 2160 plasmid PL131	130989	N.phar.DSM 2160 PL131
Other		
Uncultured bacterium plasmid pB4	79370	Uncultured pB4
Uncultured bacterium plasmid pB10	64508	Uncultured pB10
Uncultured bacterium plasmid pTB11	68869	Uncultured pTB11
Uncultured bacterium plasmid pTP6	54344	Uncultured pTP6
Genomes		
α-proteobacteria		
Agrobacterium tumefaciens str. C58 chromosome circular	2841490	*A.tumefaciens C58ci
Agrobacterium tumefaciens str. C58 chromosome linear	2075560	*A.tumefaciens C58li
Gluconobacter oxydans 621H	2702713	*G.oxydans 621H
Jannaschia sp. CCS1	4317977	*Jannaschia sp. CCS1

Table 1. Continued.

Mesorhizobium sp. BNC1	4412446	*Mesorhiz. sp. BNC1
Mesorhizobium loti MAFF303099	7036071	*M.loti MAFF303099
Nitrobacter hamburgensis X14	4406967	*N.hamburgensis X14
Nitrobacter winogradskyi Nb-255	3402093	*N.winograds. Nb-255
Novosphingobium aromaticivorans DSM 12444	3561584	*N.aromati.DSM 12444
Oligotropha carboxidovorans OM5	3745629	*O.carboxidovor.OM5
Paracoccus denitrificans PD1222 chromosome 1	2852282	*P.denitri.PD1222 c1
Rhizobium etli CFN 42	4381608	*R.etli CFN 42
Rhizobium leguminosarum bv. viciae 3841	5057142	*R.legu.bv.vic. 3841
Rhodobacter sphaeroides ATCC 17025	3217726	*R.sphaer.ATCC 17025
Rhodospirillum rubrum ATCC 11170	4352825	*R.rubrum ATCC 11170
Rickettsia felis URRWXCal2	1485148	*R.felis URRWXCal2
Roseobacter denitrificans OCh 114	4133097	*R.denitrif. OCh 114
Silicibacter pomeroyi DSS-3	4109442	*S.pomeroyi DSS-3
Sinorhizobium meliloti 1021	3654135	*S.meliloti 1021
Sphingomonas wittichii RW1	5382261	*S.wittichii RW1
β-proteobacteria		
Acidovorax sp. JS42	4448856	*Acidovorax sp JS42
Azoarcus sp. EbN1	4296230	*Azoarcus sp. EbN1
Burkholderia cenocepacia HI2424 chromosome 1	3483902	*B.cenoce.HI2424 c1
Burkholderia vietnamiensis G4 chromosome 1	3652814	*B.vietnam. G4 c1
Delftia acidovorans SPH-1	6767514	*D.acidovorans SPH-1
Methylibium petroleiphilum PM1	4044195	*M.petrolei. PM1
Nitrosomonas eutropha C91	2661057	*N.eutropha C91
Polaromonas sp. JS666	5200264	*Polaromon. sp.JS666
Polaromonas naphthalenivorans CJ2	4410291	*P.napthalen. CJ2
Ralstonia eutropha JMP134 chromosome 1	3806533	*R.eutrop. JMP134 c1
Ralstonia metallidurans CH34 chromosome 1	3928089	*R.metallidu.CH34 c1
Ralstonia solanacearum GMI1000	3716413	*R.solanacea.GMI1000
Rhodoferax ferrireducens T118	4712337	*R.ferrireduc. T118
δ-proteobacteria		
Desulfotalea psychrophila LSv54	3523383	*D.psychro. LSv54
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	3570858	*D.vul.subsp vul Hil

Table 1. Continued.

Lawsonia intracellularis PHE/MN1-00	1457619	*L.intra. PHE/MN1-00
Pelobacter propionicus DSM 2379	4008000	*P.propion. DSM 2379
γ-proteobacteria		
Acinetobacter sp. ADP1	3598621	*Acinetob. spADP1
Aeromonas salmonicida subsp. salmonicida A449	4702402	*A.salm. salm. A449
Citrobacter koseri ATCC BAA-895	4720462	*C.kose.ATCC BAA-895
Enterobacter sp. 638	4518712	*Enterobacter sp 638
Erwinia carotovora subsp. atroseptica SCRI1043	5064019	*E.caro.atr.SCRI1043
Escherichia coli K12	4639675	*E.coli K12
Escherichia coli O157:H7 EDL933	5528445	*E.coli O157:H7 EDL933
Escherichia coli UTI89	5065741	*E.coli UT189
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	5315120	*K.pne.pne.MGH 78578
Legionella pneumophila str. Paris	3503610	*L.pneumo. str.Paris
Marinobacter aquaeolei VT8	4326849	*M.aquaeolei VT8
Photobacterium profundum SS9 chromosome 1	4085304	*P.profundum SS9 c1
Pseudomonas aeruginosa PAO1	6264404	*P.aeruginosa PA01
Pseudomonas putida F1	5959964	*P.putida F1
Pseudomonas syringae pv. phaseolicola 1448A	5928787	*P.syr.pv.phas.1448A
Salmonella enterica subsp. enterica serovar Typhi str. CT18	4809037	*S.ent.ent.TyphiCT18
Salmonella typhimurium LT2	4857432	*S.typhumurium LT2
Serratia proteamaculans 568	5448853	*S.proteamaculans568
Shewanella baltica OS155	5127376	*S.baltica OS155
Shewanella oneidensis MR-1	4969803	*S.oneidensis MR-1
Shigella boydii Sb227	4519823	*S.boydii Sb227
Shigella dysenteriae Sd197	4369232	*S.dysenteriae Sd197
Shigella flexneri 2a str. 301	4607203	*S.flexn. 2a str.301
Shigella sonnei Ss046	4825265	*S.sonnei Ss046
Sodalis glossinidius str. 'morsitans'	4171146	*S.glossin.morsitans
Xanthomonas campestris pv. campestris str. ATCC 33913	5076188	*X.cam.str ATCC33913
Xylella fastidiosa 9a5c	2679306	*X.fastidiosa 9a5c
Yersinia enterocolitica subsp. enterocolitica 8081	4615899	*Y.enter.enter. 8081
Yersinia pestis biovar Microtus str. 91001	4595065	*Y.pe.bio.Micr.91001
Yersinia pestis KIM	4600755	*Y.pestis KIM

Table 1. Continued.

Yersinia pseudotuberculosis IP 32953	4744671	*Y.pseudotu.IP 32953
Actinobacteria		
Arthrobacter sp. FB24 chromosome 1	4698945	*Arthro. sp FB24 c1
Arthrobacter aurescens TC1	4597686	*A.aurescens TC1
Corynebacterium diphtheriae NCTC 13129	2488635	*C.dipthe.NCTC 13129
Mycobacterium sp. KMS	5737227	*Mycobac. sp. KMS
Mycobacterium sp. MCS	5705448	*Mycobac. sp. MCS
Mycobacterium gilvum PYR-GCK	5619607	*M.gilvum PYR-GCK
Mycobacterium ulcerans Agy99	5631606	*M.ulcerans Agy99
Nocardia farcinica IFM 10152	6021225	*N.farcin. IFM 10152
Nocardioides sp. JS614	4985871	*Nocardioid.sp.JS614
Rhodococcus sp. RHA1	7804765	*Rhodococcus sp.RHA1
Streptomyces avermitilis MA-4680	9025608	*S.avermiti. MA-4680
Streptomyces coelicolor A3(2)	8667507	*S.coelicolor A3(2)
Firmicutes		
Bacillus anthracis str. Ames	5227293	*B.anthracis Ames
Bacillus cereus ATCC 10987	5224283	*B.cereus ATCC 10987
Bacillus cereus ATCC 14579	5411809	*B.cereus ATCC 14579
Bacillus cereus E33L	5300915	*B.cereus E33L
Bacillus thuringiensis serovar konkukian str. 97-27 complet	5237682	*B.thu.ser.kon.97-27
Clostridium acetobutylicum ATCC 824	3940880	*C.acetobut.ATCC 824
Clostridium perfringens str. 13	3031430	*C.perfringen. str13
Clostridium tetani E88	1799251	*C.tetani E88
Enterococcus faecalis V583	3218031	*E.faecalis V583
Geobacillus thermodenitrificans NG80-2	3550319	*G.thermodeni.NG80-2
Lactobacillus salivarius subsp. salivarius UCC118	1827111	*L.saliv.sali.UCC118
Lactococcus lactis subsp. lactis Il1403	2365589	*L.lactis lac.Il1403
Listeria innocua Clip11262	3011208	*L.innocua Clip11262
Staphylococcus aureus subsp. aureus COL	2809422	*S.aureus aureus COL
Cyanobacteria		
Anabaena variabilis ATCC 29413	6365727	*A.variab.ATCC 29413
Nostoc sp. PCC 7120	6413771	*Nostoc sp. PCC 7120
Synechocystis sp. PCC 6803	3573470	*Synecho.sp.PCC 6803

Table 1. Continued.

Spirochetes		
Borrelia afzelii PKo	905394	*B.afzelii PKo
Borrelia burgdorferi B31	910724	*B.burgdoferi B31
Borrelia garinii PBi chromosome linear	904246	*B.garinii PBi
Deinococcus-Thermus		
Deinococcus geothermalis DSM 11300	2467205	*D.geother.DSM 11300
Deinococcus radiodurans R1 chromosome 1	2648638	*D.radiodurans R1 c1
Thermus thermophilus HB27	1894877	*T.thermophilus HB27
Bacteroidetes		
Bacteroides vulgatus ATCC 8482	5163189	*B.vulgat. ATCC 8482
Archaea		
Haloarcula marismortui ATCC 43049 chromosome I	3131724	*H.marismo.ATCC43049
Halobacterium sp. NRC-1	2014239	*Halobact. sp. NRC-1
Natronomonas pharaonis DSM 2160	2595221	*N.pharaonis DSM2160
Other		
Acidobacteria bacterium Ellin345	5650368	*A.bacteri. Ellin345
Aquifex aeolicus VF5	1551335	*A.aeolicus VF5
Chlamydia trachomatis A/HAR-13	1044459	*C.trachom. A/HAR-13
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	2174500	*F.nuc.nuc.ATCC25586
Rhodopirellula baltica SH 1	7145576	*R.baltica SH 1
Roseiflexus sp. RS-1	5801598	*Roseiflexus sp.RS-1
Thermosipho melanesiensis BI429	1915238	*T.melanesien. BI429

Table 2. Outliers from the Dendrogram in Figure 1. Outliers are plasmids clustering outside the branch of their taxanomical host. Mahalanobis distance is given for the three closest chromosomes.

Plasmid	Dendrogram Clustering	Mahalanobis Clustering	Mahalanobis Distance (D)
L. pneumonia str Paris pLPP	Firmicutes	*L. pneumonia* str. Corby	1.45
	Cyanobacteria	*E. faecalis* V583	2.79
		B. cereus ATCC 10987	4.3
N.eutropha C91 plasmid 1	γ-proteobacteria	*N. eutropha* C91	1.99
		E. coli O157:H7 EDL933	2.17
		Enterobacter sp. 638	3.02
C. striatum pTP10	γ-proteobacteria	*C. diptheriae NCTC 13129*	3.65
		Acidovorax sp. JS42	4.48
		Polaromonas sp. JS666	4.71
R. felis URRWXCal2 pRF	γ-proteobacteria	*R. felis URRWXCal2*	6.13
	Firmicutes	*C. perfringens* str. 13	6.39
	Cyanobacteria	*C. tetani* E88	6.86
M. jannaschii extrachrom large	Firmicutes	*M. jannaschii* DSM 2160	2.74
	Spirochaetes	*C. tetani* E88	3.55
		F. nucleatum subsp. nucleatum ATCC 25586	4.19
C. tetani pE88	Firmicutes	*C. perfringens* str. 13	4
	Spirochaetes	*C. tetani* E88	4.34
		F. nucleatum subsp. nucleatum ATCC 25586	6.57

Table 3. Mahalanobis distance (D) for *Nitrosomonas eutropha* C91 plasmid 1 to all chromosomes in our dataset.

Chromosome	**Host classification**	**D**
N. europaea ATCC 19718	β-Proteobacteria	2.04
E. coli O157:H7 EDL933	γ-Proteobacteria	2.17
X. fastidiosa 9a5c	γ-Proteobacteria	2.46
S. flexneri 2a str. 301	γ-Proteobacteria	2.75
E. coli UTI89	γ-Proteobacteria	2.78
E. carotovora subsp. atroseptica SCRI1043	γ-Proteobacteria	2.90
Enterobacter sp. 638	γ-Proteobacteria	3.02
S. boydii Sb227	γ-Proteobacteria	3.02
S. sonnei Ss046	γ-Proteobacteria	3.06
S. dysenteriae Sd197	γ-Proteobacteria	3.24

Table 4. Genes identified in the chromosome of *Nitrosomonas eutropha* C91 which have at least 90% sequence identity to γ-proteobacteri and are involved in replication, recombination, and repair.

Gene Name	Percent identity
Transposase Tn3 family protein	100
Resolvase, N-terminal domain	100
Resolvase, N-terminal domain	100
Putative DNA helicase	100
Phage integrase family protein	100
Resolvase, N-terminal domain	100
Transposase Tn3 family protein	100
Phage integrase family protein	100
Putative DNA helicase	100
Resolvase, N-terminal domain	100

Table 5. Proteins in *N. eutropha* C71 with best hit nucleotide intentidy of at least 90% to a γ-proteobacterial sequence.

Gene Name	Percent Identity	Cog Function
Transposase Tn3 family protein	100	Replication, recombination, and repair
Resolvase, N-terminal domain	100	Replication, recombination, and repair
Conserved hypothetical protein	100	
MerE family protein	100	
Transcriptional regulator, MerR family	100	Transcription
MerE family protein	93	
Conserved hypothetical protein	90	Intracellular trafficking, secretion, and vesicular transport
Conserved hypothetical protein	100	
Resolvase, N-terminal domain	100	Replication, recombination, and repair
Transcriptional regulator, ArsR family protein	100	Transcription
NADPH-dependent FMN reductase	100	General function prediction only
Protein tyrosine phosphatase	100	Signal transduction mechanisms
Bile acid:sodium symporter	100	Inorganic ion transport and metabolism
Conserved hypothetical protein	98	
TrbL/VirB6 plasmid conjugal transfer protein	98	Intracellular trafficking, secretion, and vesicular transport
Putative mating pair formation protein	100	Intracellular trafficking, secretion, and vesicular transport
Putative stabilization protein	100	General function prediction only
Replication C family protein	100	
Putative DNA helicase	100	Replication, recombination, and repair
Phage transcriptional regulator, AlpA	100	Transcription
Phage integrase family protein	100	Replication, recombination, and repair
Transcriptional regulator, MerR family protein	100	Transcription
MerE family protein	100	
Conserved hypothetical protein	100	
Resolvase, N-terminal domain	100	Replication, recombination, and repair

Table 5. Continued.

Transposase Tn3 family protein	100	Replication, recombination, and repair
ATP synthase F0, C subunit	91	Energy production and conversion
Phage integrase family protein	100	Replication, recombination, and repair
Phage transcriptional regulator, AlpA	100	Transcription
Putative DNA helicase	100	Replication, recombination, and repair
Replication C family protein	100	
Putative stabilization protein	100	General function prediction only
Putative mating pair formation protein	100	Intracellular trafficking, secretion, and vesicular transport
TrbL/VirB6 plasmid conjugal transfer protein	98	Intracellular trafficking, secretion, and vesicular transport
Conserved hypothetical protein	98	
Bile acid:sodium symporter	100	Inorganic ion transport and metabolism
Protein tyrosine phosphatase	100	Signal transduction mechanisms
NADPH-dependent FMN reductase	100	General function prediction only
Transcriptional regulator, ArsR family protein	100	Transcription
Resolvase, N-terminal domain	100	Replication, recombination, and repair
Conserved hypothetical protein	100	

Table 6. Closest nine chromosomes to *Legionella pneumophila* str. Corby plasmid pLPL based on Mahalanobis distance to all chromosomes in our dataset (shown in Table 1).

Chromosome	**Host classification**	**D**
Legionella pneumophila str. Corby	γ-Proteobacteria	2.58
Acinetobacter sp. ADP1	γ-Proteobacteria	3.18
Enterococcus faecalis V583	Firmicutes	3.78
Bacteroides vulgatus ATCC 8482	Bacteroidetes	4.12
Lactococcus lactis subsp. lactis Il1403	Firmicutes	4.28
Bacillus cereus ATCC 10987	Firmicutes	5.10
Listeria innocua Clip11262	Firmicutes	5.35
Anabaena variabilis ATCC 29413	Cyanobacteria	5.61
Bacillus cereus ATCC 14579	Firmicutes	5.66

Table 7. Mahalanobis distances (D) of *M. jannaschii* extrachromosomal element and *R. felis* plasmids pRF to the closest 9 chromosomes in our dataset.

M. jannaschii extrachromosomal element

Chromosome	**Host classification**	D
C. tetani E88	Firmicutes	3.55
F. nucleatum subsp. nucleatum ATCC 25586	Other	4.19
C. perfringens str. 13	Firmicutes	4.38
C. acetobutylicum ATCC 824	Firmicutes	4.47
T. melanesiensis BI429	Other	5.10
L. salivarius subsp. salivarius UCC118	Firmicutes	5.52
B. garinii PBi chromosome linear	Spirochaetes	6.15
B. burgdorferi B31	Spirochaetes	6.22
B. afzelii Pko	Spirochaetes	6.34

R. felis URRWXCal2 plasmid pRF

Chromosome	**Host classification**	D
L. intracellularis PHE/MN1-00	δ-Proteobacteria	4.61
C. acetobutylicum ATCC 824	Firmicutes	5.15
L. salivarius subsp. salivarius UCC118	Firmicutes	5.53
R. felis URRWXCal2	α-Proteobacteria	6.13
C. perfringens str. 13	Firmicutes	6.39
C. tetani E88, complete genome.	Firmicutes	6.86
B. thuringiensis serovar konkukian str. 97-27	Firmicutes	7.25
B. cereus ATCC 14579	Firmicutes	7.27
E. faecalis V583	Firmicutes	7.59

Table 8. Mahalanobis distance of unclutured plasmids pB10, pTP6, pTB11, and pB4 to the dataset of genomes.

pTB10

Chromosome	**Host classification**	D
R. solanacearum GMI1000	β-Proteobacteria	2.53
R. metallidurans CH34 chromosome 1	β-Proteobacteria	2.94
X. campestris pv. campestris str. ATCC 33913	γ-Proteobacteria	2.96
R. eutropha JMP134 chromosome 1	β-Proteobacteria	2.98
Acidovorax sp. JS42	β-Proteobacteria	3.02
Polaromonas sp. JS666	β-Proteobacteria	3.08
P. aeruginosa PAO1.	γ-Proteobacteria	3.42
N. aromaticivorans DSM 12444	α-Proteobacteria	3.47
Roseiflexus sp. RS-1	Other	3.72
M. loti MAFF303099	α-Proteobacteria	3.76

pTB11

Chromosome	**Host classification**	D
R. solanacearum GMI1000	β-Proteobacteria	2.82
R. metallidurans CH34 chromosome 1	β-Proteobacteria	2.92
R. eutropha JMP134 chromosome 1	β-Proteobacteria	3.00
Polaromonas sp. JS666	β-Proteobacteria	3.24
Roseiflexus sp. RS-1	Other	3.34
M loti MAFF303099	α-Proteobacteria	3.35
N. aromaticivorans DSM 12444	α-Proteobacteria	3.36
X. campestris pv. campestris str. ATCC 33913	γ-Proteobacteria	3.41
Acidovorax sp. JS42	β-Proteobacteria	3.79
P. naphthalenivorans CJ2	β-Proteobacteria	3.95

Table 6 Continued.

pTP6

Chromosome	**Host classification**	D
R. solanacearum GMI1000	β-Proteobacteria	2.57
R. eutropha JMP134 chromosome 1	β-Proteobacteria	2.96
X. campestris pv. campestris str. ATCC 33913	γ-Proteobacteria	2.99
Acidovorax sp. JS42	β-Proteobacteria	3.00
R. metallidurans CH34 chromosome 1	β-Proteobacteria	3.10
P. aeruginosa PAO1	γ-Proteobacteria	3.31
Polaromonas sp. JS666	β-Proteobacteria	3.35
P. naphthalenivorans CJ2	β-Proteobacteria	3.85
Roseiflexus sp. RS-1	Other	3.87
N. aromaticivorans DSM 12444	α-Proteobacteria	3.87

pB4

Chromosome	**Host classification**	D
X. campestris pv. campestris str. ATCC 33913	γ-Proteobacteria	2.61
Acidovorax sp. JS42	β-Proteobacteria	2.81
Polaromonas sp. JS666	β-Proteobacteria	2.86
R. metallidurans CH34 chromosome 1	β-Proteobacteria	3
R. eutropha JMP134 chromosome 1	β-Proteobacteria	3.08
R. solanacearum GMI1000	β-Proteobacteria	3.20
Roseiflexus sp. RS-1	Other	3.35
P. naphthalenivorans CJ2	β-Proteobacteria	3.54
P. aeruginosa PAO1	γ-Proteobacteria	3.8
M. loti MAFF303099	α-Proteobacteria	4.02

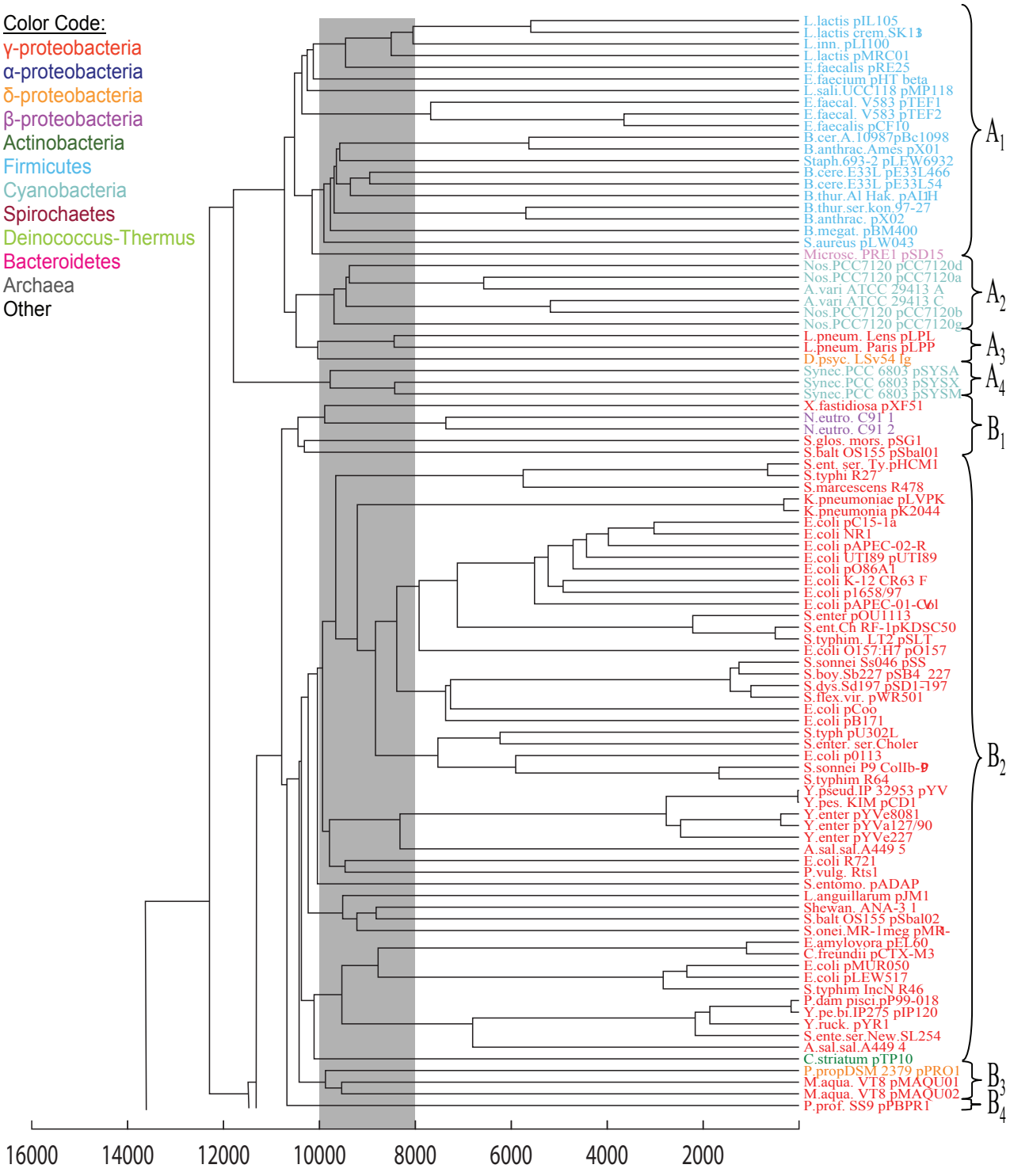

Figure 1. The dendrogram was constructed by the integrated method using plasmid sequences great than 50 kb in length. The gray bar represents the transition from amino acid similarity to genome signature differences. This gray area is where most of the connections arise, demonstrating the ability of genome signature differences to distinguish taxonomical groups. Clustering below 5000 is dominated by NUCmer, 5000-9000 by PROmer, and >9000 by genome signature.

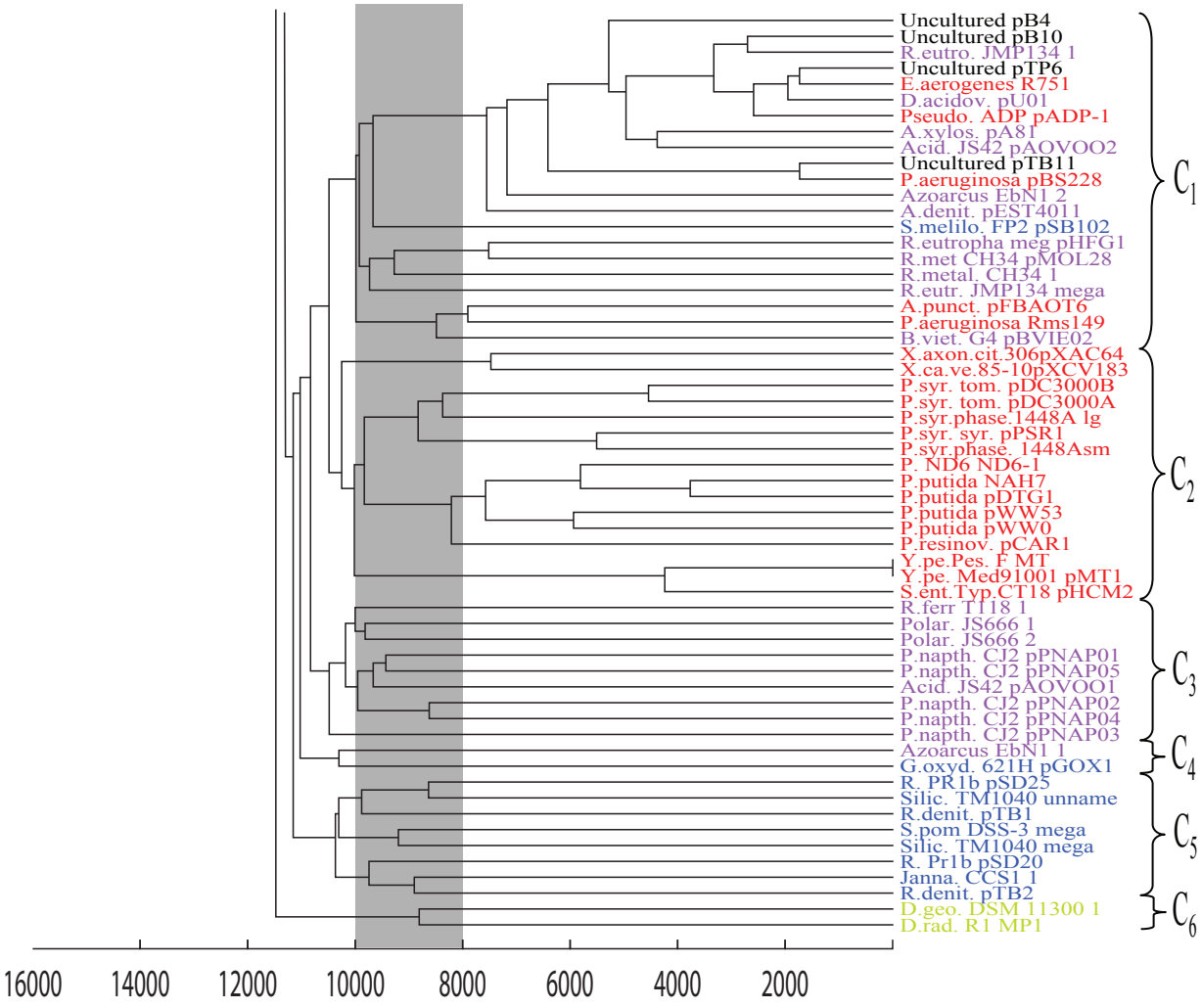

Figure 1. Continued.

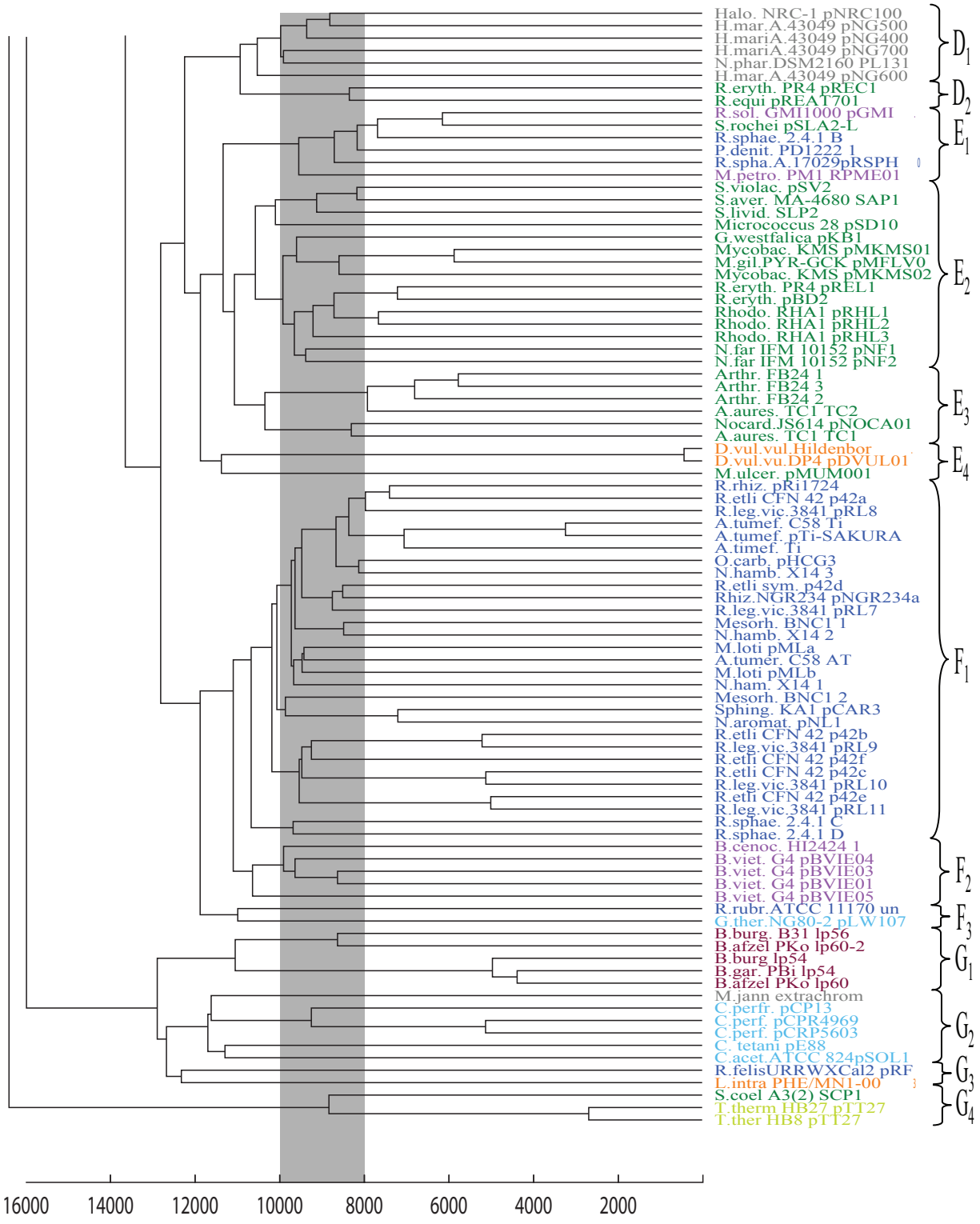

Figure 1. Continued.

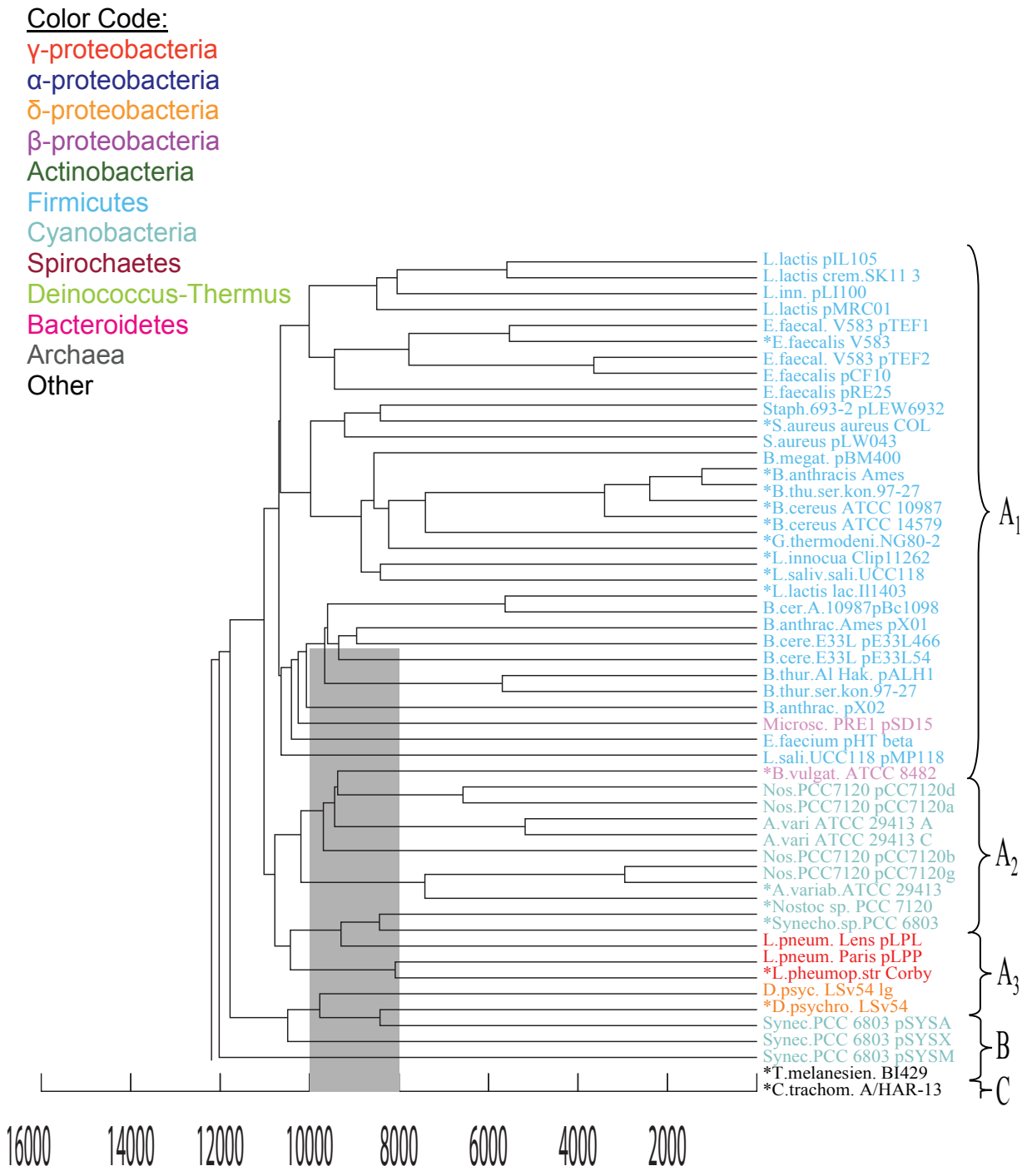

Figure 2. The dendrogram was constructing using the integrated method. Plasmid host chromosomes were included when a full sequence was available. The gray bar represents the transition between amino acid similarity and genome signature differences. * denotes a chromosome sequence, all other sequences are plasmids. Clustering below 5000 is dominated by NUCmer, 5000-9000 by PROmer, and >9000 by genome signature.

Figure 2. Continued

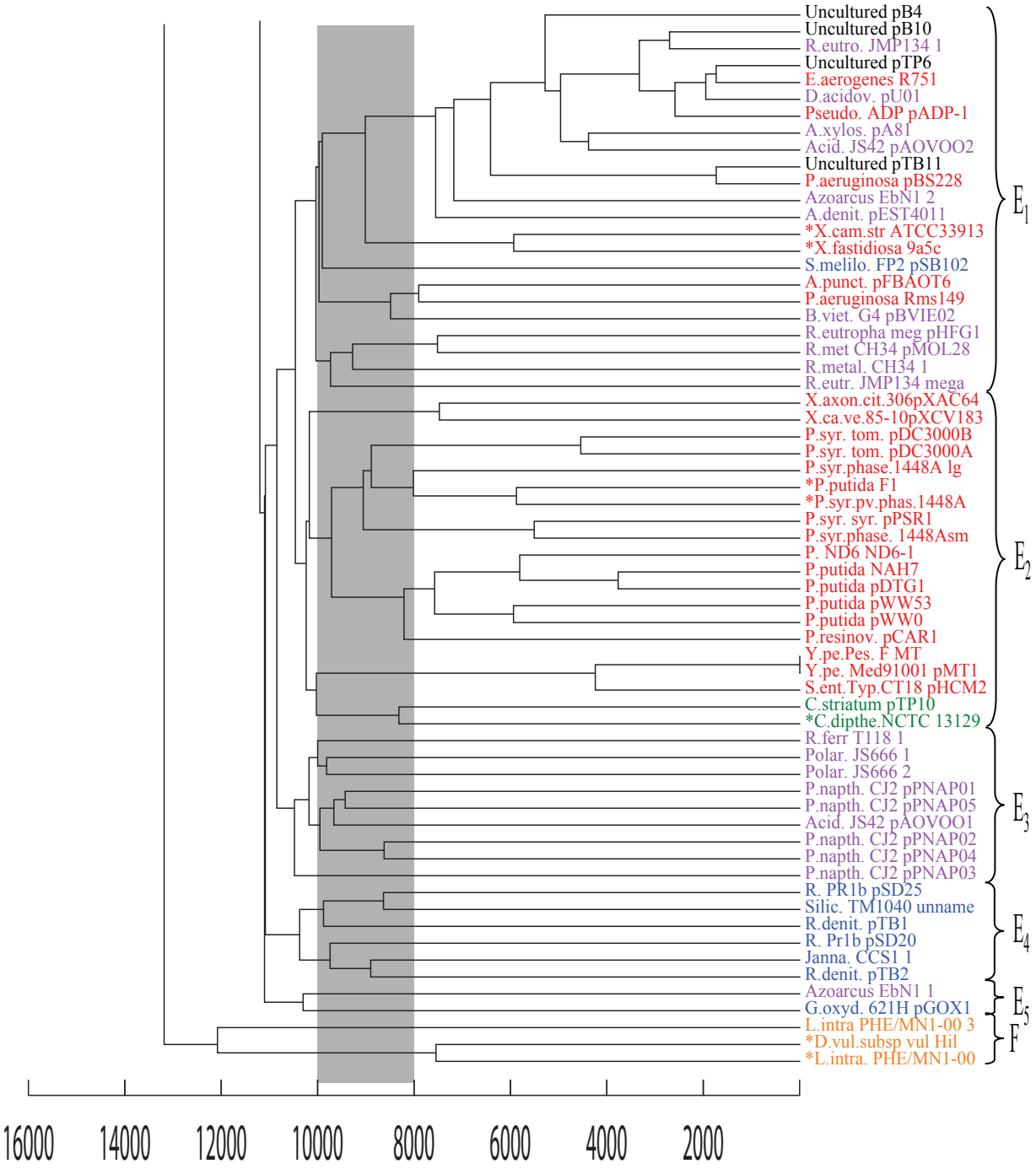

Figure 2. Continued

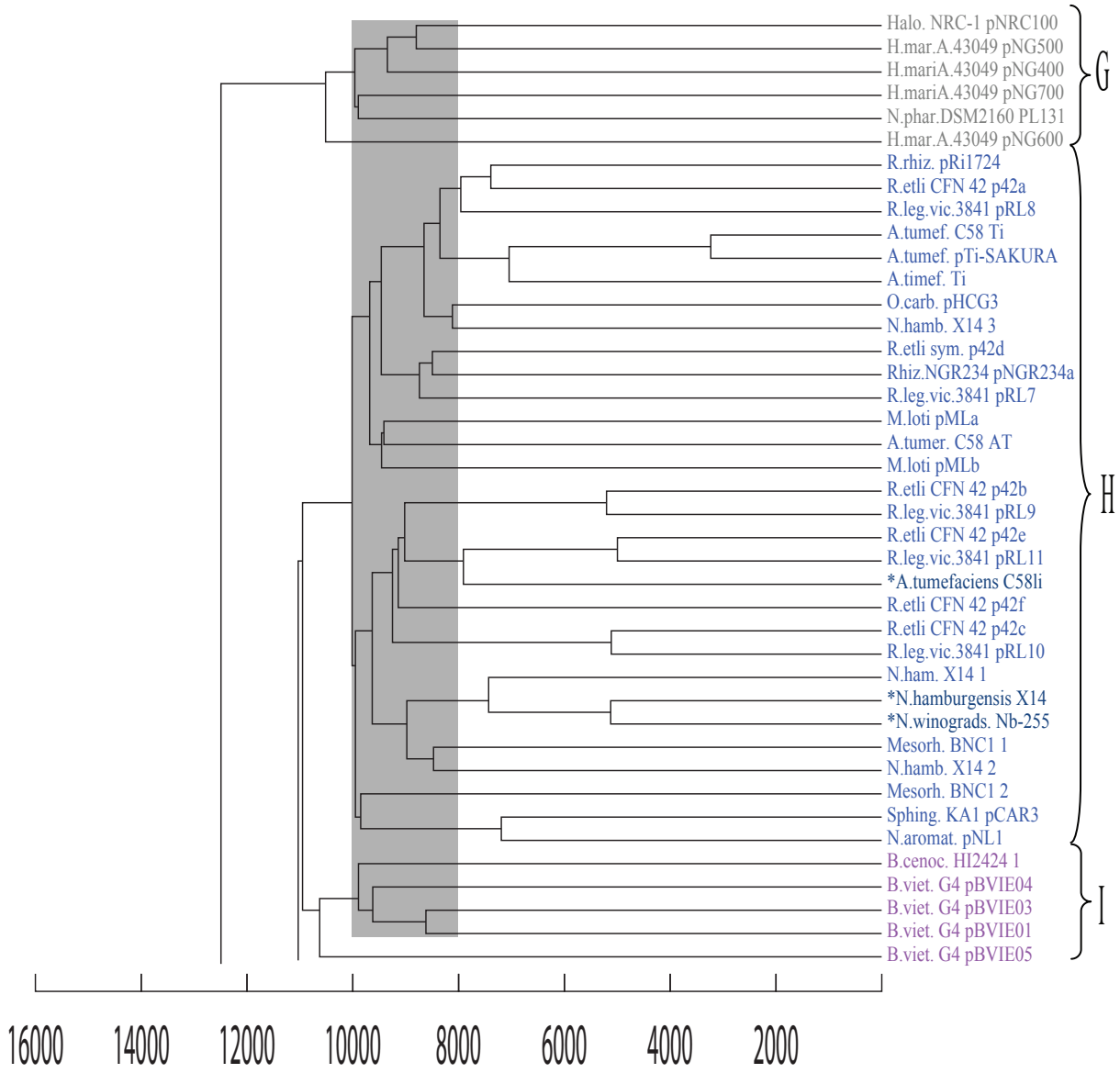

Figure 2. Continued

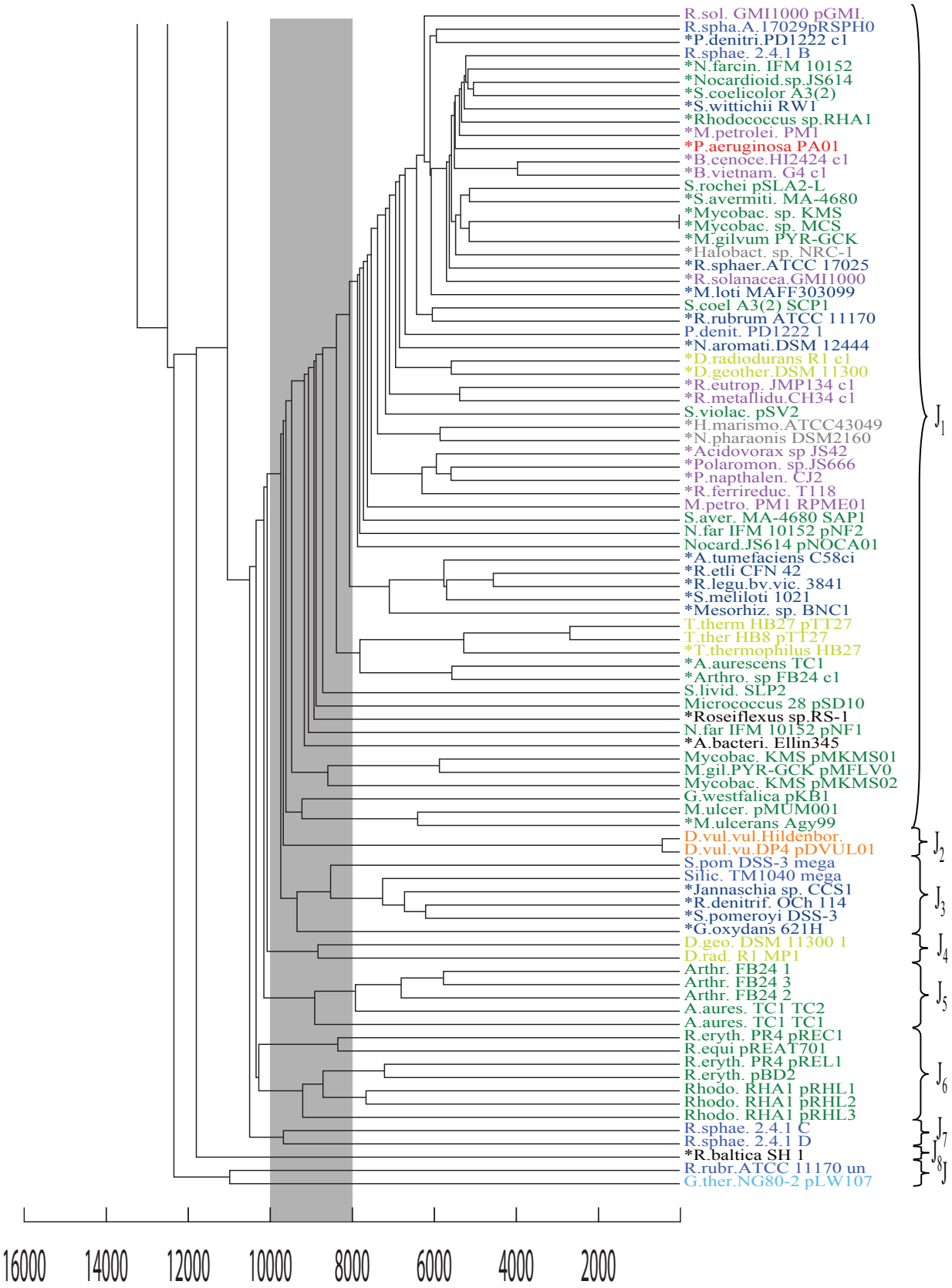

Figure 2. Continued

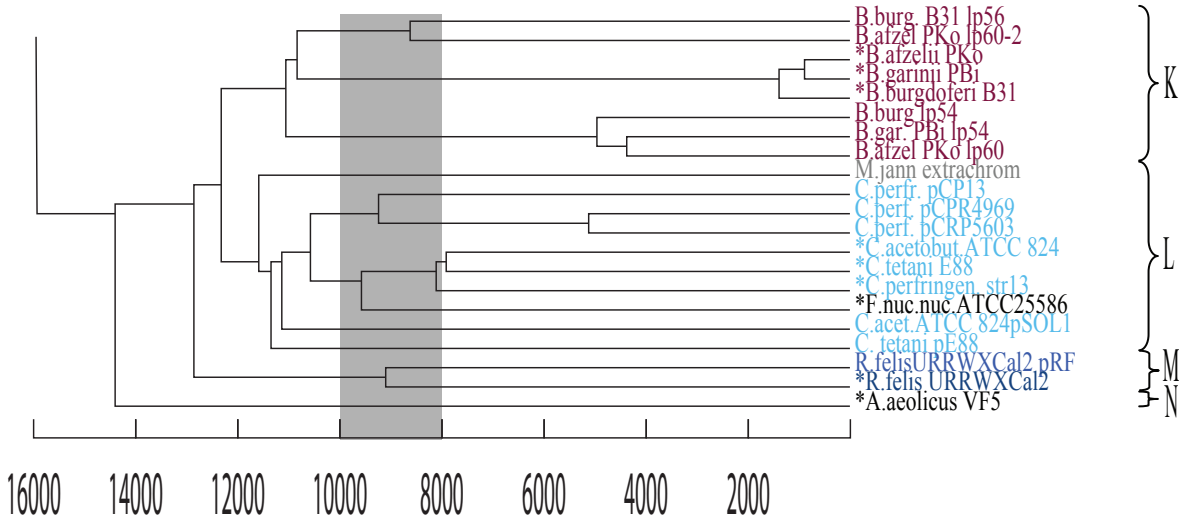

Figure 2. Continued

CHAPTER 3

CONCLUSIONS AND FUTURE DIRECTIONS

This study focused on developing a method suitable for comparison of DNA sequences that do not necessarily share homologous genes. Plasmids were used to demonstrate the effectiveness of the method. However, it is also appropriate for viruses, phages, and other mobile elements where standard homology-based methods have limited use. To accomplish this task, resolution must be seen between sequences that are closely related while at the same time placing distantly related sequences together as a deep-rooted branch. Combining NUCmer, PROmer, and genome signature realizes this goal.

NUCmer and PROmer connect sequences that are more closely related using nucleotide and amino acid identity, respectively. Genome signature comparison offers a link for sequences that do not share homologous sequences and represents the backbone of the dendrogram. Many of the branch nodes are created with genome signature differences and the dendrogram shows that this method can identify related sequences as most taxonomical groups are placed together by genome signature data.

Further optimization of the integrated technique and its parameters for specific types of applications could provide improvement for comparisons that have already proved to be useful. Optimization of genome signature using various oligonucleotides to find the best components may provide a better compositional comparison. Dinucleotide relative abundance has already been shown to give the best comparison over tri- and tetra-nucleotide comparisons. Conceivably, a combination of components will prove to have the best resolution for DNA sequences.

Principal component analysis has great potential in identifying the most informative variables for genome comparisons. While the method is somewhat labor intensive, much information can be learned from the results. The result reflects each of the components and therefore gives a comprehensive visual of the DNA sequence comparison. With this application, one can elucidate the oligonucleotides and alphabets that create the best organization of organisms based on standard taxonomy by separating distantly related DNA sequences but conserving the similarity of closely related sequences.

The integrated method is useful in studying the evolution of plasmids, especially in discovering possible instances of recent plasmid transfer between different hosts. Many instances of suggested plasmid transfer were found without our dataset (Table 2) and can be further investigated to determine if a transfer event occurred. Further, unknown sequences can be included in the comparison dataset and possible hosts and related sequences can be deduced.

REFERENCES

1. **Belda, E., Moya, A., Silva, F.J.** (2005) Genome rearrangement distances and gene order phylogeny in gamma-Proteobacteria. Mol Biol Evol **22**:1456-1467.

2. **Brussow, H.** (2009). "The not so universal tree of life *or* the place of viruses in the living world." Phil. Trans. R. Soc. B **364**(1627): 2263-2274.

3. **Campbell, A., J. Mrazek, and S. Karlin.** 1999. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. **96:**9184-9189.

4. **Canchaya, C., G. Fournous, S. Chibani-Chennoufi, M. L. Dillmann, and H. Brussow.** 2003. Phage as agents of lateral gene transfer. Curr. Opin. Microbiol. **6:**417-24.

5. **Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., Bork, P.** (2006) Toward automatic reconstruction of a highly resolved tree of life. Science **311**:1283-1287.

6. **Cohen, S. N., A. C. Chang, and L. Hsu.** 1972. Nonchromosomal antibiotic resistance in bacteria: genetic transformation of *Escherichia coli* by R-factor DNA. Proc. Natl. Acad. Sci. U. S. A. **69:**2110-4.

7. **Cotuk, A., N. Dogruoz, et al.** (2005). "The effects of *Pseudomonas* and *Aeromonas* strains on *Legionella pneumophila* growth." Annals of Microbiol. **55**(3): 219-224.

8. **Delcher, A.L., Phillippy, A., Carlton, J., Salzberg, S.L.** (2002). "Fast algorithms for large-scale genome alignment and comparison." Nucleic Acids Research **30**(11): 2478-2483.

9. **Doolittle, W.F.** (1999) Phylogenetic classification and the universal tree. Science **284**:2124-2129.

10. **Echols, H., Goodman, M.** (1991) "Fidelity mechanisms in DNA replication." Annu. Rev. Biochem. **60**: 477-511.

11. **Fitz-Gibbon, S.T., House, C.H.** (1999) Whole genome-based phylogenetic analysis of free-living microorganisms. Nucleic Acids Res. **27**:4218-4222.

12. **Foster, T. J.** 1983. Plasmid-determined resistance to antimicrobial drugs and toxic metal ions in bacteria. Microbiol. Rev. **47:**361-409.

13 **Frost, L. S., R. Leplae, A. O. Summers, and A. Toussaint.** 2005. Mobile genetic elements: the agents of open source evolution. Nat. Rev. Microbiol. **3:**722-32.

14. **Henz, S. R., D. H. Huson, et al**. (2005). "Whole-genome prokaryotic phylogeny." Bioinformatics **21**(10): 2329-2335.

15. **Kapatral, V., I. Anderson, et al**. (2002). "Genome sequence and analysis of the oral bacterium *Fusobacterium nucleatum* strain ATCC 25586." J. Bacteriol. **184**(7): 2005-2018.

16. **Karlin, S., Burge, C.** (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet **11**:283-290.

17. **Konstantinidis, K.T., Tiedje, J.M.** (2005) Towards a genome-based taxonomy for prokaryotes. J Bacteriol **187**:6258-6264.

18. **Korbel, J.O., Snel, B., Huynen, M.A., Bork, P.** (2002) SHOT: a web server for the construction of genome phylogenies. Trends Genet **18**:158-162.

19. **Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg** (2004). Versatile and open software for comparing large genomes. Genome Biol. **5:**R12.

20. **Mrázek, J.** (2009) Phylogenetic signals in DNA composition: limitations and prospects. Mol Biol Evol **26**:1163-1169.

21. **Novick, R. P.** 1969. Extrachromosomal inheritance in bacteria. Bacteriol. Rev. **33:**210-63.

22. **Novick, R. P., and C. Roth.** 1968. Plasmid-linked resistance to inorganic salts in *Staphylococcus aureus*. J. Bacteriol. **95:**1335-42.

23. **Ogata, H., Renesto, P., Audic, S., Robert, C., Blanc, G., Fournier, PE., Parinello, H., Claverie,. PE., Raoult, D**. "The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite." PLoS Biol. **3**(8): 1391-1402.

24. **Olsen, G.J., Woese, C.R.** (1993) "Ribosomal RNA: a key to phylogeny." FASEB J **7**:113-123.

25. **Olsen, G.J., Woese, C.R., Overbeek, R.** (1994). "The winds of (evolutionary) change: breathing new life into Microbiology." Journal of Bacteriology **176**(1): 1-6.

26. **Paz A., Kirzhner V., Nevo E., and Korol A.** (2006). Coevolution of DNA-interacting proteins and genome "dialect". Mol Biol Evol **23**(1):56-64.

27. **Sankoff, D.** (1997). "Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome." PNAS **89**(14): 6575-6579.

28. **Smith, D.G.E. and Lawson, G.H.K.** Lawsonia intracelluaris: getting inside the pathogenesis of proliferative enteropathy (2001). Vet. Microbiol. **82**(4):331-345.

29. **Snel, B., Bork, P., Huynen, M.A.** (1999) Genome phylogeny based on gene content. Nat Genet **21**:108-10.

30. **Suzuki, H., M. Sota, C. J. Brown, and E. M. Top.** 2008. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. Nucleic Acids Res. **36:**e147.

31. **Tison, D. L., D. H. Pope, et al**. (1980). "Growth of *Legionella pneumophila* in association with blue-green algae (cyanobacteria)." Appl. Environ. Microbiol. **39**(2): 456-459.

32. **Woese, C.** (1987). "Bacterial Evolution." Microbio. Reviews **51**(2): 221-271.

33. **Zhang, Z., S. Schwartz, et al**. (2000). "A greedy algorithm for aligning DNA sequences." J Comput Biol **7**(1-2): 203-214.