An investigation into the contentious phylogenetic results of the *Mixta* species

Kim N. Hinz, John Stavrinides1

Department of Biology  
University of Regina  
3737 Wascana Parkway  
Regina, Saskatchewan, Canada  
S4S0A2

1**Correspondence**:  
John Stavrinides  
Department of Biology  
University of Regina  
3737 Wascana Parkway  
Regina, Saskatchewan  
S4S0A2  
Ph: (306) 337-8478  
Email: [john.stavrinides@uregina.ca](mailto:john.stavrinides@uregina.ca)

**Keywords:** *Mixta*, phylogenetic analyses, average nucleotide identity, average amino acid identity

# Abstract

*Mixta* is a recently described genus within the bacterial family *Erwiniaceae* and is closely related to *Pantoea*, *Erwinia*, and *Tatumella*. Depending on the genes and method used for phylogenetic analysis, the genus with which *Mixta* shares the most recent ancestor differs. This study aimed to determine the cause behind these contentious results. Ten complete or partial genomes – two *Mixta* species, six species from other *Erwiniaceae* genera, and two outgroups – were retrieved from NCBI and annotated. Homologous genes were extracted yielding a dataset of 799 genes. The genes were aligned with the ClustalW algorithm, and MEGA-CC was used to calculate the most appropriate model, distance matrices, and phylogenetic trees for both nucleotide and amino acid sequences. Nucleotide and amino acid identity analyses were also done using the programming language R. *Pantoea* was the closest relative to the *Mixta* species in most analyses; however, results were not consistent. Some genes were also found to be more similar to other, non-*Pantoea* species. Diligence must be given to selecting genes for phylogenetic analysis and the method chosen to prevent any xenologous signal from distorting the actual relationships. Furthermore, future research should consider that different phylogenetic analyses may provide different results.

# Introduction

The bacterial family *Erwiniaceae* includes diverse Gram-negative [1, 2] organisms, including the genera *Mixta*, *Pantoea*, *Erwinia*, *Tatumella*, *Buchnera*, *Wigglesworthia*, *Phaseolibacter*, and *Kalamiella* [REF CRAIG]. *Mixta* is a relatively newer genus first identified by Palmer et al. in 2017 [3] and later described by the same researchers in 2018 [4]. Currently, there are only five *Mixta* species validly published [5]. *M. calida* and *M. gaviniae* (previously species of *Pantoea*) were initially found in infant formula and its production environment [6]. Fritz et al. [7] demonstrated that some strains of *M. calida* were linked with post-surgical meningitis in humans. The other *Mixta* species include *M. theicola* which was isolated from black tea extract [8], *M. intestinalis* which was isolated from a healthy human fecal sample [2], and *M. tenebrionis* which was isolated from the gut microbiota of a plastic-eating mealworm [9]. The first two were previously named as *Pantoea* species.

*Mixta*’s position relative to other members of the *Erwiniaceae*, in particular to *Pantoea*, *Erwinia*, *Tatumella*, differs vastly between studies [2–4, 10–14]. Prakash et al. [2], Palmer et al. [3], Rezzonico et al. [12], and Gueule et al. [14] all used *gyrB*, *rpoB*, *atpD*, and *infB* nucleotide gene sequences in their phylogenetic analyses. Prakash et al. [2] conducted a neighbour-joining statistical method while the others used a maximum-likelihood (ML) method [3, 12, 14]. Three of the studies – Rezzonico et al. [12], Prakash et al. [2], and Gueule et al. [14] – placed the *Mixta* species within the *Pantoea* genus. In contrast, the analysis in Palmer et al. [3] identified *Mixta* outside of the other three genera. A separate study by Palmer et al. [4], Rezzonico et al. [12], and Brady et al. [13] also ran analyses with the four MLSA (multi-locus sequence analysis) genes as amino acid sequences. *Mixta* was found to be related closest to *Pantoea* by Palmer et al. [4]. Both Rezzonico et al. [12] and Brady et al. [13] suggested that *Mixta* was a part of the *Pantoea* genus and not separate. Additionally, Palmer et al. [3] computed two ML trees; one used 1039 core amino acid sequences, and the other used 52 ribosomal MLSA nucleotide sequences. The 1039 core gene tree suggested *Pantoea* and *Tatumella* were sister taxa followed by two *Mixta* species. However, the ribosomal MLSA gene tree proposed *Pantoea* and *Tatumella* were sister taxa followed by *Erwinia* and then by *Mixta*.

What remains unclear is the relative contribution of disaparate evolutionary histories of *Mixta* genes, tree-building methods, and sequence evolution on the phylogenetic position of the genus. This paper aims to determine the cause behind the *Mixta* species’ contentious phylogenetic results. Ten complete or partial genomes – two for each of *Mixta*, *Pantoea*, *Erwinia*, and *Tatumella* and two outgroups – were retrieved from a public database. Homologous genes were extracted from these genomes in both the nucleotide and amino acid sequences. Distance matrices were exported from the phylogenetic trees, and these were used to determine which species were most closely related to *Mixta* for each gene. Nucleotide and amino acid identity analyses were also conducted for each gene to compare the results between evolutionary history and sequence similarity. Phylogenetic trees were constructed for combinations of different gene sequences.

# Materials and Methods

## Genomes and Gene Sets

Ten publicly available full or partial genomes of the species’ type strains were retrieved from the National Center for Biotechnology Information (NCBI). Two representative genomes were used for each of the four *Erwiniaceae* genera while *Enterobacter cloacae* (*Enterobacterales*) and *Pseudomonas syringae* (*Pseudomonaceae*) were used as outgroups (Table 1). The genomes were annotated with Prokka v1.14.1 [15] using default parameters, and homologous genes were extracted using GET\_HOMOLOGUES software package [16] with the bidirectional best-hit search algorithm using default parameters. GET\_HOMOLOGUES extracted 954 homologous genes between the ten species.

Table 1. Genomes used in this study.

|  |  |  |  |
| --- | --- | --- | --- |
| Species | Strain | GenBank Accession No. | Level |
| Mixta calida | DSM\_22759 | GCA\_002953215.1 | Complete |
| Mixta gaviniae | DSM\_22758 | GCA\_002953195.1 | Complete |
| Pantoea agglomerans | NBRC\_102470 | GCA\_001598475.1 | Contigs |
| Pantoea septica | LMG\_5345 | GCA\_002095575.1 | Contigs |
| Erwinia amylovora | CFBP\_1232 | GCA\_000367625.2 | Contigs |
| Erwinia tasmaniensis | ET1/99 | GCA\_000026185.1 | Complete |
| Tatumella ptyseos | NCTC\_11468 | GCA\_900478715.1 | Complete |
| Tatumella saanichensis | NML\_06-3099 | GCA\_000439375.1 | Contigs |
| Enterobacter cloacae subsp cloacae | ATCC\_13047 | GCA\_000025565.1 | Complete |
| Pseudomonas syringae pv syringae | ICMP\_3023 | GCA\_001401075.1 | Scaffold |

In R [17], two filter functions were written for the nucleotide sequence gene files. The first filter kept genes files that contained ten, single-copy sequences that were approximately the same length. The shortest gene length allowed was 90% the length of the longest sequence in the gene file. Of the 954 homologous gene sets, 38 multi-copy genes were excluded, and another 175 were excluded for not meeting the length requirements.

A second filter was applied to these 175 genes to identify any genes wherein both representatives of each genera were approximately the same length. The filter split the genes into two groups: sequences that met the original 90% length cutoff and sequences that were between 80 and 90% the length of the longest sequence. If both representatives of a genus were in the same group, then the gene was added to the working dataset. If at least one of the sequences did not meet the 80% cutoff, then the gene was excluded. An additional 58 genes were added to the working dataset, resulting in 799 genes. The amino acid sequences for these genes were subsequently downloaded.

## Gene Alignment

Gene alignments were done in R using a ClustalW algorithm. The fasta files were read into the session using the readDNAStringSet() function (package: Biostrings) [18]. The function called to do the alignments was the msaClustalW() function (package: msa) [19] using default parameters and 100 iterations. The aligned sequences were then converted into a comma-separated values (CSV)-compatible format using msaConvert() (package: msa)[19] and written into a new fasta file using the write.fasta() function (package: seqinr) [20]

## Model Selection

The best model of sequence evolution was identified for each nucleotide and amino acid alignment using Model Selection as implemented in MEGA-CC [21]. Default parameters were used for model testing. The output was a CSV file for each gene. These CSV files were read into R, and the model for phylogenetic analysis was extracted based on the lowest Bayesian Information Criterion (BIC).

## Phylogenetic Analysis

Phylogenetic analysis was performed in MEGA-CC using maximum likelihood with 500 bootstrap replications, the model, and rate patterns to create a phylogenetic tree for each gene. If a model required a Gamma distribution, the number of discrete Gamma categories was set to 5, which is the default.

Distance matrices were extracted from each of the phylogenetic trees and saved as Excel files. These files were then read into the R session. The genetic distances between *Mixta* and the other species were extracted for further analysis. Tables were created in R Markdown using the kable() function (package: knitr) [22] and figures were created using the ggplot() function (package: ggplot2) [23].

## Nucleotide and Amino Acid Identity

Each of the non-*Mixta* sequences were aligned to one of the corresponding *Mixta* sequences in turn. Alignments were performed in R as described above, and the aligned sequences were then split into the individual sites using the str\_split() function (package: stringr) [24]. A function was created using base R code to calculate the number of identical sites between each sequence and the *Mixta* homologues. Values that equal or are close to 100% are more identical to the *Mixta* sequences.

# Results

After the two filters, 799 of 954 homologous genes remained. The genes included those encoding rRNA sequences, enzymes, transporters, and housekeeping genes and were scattered across the M. calida and M. gaviniae genomes.

## Best Model vs Best Available Model

In MEGAX [25], the user can either calculate genetic distance or estimate a phylogenetic tree. However, although model testing assesses all possible models available in the software, all models are not available for genetic distance analysis. The GTR (General Time Reversible) and HKY (Hasegawa-Kishino-Yano) models and the +I (invariant sites) parameter are available only for phylogenetic tree analysis.

The genetic distances for the nucleotide sequences (NTS) were estimated using both the recommended (RM) and best available models (BAM) for distance analysis to determine if different results would arise when using a model that does not explain most of the genetic variance. The RM is the model with the lowest Bayesian Information Criterion (BIC) among the available options. First, phylogenetic gene trees were computed using the recommended model from model testing. The software suggested twelve unique models for this method (Table 2). Genetic distance matrices were then extracted from the trees. Second, distance matrices were calculated directly using the BAM for calculating pairwise distances. Only four models were recommended for genetic distance analysis (Table 2). Of the 799 NTS, 242 genes required a different model between the two analyses, and the mean BIC difference between the two models was 13.8 (range: 0.1, 122.6).

Table 2. The number of genes that required each model according to the best model with the lowest BIC and the best available model for genetic distance analysis.

|  |  |  |
| --- | --- | --- |
| Model | Best Model | Best Available Model |
| GTR+G | 89 | NA |
| GTR+G+I | 42 | NA |
| HKY+G | 22 | NA |
| HKY+G+I | 1 | NA |
| K2 | 1 | 3 |
| K2+G | 94 | 98 |
| K2+G+I | 1 | NA |
| K2+I | 2 | NA |
| T92+G | 313 | 398 |
| T92+G+I | 33 | NA |
| TN93+G | 149 | 300 |
| TN93+G+I | 52 | NA |

The closest relatives to both of the *Mixta* species were extracted from RM and BAM distance matrices. When the RM was used. *P. septica* was the closest relative to 66.3 and 66.5% of the *M. calida* and *M. gaviniae* sequences, respectively (Figure 1). These percentages increase by 12.6 and 14.4% for the *M. calida* and *M. gaviniae* sequences, respectively, when the BAM was used (Figure 1). Instead of *P. septica* when the RM is used, more genes were more closely related to *P. agglomerans*, *E. tasmaniensis*, *E. amylovora*, *T. saanichensis*, and *E. cloacae* (Figure 1). Additionally, there was more disparity between *M. calida* and *M. gaviniae* when the BAM was used (Figure 1).

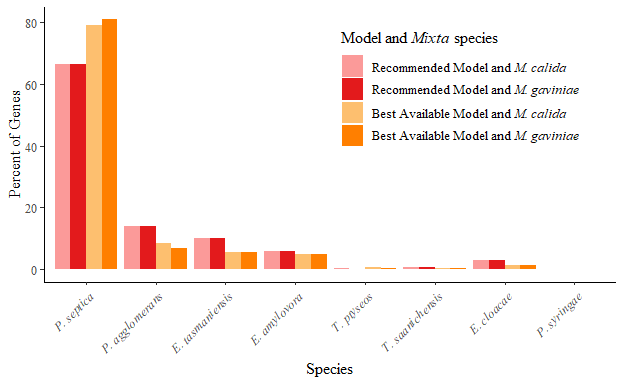


Figure 1. The percentage of genes (n = 799) that were most closely related to any of the eight non-*Mixta* species when using the recommended model from model testing versus the best available) model for genetic distance analysis in MEGAX. The sequences were nucleotide sequences. The species with the shortest genetic distance from the *Mixta* species is considered to be the closest relative. The pink and orange bars represent the percentage of *Mixta* genes that were most closely related to any of the non-*Mixta* species when using the recommended model and the best available model, respectively. The lighter shades represent *M. calida* and the darker shades represent *M. gaviniae*.

The RM and BAM were different for approximately 30.3% (n = 242) of the genes. Of these 242 genes, 20.2% (n = 49) were said to have different closest relatives between the two analyses. However, when the RM and BAM were the same (n = 557), 19.9% (n = 111) had different closest relatives.

## Closest Relative from Genetic Distance

Genetic distance gives a numerical estimate of evolutionary change between species, and the taxa with which a species of interest has the shortest distance is regarded as the closest relative. Since the phylogenetic trees were calculated using the RM, the distance matrices exported from the trees were kept for future analyses. *P. septica* was consistently the closest relative for most of the *Mixta* sequences (Figure 2). However, the number of genes most closely related *P. septica* decreased from 66.3 and 66.5% (*M. calida* and *M. gaviniae*, respectively) with the NTS by 23.7% with the amino acid sequences (AAS; Figure 2). Instead, the *Mixta* AAS were most closely related to the other taxa within the dataset except for *P. syringae* (Figure 2). Interestingly, the number of genes most closely related to *E. cloacae* increased between the NTS and the AAS from 2.9% by 3.1% (Figure 2). Additionally, 55.1 and 55.2% of the *M. calida* and *M. gaviniae* sequences, respectively, had the same closest relatives between the NTS and the AAS analyses.

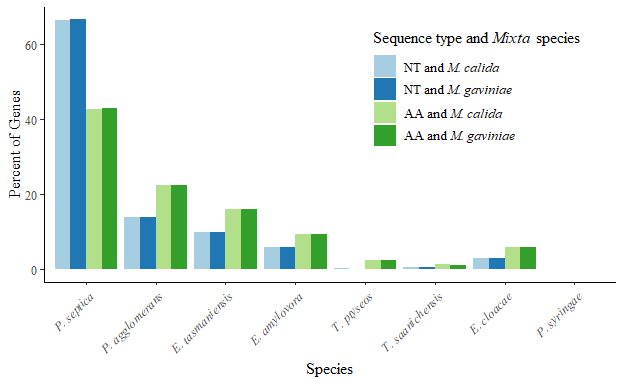


Figure 2. The percentage of genes (n = 799) that were most closely related to any of the eight non-*Mixta* species using the recommended models and genetic distances. The species with the shortest genetic distance from the *Mixta* species is considered to be the closest relative. The percentage of *Mixta* genes that were most closely related to any of the non-*Mixta* species are represented by the blue and green bars for the nucleotide and amino acid sequences, respectively. The lighter shades represent *M. calida* and the darker shades represent *M. gaviniae*.

## Spatial Patterns in the Closest Relatives from Genetic Distances

Circular plots were generated using the *Mixta* gene IDs to simulate their relative positions on the *Mixta* chromosomes and uncover positional and spatial patterns. The bars’ lengths and colours correspond to the species that was the closest relative to the *Mixta* gene at that position. The results between the two *Mixta* species for the closest relatives were the same except for one (*ttuB*) and two (*ttuB* and *secY*) genes for the NTS and the AAS, respectively; therefore, only the results for *M. calida* are shown.

Many of the genes most closely related to any of the other taxa in this study were evenly positioned around the *Mixta* chromosome in the NTS (Figure 3A). The exception is for the *Tatumella* species (pinks) and *E. amylovora* (dark green; Figure 3A). The former and most of the genes closely related to the latter were located on the right side of the chromosome (Figure 3A). Unexpectedly, genes closely related to *E. cloacae* (orange) were distributed throughout the *Mixta* genome.

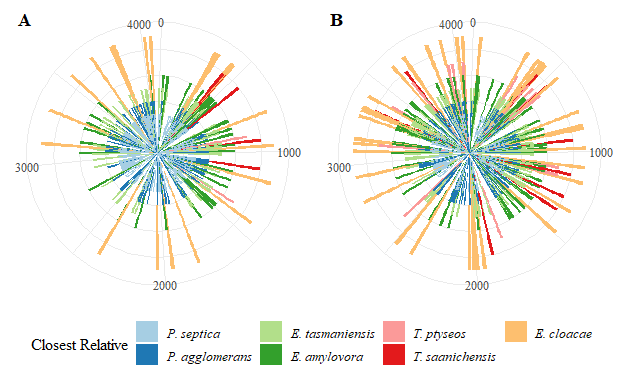


Figure 3. The closest relatives to *M. calida* genes for (**A**) nucleotide and (**B**) amino acid sequences around the *M. calida* chromosome. The length and colour of the bars represent the different species to which the gene is most closely related according to genetic distance. The *M. calida* genome has 4092 genes.

In the AAS analysis, the genes closely related to the *Tatumella* species were not exclusive to the right half (Figure 3B). Additionally, the genes closely related to the *Erwinia* species and *E. cloacae* were more evenly distributed across the genome (Figure 3B). Similar to the NTS analysis, no gene clusters related to a single taxon were evident (Figure 3B).

## Closest Relative from Average Site Identity

Site identity is the amount of similarity between two sequences and accounts for the sequence length. The non-*Mixta* sequences were aligned to each of the *Mixta* homologues, so the percent identity was relative to *Mixta*. The closest relative to both *Mixta* species is defined as the taxon with the highest similarity to *Mixta*.

Similar to the closest relative according to genetic distances, *P. septica* was the closest relative for most *Mixta* sequences (Figure 4). With the NTS, *P. septica* was the closest relative for 79.3 and 80.6% of the *M. calida* and *M. gaviniae* sequences, respectively (Figure 4). However, these percentage decreased by 31.5 and 33.0%, respectively, with the AAS (Figure 4). The *Mixta* AAS were instead more closely related to other taxa except for *P. syringae* (Figure 4). Unexpectedly, 1.9 and 1.3% of the *M. calida* and *M. gaviniae* sequences, respectively, were most similar to the *E. cloacae* sequences (Figure 4). Additionally, these percentages increased by 3.1 and 2.5% for *M. calida* and *M. gaviniae*, respectively, for the AAS (Figure 4).

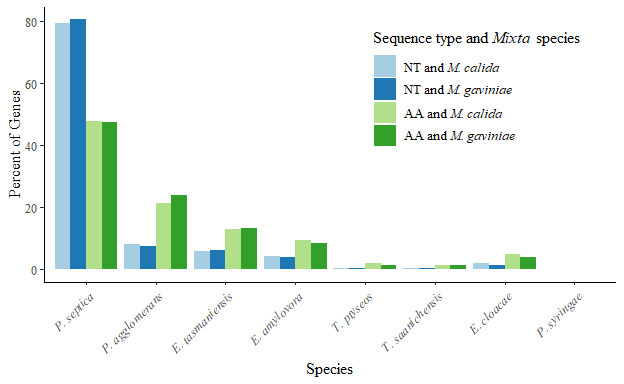


Figure 4. The percentage of genes (n = 799) that were most closely related to any of the eight non-*Mixta* species using percent site identity. The species with the highest percent identity to the *Mixta* species is considered to be the closest relative. The percentage of *Mixta* genes that were most closely related to any of the non-*Mixta* species are represented by the blue and green bars for the nucleotide and amino acid sequences, respectively. The lighter shades represent *M. calida* and the darker shades represent *M. gaviniae*.

When comparing the site identity analysis results, 89.2% of the NTS and 85.2% of the AAS agreed on the same closest relative between the two *Mixta* species. In contrast, only 55.1 and 53.9% of the *M. calida* and *M. gaviniae* sequences, respectively, had the same results between the NTS and AAS analyses.

## Spatial Patterns from Average Site Identity

Positional effects were investigated in the site identity results. Only the results of *M. calida* are shown because the *M. gaviniae* results showed similar patterns. The genes most similar to either of the *Pantoea* species and *E. cloacae* were evenly dispersed across the *Mixta* chromosome for both the NTS and AAS (Figure 5A and 5B). Similar to the genetic distance results, genes that are labelled as closely related to either of the *Tatumella* species are located only on the right side of the chromosome, and genes closely related to *E. amylovora* are predominantly on the right side as well (Figure 5A). Unexpectedly, the lower left quadrant of the chromosome primarily contains genes closely related to either of the *Pantoea* species, with only a few exceptions (Figure 5A). In the AAS analysis, genes closely related to any of the other taxa were more evenly distributed across the genome (Figure 5B).

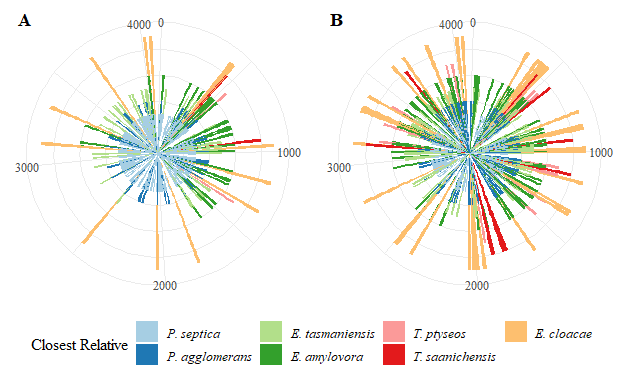


Figure 5. The closest relatives to *M. calida* genes for (**A**) nucleotide and (**B**) amino acid sequences around the *M. calida* chromosome. The length and colour of the bars represent the different species to which the gene is most closely related according to site identity. The *M. calida* genome has 4092 genes.

## A Comparison of the Four Analyses

The four above analyses (genetic distance and site identity, both using NTS and AAS) gave slightly different results for the *Mixta* sequences. The four analyses agreed upon the closest relative for approximately 38.7% of each of the *M. calida* and *M. gaviniae* sequences. The number of sequences closely related to one of the other taxa differed between the two *Mixta* species (Table 3). Additionally, although these sequences had a consistent closest relative, 30 genes differed between them. For these 30 genes, there was a consensus for only one of the *Mixta* species and not the other. Nevertheless, 34.9% of the *M. calida* and *M. gaviniae* homologues agreed on the same species.

Table 3. The number of *M. calida* and *M. gaviniae* genes wherein all analyses gave the same species as the closest relative.

|  |  |  |
| --- | --- | --- |
| Species | # of M. calida genes | # of M. gaviniae genes |
| P. septica | 255 | 252 |
| P. agglomerans | 22 | 22 |
| E. tasmaniensis | 17 | 20 |
| E. amylovora | 8 | 7 |
| T. saanichensis | 1 | 1 |
| E. cloacae | 6 | 7 |

Displaying these results in a circular plot showed that genes closely related to any of the taxa were dispersed across the *M. calida* and *M. gaviniae* chromosomes (Figures 6A and 6B, respectively). Given that similar patterns were found in the four analyses, genes with a consistent closest relative are randomly distributed across the genome. The position within the genome does not affect whether the analyses agree (Figures 6A and 6B).

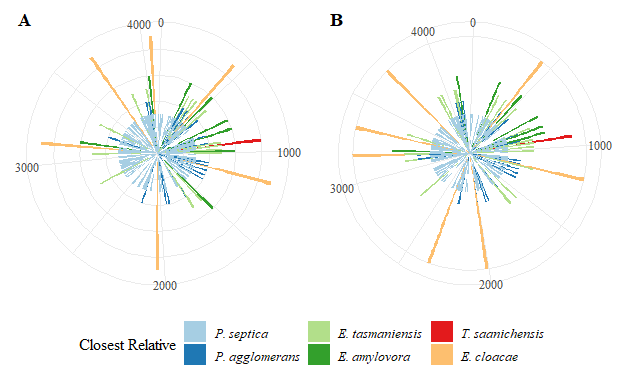


Figure 6. The closest relatives to the *M. calida* (**A**) and *M. gaviniae (****B****) genes that were agreed upon by all four analyses that were conducted in this study. The circular plot simulates the circular chromosome of the respective species. The length and colour of the bars represent the different species to which the gene is most closely related. The* M. calida\* and *M. gaviniae* genomes have 4092 and 4242 genes, respectively.

Interestingly, the four analyses agreed on a genus being the closest relative for a total of 63.4 and 63.2% of the *M. calida* and *M. gaviniae* sequences, respectively (Table 4). Unexpectedly, only 39 of the *M. calida* and the *M. gaviniae* homologues did not agree on which genus was the closest relative. Of these genes, there was a consensus for one of *M. calida* or *M. gaviniae* and not the other for 38 homologues. Only one gene (*ttuB*) did the analyses agree on *Tatumella* for *M. calida* and *Pantoea* for \*M. gaviniae.

Table 4. The number of *M. calida* and *M. gaviniae* genes wherein all analyses gave the same genus as the closest relative.

|  |  |  |
| --- | --- | --- |
| Genus | # of M. calida genes | # of M. gaviniae genes |
| Pantoea | 454 | 457 |
| Erwinia | 44 | 39 |
| Tatumella | 3 | 2 |
| Enterobacter | 6 | 7 |

## Phylogenetic Gene Trees for *secY* and *ttuB*

In the genetic distance analysis, only one (*ttuB*) and two (*secY* and *ttuB*) genes differed between the two *Mixta* species for the NTS and AAS, respectively. The phylogenetic trees for the *secY* gene showed that the gene is well-conserved across the *Erwiniaceae* taxa (Figures 7A and 7B). When using the NTS, the *Mixta* species grouped together and fell outside of the other three *Erwiniaceae* taxa (Figure 7A). When using the AAS, the *Mixta* species did not group together (Figure 7B). Additionally, *M. calida* had a slightly shorter branch length to the *Erwinia* species, whereas *M. gaviniae* had a somewhat shorter branch length to *P. septica* (Figure 7B). However, the nodes separating the *Mixta* species from the other taxa were not well-supported (Figures 7A and 7B). Unexpectedly, *P. septica*’s homologue of *secY* was most closely related to *E. cloacae* for the AAS (Figure 7B) whereas *E. cloacae* fell outside of the *Erwiniaceae* for the NTS (Figure 7A).

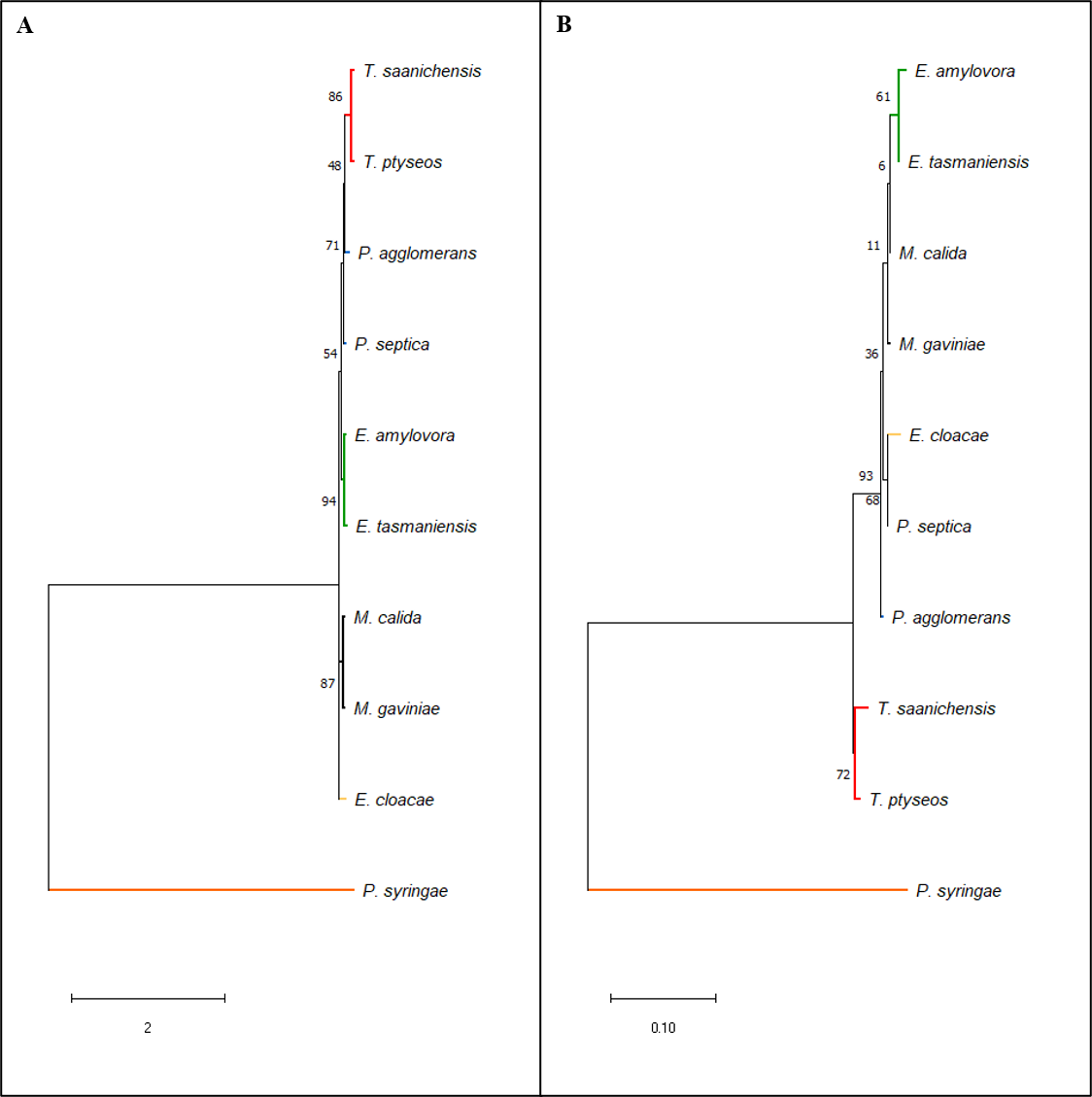


Figure 7. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the nucleotide (**A**) and amino acid (**B**) sequences of the gene *secY*. HKY+G and LG+G models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

The phylogenetic trees for the *ttuB* gene showed a more apparent separation between the two *Mixta* homologues (Figures 8A and 8B). For both the NTS and AAS, *M. calida* was more closely related to *Tatumella* followed by *E. cloacae* and *M. gaviniae* was more closely related to *Pantoea* followed by *Erwinia* (Figures 8A and 8B). Unlike with *secY*, many of the nodes were well-supported (Figures 8A and 8B).

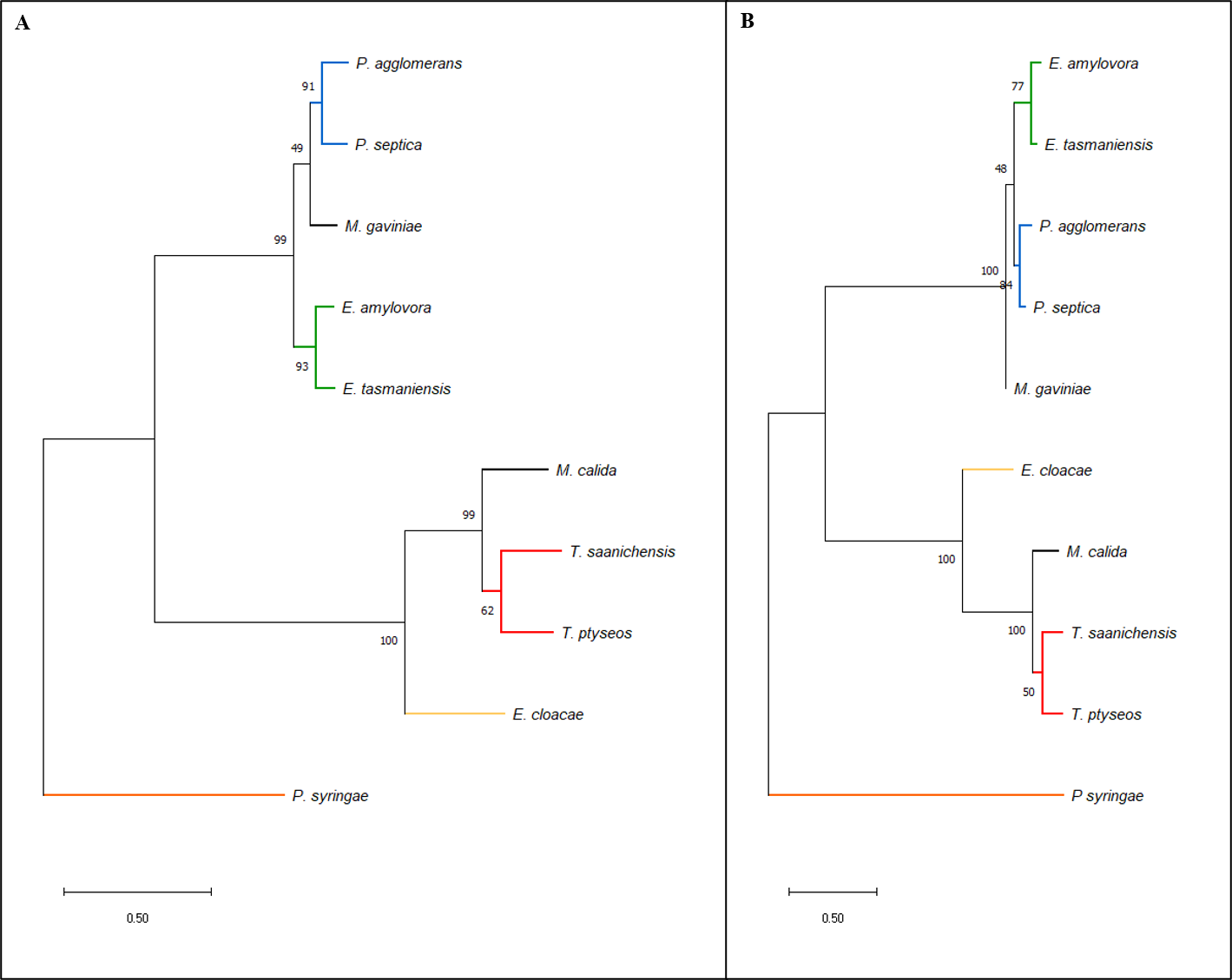


Figure 8. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the nucleotide (**A**) and amino acid (**B**) sequences of the gene *ttuB*. GTR+G and LG+G models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

## Phylogenetic Trees from MLSA Genes

Phylogenetic analyses often use several concatenated essential gene sequences to infer relatedness, namely *atpD*, *gyrB*, *infB*, and *rpoB*. However, *atpD* and *gyrB* did not meet the length requirements during the filtering process. Therefore, genes encoding the other subunits of ATP synthase (*atp*), translation initiation factor IF (*inf*), and DNA-directed RNA polymerase (*rpo*) were concatenated to create phylogenetic trees. The resulting trees showed that *Mixta* are more closely related to *Pantoea* followed by *Erwinia* (Figures 9A and 9B). With the NTS, a well-supported bifurcation occured between *Mixta* and *Pantoea* (Figure 9A). However, this did not happen with the AAS (Figure 9B). Instead, a somewhat-supported bifurcation occured first between *Pantoea* and *Erwinia* and then a weakly-supported bifurcation between these two and *Mixta* (Figure 9B).

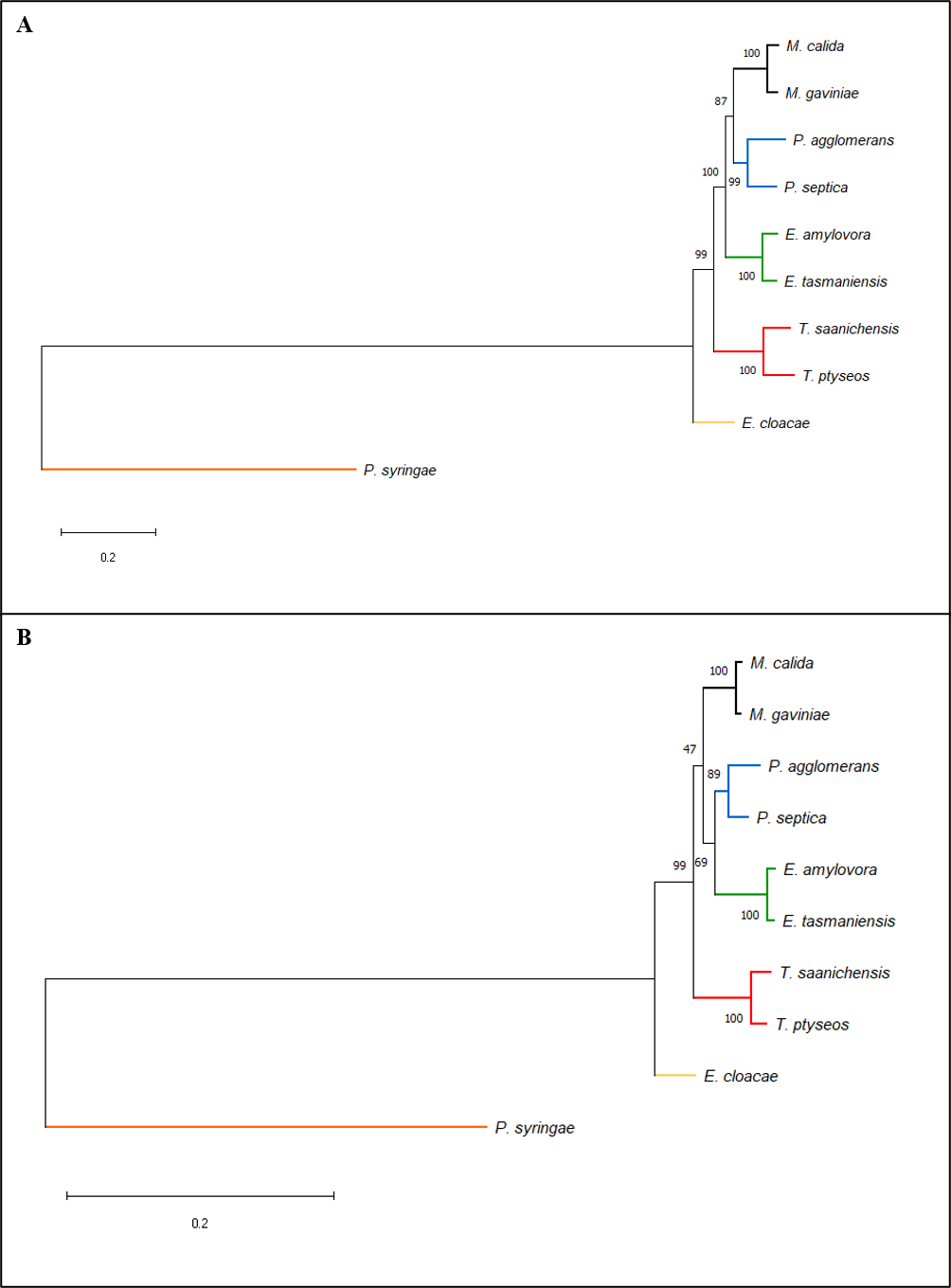


Figure 9. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the concantenated nucleotide (**A**) and amino acid (**B**) sequences of the genes *atpA*, *atpB*, *atpC*, *atpE*, *atpF*, *atpG*, *atpH*, *infA*, *infB*, *rpoA*, *rpoB*, *rpoC*, *rpoD*, and *rpoH*. GTR+G+I and LG+G+F models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. The nucleotide model allowed for some sites to be evolutionary invariable. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

The four analyses conducted in this study agreed on the closest relative for only four of these genes – *atpA*, *atpG*, *rpoB*, and *rpoC* – for both *M. calida* and *M. gaviniae* (Figures 10A and 10B). This closest relative to these genes were *P. septica*. The other genes were more closely related to *P. agglomerans*, *E. tasmaniensis*, *E. amylovora*, and *E. cloacae* and these varied amongst the analyses (Figures 10A and 10B). One only of the genes, *rpoD*, was closely related to *E. cloacae* by one of the analyses; the NTS genetic distance analysis labelled both the *Mixta* species as being closely related to *E. cloacae* and the remaining three analyses labelled this gene as being closely related to *P. septica* (Figures 10A and 10B).

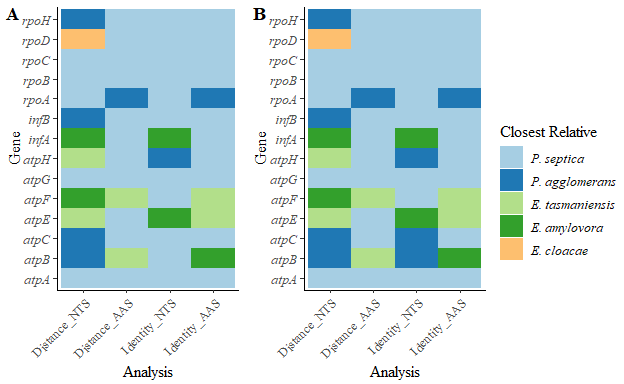


Figure 10. The closest relatives to the *M. calida* (**A**) and \*M. gaviniae (**B**) MLSA genes for each of the four analyses. The colour of the tiles represent the different species to which the gene is most closely related.

## Phylogenetic Trees Showing Other Results

Several genes were found to be more closely related to taxa other than *P. septica* or *P. agglomerans*. Thus, gene sequences wherein the four analyses agreed upon each of the other taxa were concatenated to compute phylogenetic trees. There was a consensus for 32 genes that the closest relative was not either of the *Pantoea* species: 17 were more closely related to *E. tasmaniensis*, 8 to *E. amylovora*, 1 to *T. saanichensis*, and 6 to *E. cloacae*.

The genes most closely related to *E. tasmaniensis* were *fabG*, *folK*, *greB*, *hemE*, *hscA\_2*, *ibaG*, *proA*, *rppH*, *sbp*, *tusA*, *tusE*, *uppS*, *ycfH*, *yffB*, and 3 hypothetical proteins. The phylogenetic trees resulting from the concatenated NTS (Figure 11A) and AAS (Figure 11B) showed strong support for the *Mixta* species being more closely related to *Erwinia*. Interestingly, *E. cloacae* fell outside of the *Erwiniaceae* for only the AAS (Figure 11B). However, the bifurcations between *E. cloacae* and the other taxa are not well-supported for either sequence type (Figures 11A and 11B).

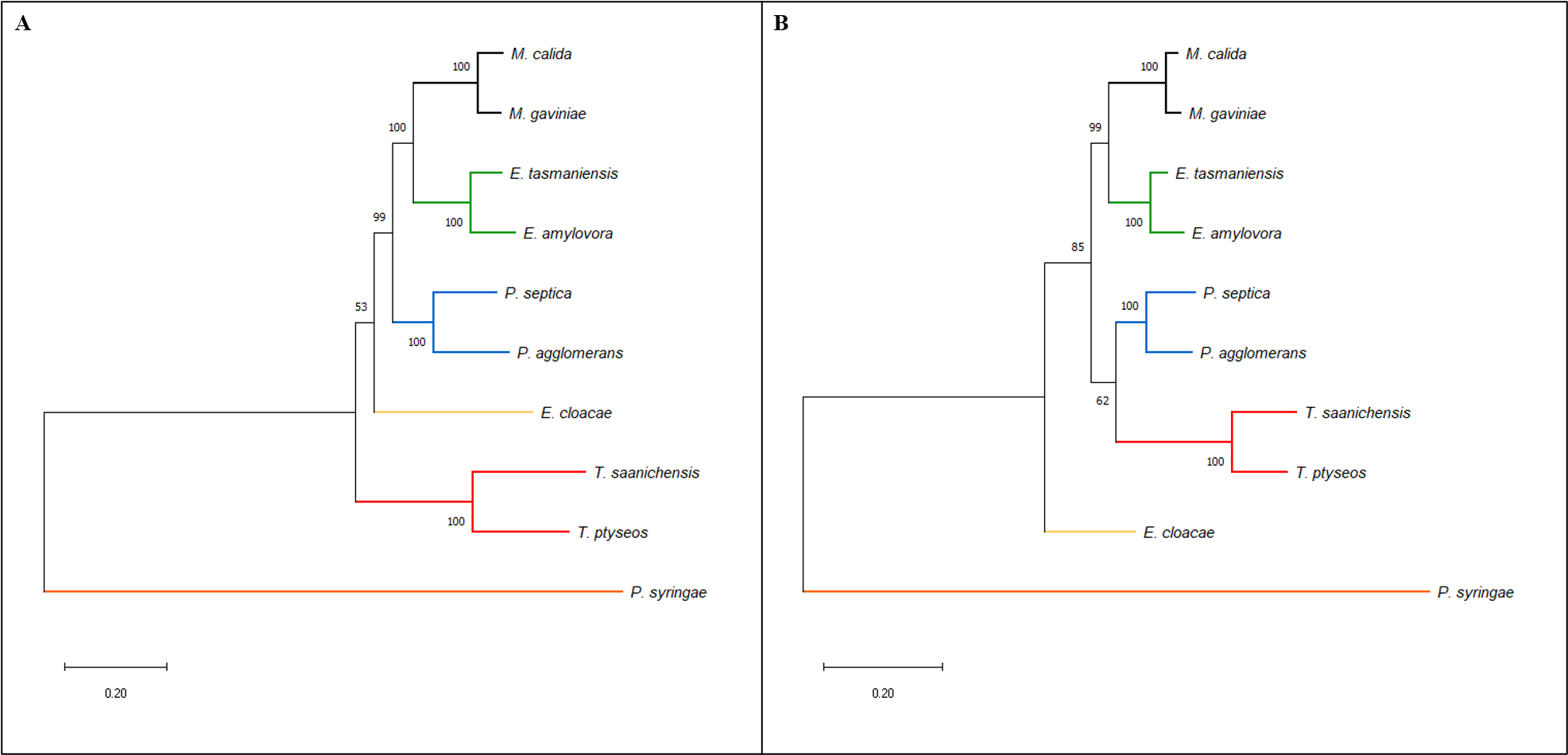


Figure 11. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the concantenated nucleotide (**A**) and amino acid (**B**) sequences of the genes *fabG*, *folK*, *greB*, *hemE*, *hscA\_2*, *ibaG*, *proA*, *rppH*, *sbp*, *tusA*, *tusE*, *uppS*, *ycfH*, *yffB*, and 3 hypothetical proteins. The four analyses in this study agreed on *E. tasmaniensis* as being the closest relative for these 17 genes. GTR+G+I and LG+G+F models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. The nucleotide model allowed for some sites to be evolutionary invariable. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

The eight genes most closely related to *E. amylovora* were *acuI*, *coaBC*, *dapB*, *grpE*, *mtnC*, *rlmJ*, *ssuE*, and a hypothetical protein. Similar to the results for *E. tasmaniensis*, the bifurcations between *Mixta* and *Erwinia* are well-supported for both the NTS (Figure 12A) and AAS (Figure 12B). Additionally, *E. cloacae*’s position relative to *Mixta* and *Erwinia* is well-supported for the NTS, but its position is ambiguous for AAS (Figures 12A and 12B).

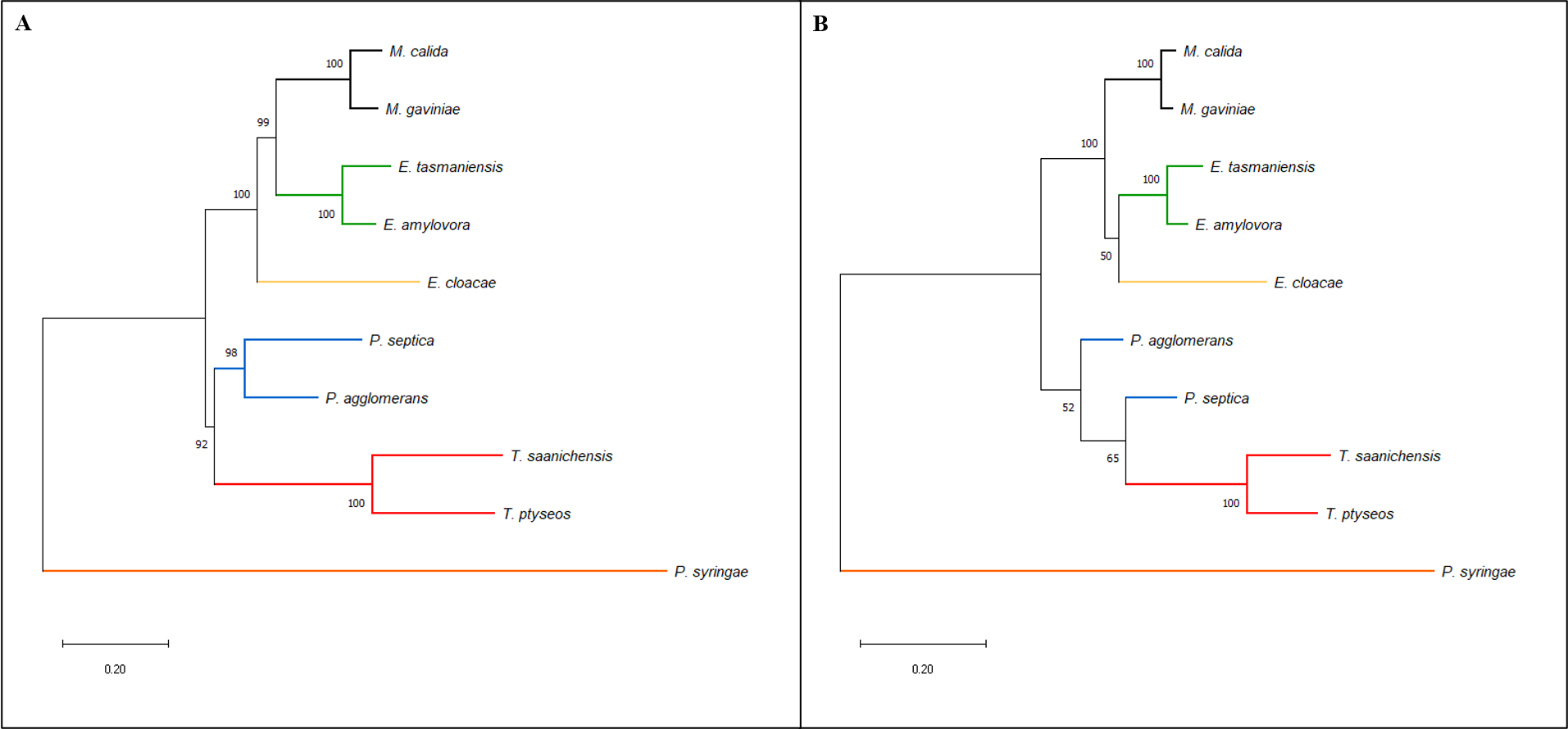


Figure 12. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the concantenated nucleotide (**A**) and amino acid (**B**) sequences of the genes *acuI*, *coaBC*, *dapB*, *grpE*, *mtnC*, *rlmJ*, *ssuE*, and a hypothetical protein. The four analyses in this study agreed on *E. amylovora* as being the closest relative for these 8 genes. GTR+G+I and LG+G+F models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. The nucleotide model allowed for some sites to be evolutionary invariable. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

Only one gene, *tsf*, was most closely related to *T. saanichensis*. Nevertheless, this relationship is well-supported for both the NTS and AAS (Figures 13A and 13B). Unexpectedly, the two *Tatumella* species were not grouped for the AAS (Figure 13B).

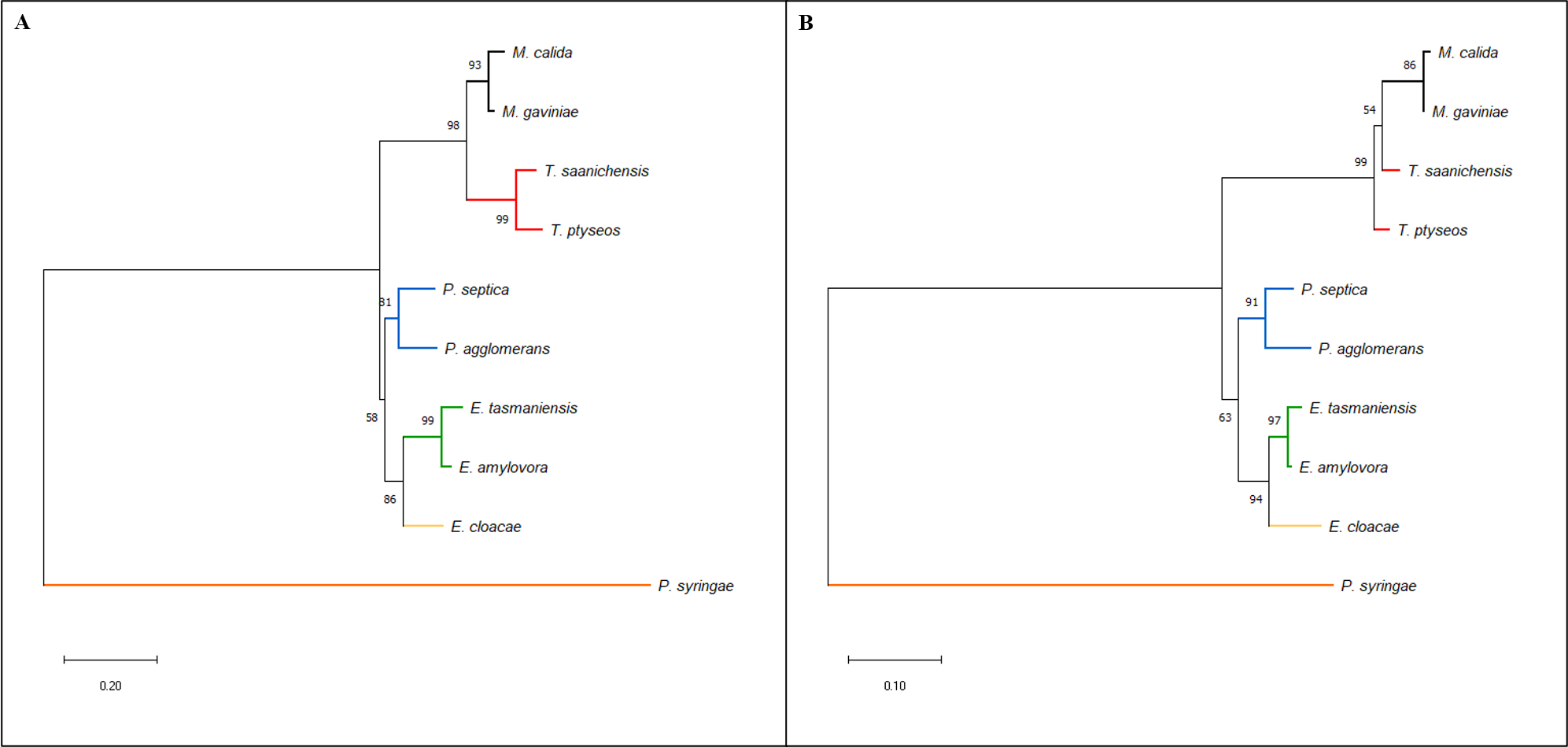


Figure 13. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the concantenated nucleotide (**A**) and amino acid (**B**) sequences of the gene *tsf*. The four analyses in this study agreed on *T. saanichensis* as being the closest relative for this one gene. K2+G and LG+G models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

The six genes most closely related to the outgroup species *E. cloacae* were *bioD1*, *cueR*, *hemN\_2*, *iscS*, *rpsM*, and *yqjC*. The relationship between *Mixta* and *E. cloacae* is well-supported in the phylogenetic trees computed from the concatenated NTS (Figure 14A) and AAS (Figure 14B). Both of the trees showed similar results as the *Tatumella* species were the next closest relative to *Mixta* with the same main difference being weaker branch support for the AAS (Figures 14A and 14B). Additionally, *Pantoea* and *Erwinia* fell into their own group (Figures 14A and 14B).

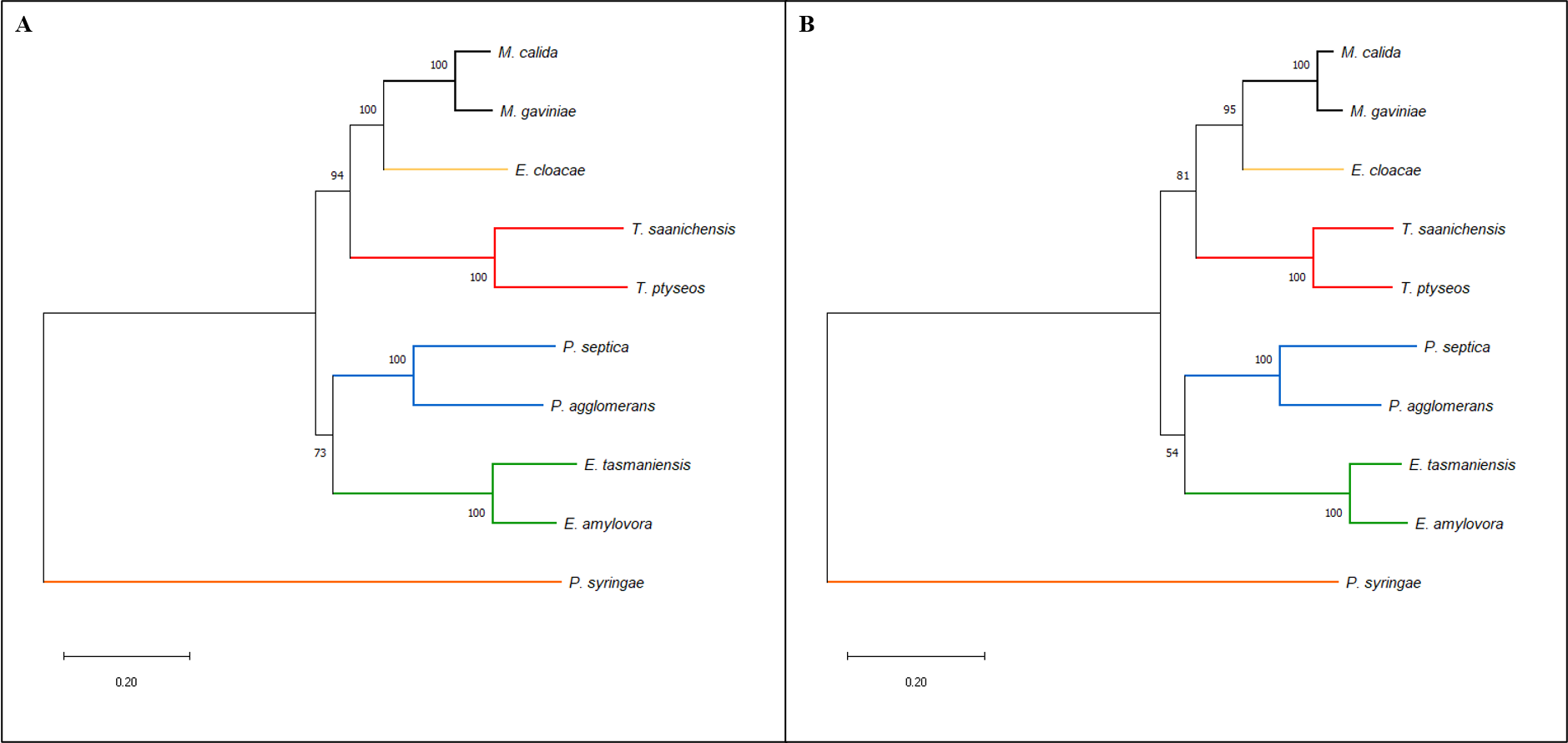


Figure 14. Maximum-likelihood phylogenetic trees for members of the *Erwiniaceae* with *E. cloacae* and *P. syringae* as outgroups. The trees were reconstructed from the concantenated nucleotide (**A**) and amino acid (**B**) sequences of the genes *bioD1*, *cueR*, *hemN\_2*, *iscS*, *rpsM*, and *yqjC*. The four analyses in this study agreed on *E. cloacae* as being the closest relative for these 6 genes. GTR+G+I and LG+G+F models were used for the nucleotide and amino acid sequences, respectively. Both trees required a discrete Gamma distribution (5 categories) to model the evolutionary rate differences among sites. The nucleotide model allowed for some sites to be evolutionary invariable. Branch support was inferred from 500 bootstrap replicates; the percentage of trees in which the associated taxa clustered together are adjacent to the branches. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. Evolutionary analyses were conducted in MEGAX.

# Discussion

Phylogenetic software programs vary on which models and parameters they have available, and these may vary within a program depending on the chosen method. MEGAX [25], for example, offers a wide range of models and additional parameters for the building of phylogenetic trees; however, the options are reduced for genetic distance calculation. A vital practise of selecting the most accurate model for your data is by conducting model testing. Model testing provides a list of models with additional parameters in the order of which model explains the most variation in the sequences’ differences. Selecting the most accurate model is always recommended; however, it is not always possible. My results show that slight differences in the taxa’s relationships occur when the more precise model is not used. For some genes, the phylogenetic signal of which taxon was the closest relative to *Mixta* was strong enough despite inaccuracies in the model. Nonetheless, there is a risk that the wrong species is labelled as the closest relative, and this poses problems if that species belongs to a separate genus or order. The species of interest may be wrongfully defined as belonging to a specific genus or order, or the genes may be inappropriately labelled as results from horizontal gene transfer (HGT).

Additionally, the chosen method may contribute to contentious results. In order to use the most accurate model, I had to first run a phylogenetic analysis and then extract the genetic distance matrices. For the best available model for genetic distance, I ran the genetic distance analysis directly and thus, forwent the phylogenetic tree step. Even when the most accurate and the best available models were the same, the results for the closest relative were different. This suggests that the two methods calculate taxa relatedness differently despite using the same models and parameters. One explanation is the Gamma parameter (+G). The genetic distance method uses a continuous distribution in which 1.00 is the default, while the phylogenetic tree method uses a discrete Gamma parameter set to 5 categories. Therefore, diligence must be given when selecting software programs, methods, and models.

Phylogenetic analysis is an accepted method for understanding the evolutionary relatedness [26], while site identity is used to confirm identification and taxon boundaries [27]. These two methods were conducted in this study as both nucleotide (NTS) and amino acid (AAS) sequences. *P. septica* was the closest relative to the majority of *Mixta* homologues in all four analyses. This lends strong support that *Mixta* and *P. septica* share a recent common ancestor, followed by other *Pantoea* species, *Erwinia*, and then *Tatumella*. This result agrees with *Mixta* being previously known as part *Pantoea*. Additionally, there were no *Mixta* sequences that were closely related to *P. syringae*, which demonstrates that this species is a suitable outgroup.

For both methods, fewer genes were labelled as being closely related to *P. septica* when using the AAS versus the NTS. This difference may be to a loss of phylogenetic signal when using amino acid sequences caused by codon degeneracy. Nevertheless, sequence type and different evolutionary models between NTS and AAS affect the results. AAS may be better suited for uncovering taxonomic relationships at the genus or higher level, and NTS may be more appropriate for the species level.

When considering the spatial effects of the closest relatives, the uniform pattern of the genes related to *P. septica* further lends support that this species and *Mixta* share a recent common ancestor. Additionally, the other taxa followed the same pattern. If the genes closely related to non-*Pantoea* species are a result of HGT, this suggests that HGT may be a common occurrence within the *Erwiniaceae*. A second explanation for the consistent distribution patterns in the closest relatives is that the genes are highly conserved. Two or more species may have been tied for the closest relative, and the R program may have selected one based on another characteristic, such as alphabetical order. Nevertheless, further research into these genes and the frequency of HGT events with *Mixta* and other genera within the *Erwiniaceae* is needed. Furthermore, for 279 (34.9%) homologues, the closest relative was consistent across all four analyses and between the two *Mixta* species. These results are more likely to be accurate because there was a strong phylogenetic signal and conserved site identity between the NTS and AAS. An additional 30 genes differed between *M. calida* and *M. gaviniae* on the closest species relative where there was a consensus for only one of the *Mixta* species. Codon degeneracy, codon bias, and HGT to one of the *Mixta* species may explain contentious results between the analyses and between the *Mixta* species.

The closest genus relative was also considered to account for sequence conservation and similar identity in recently diverged species. There was an agreement on the closest genus relative across the analyses for approximately 63% of the *Mixta* sequences. Of these, only 39 homologues differed between the *Mixta* species, with only one gene, *ttuB*, having a consensus on the closest genus relative for both of the *Mixta* species. A gene tree for *ttuB* reveals a possible HGT event wherein *M. calida* received a homologue from a *Tatumella* species. Because *ttuB* is a single-copy gene, the original version may have been deleted. Additionally, this explains why the genetic distance results for *M. calida* and *M. gaviniae* differed for both the NTS and AAS. A second gene, *secY*, also gave contentious results between the *Mixta* species in the AAS genetic distance analysis; however, this is likely due to its highly conservative nature given the short branch lengths in its gene tree.

Many previous studies into the *Erwiniaceae* have conducted phylogenetic analysis using four core genes: *atpD*, *gyrB*, *infB*, and *rpoB* [2, 3, 12, 14]. Two genes, *atpD* and *gyrB*, were removed from this study’s dataset during the filtering process; however, *infB*, *rpoB*, and twelve other related subunits were available. The concatenated NTS sequences produced a similar topology tree to Brady et al.’s [11] maximum-likelihood (ML) tree based on 16S rRNA gene sequences, and none of the trees based on the four genes. The differences may be caused by different software programs, phylogenetic models, and the addition of genes in my study. Additionally, the topology of the AAS sequences of the 14 core genes was not found by any previous studies [2–4, 11–14]. When the four genes were concatenated as AAS, Palmer et al. [4] found *Pantoea* and *Tatumella* formed sister taxa followed by *Mixta*. My analysis identified *Pantoea* and *Erwinia* as sister taxa followed by *Mixta*. Nonetheless, the lack of branch support leaves the true relationships ambiguous for the AAS.

Interestingly, phylogenetic trees showing a close relationship between *Mixta* and a non-*Pantoea* species were highly supported for both the NTS and AAS. Genes wherein there was a consensus between the four analyses were concatenated to create phylogenetic trees for each of *E. tasmaniensis*, *E. amylovora*, *T. saanichensis*, and *E. cloacae*. The strong branch support and long branch lengths suggest that these genes are likely candidates for HGT. Furthermore, the inclusion of these genes into future phylogenetic studies of the *Erwiniaceae* may result in contentious results and add a phylogenetic signal that is not due to vertical evolutionary history.

In conclusion, in phylogenetic analyses of the *Erwiniaceae*, diligence must be given to the software program, evolutionary model, sequence type, and included genes. Different programs may offer other methods, models, and parameters. Additionally, model testing is a valuable tool to ensure the most accurate estimation of evolutionary history is calculated. The sequence type – nucleotides versus amino acids – further play a role due to the difference in sequence lengths and codon degeneracy. The selected genes may also affect phylogenetic results as conserved genes may not conclusively separate species, and HGT genes may skew the taxa’s relationships. Traditional phylogenetic methods (i.e., genetic distance and phylogenetic trees) versus site identity may also give contentious results. Phylogenetic analyses estimate evolutionary history, and some models penalise specific mutations. However, site identity analyses provide the similarity between sequences and operate on the assumption that members of a clade should be genetically similar. Ultimately, it is best practice for future studies to be transparent in how they conducted their phylogenetic analyses to ensure reproducibility.

# Figure Legends

# Tables

# References

1. **Winslow C-EA, Broadhurst J, Buchanan RE, Krumwiede C, Rogers LA *et al.*** The families and genera of the Bacteria. Preliminary report of the Committee of the Society of American Bacteriologists on characterization and classiﬁcation of bacterial type. *Journal of Bacteriology* 1917;2:505.

2. **Prakash O, Nimonkar Y, Vaishampayan A, Mishra M, Kumbhare S *et al.*** Pantoea intestinalis sp. nov., isolated from the human gut. *International Journal of Systematic and Evolutionary Microbiology* 2015;65:3352–3358.

3. **Palmer M, Steenkamp ET, Coetzee MPA, Chan W-Y, Zyl E van *et al.*** Phylogenomic resolution of the bacterial genus Pantoea and its relationship with Erwinia and Tatumella. *Antonie van Leeuwenhoek* 2017;110:1287–1309.

4. **Palmer M, Steenkamp ET, Coetzee MPA, Avontuur JR, Chan WY *et al.*** Mixta gen. Nov., a new genus in the Erwiniaceae. *International Journal of Systematic and Evolutionary Microbiology* 2018;68:1396–1407.

5. **LPSN**. Genus Mixta. <https://lpsn.dsmz.de/genus/mixta> (2018).

6. **Popp A, Cleenwerck I, Iversen C, De Vos P, Stephan R**. Pantoea gaviniae sp. nov. and Pantoea calida sp. nov., isolated from infant formula and an infant formula production environment. *International Journal of Systematic and Evolutionary Microbiology* 2010;60:2786–2792.

7. **Fritz S, Cassir N, Noudel R, De La Rosa S, Roche P-H *et al.*** Postsurgical Pantoea calida meningitis: a case report. *Journal of Medical Case Reports* 2014;8:195.

8. **Kato Tanaka Y, Horie N, Mochida K, Yoshida Y, Okugawa E *et al.*** Pantoea theicola sp. nov., isolated from black tea. *International Journal of Systematic and Evolutionary Microbiology* 2015;65:3313–3319.

9. **Xia M, Wang J, Huo Y-X, Yang Y**. Mixta tenebrionis sp. nov., isolated from the gut of the plastic-eating mealworm Tenebrio molitor L. *International Journal of Systematic and Evolutionary Microbiology* 2020;70:790–796.

10. **Adeolu M, Alnajar S, Naushad S, Gupta RS**. Genome-based phylogeny and taxonomy of the ‘Enterobacteriales’: Proposal for enterobacterales ord. nov. divided into the families Enterobacteriaceae, Erwiniaceae fam. nov., Pectobacteriaceae fam. nov., Yersiniaceae fam. nov., Hafniaceae fam. nov., Morgane. *International Journal of Systematic and Evolutionary Microbiology* 2016;66:5575–5599.

11. **Brady CL, Goszczynska T, Venter SN, Cleenwerck I, Vos P de *et al.*** Pantoea allii sp. nov., isolated from onion plants and seed. *International Journal of Systematic and Evolutionary Microbiology* 2011;61:932–937.

12. **Rezzonico F, Smits THM, Born Y, Blom J, Frey JE *et al.*** Erwinia gerundensis sp. nov., a cosmopolitan epiphyte originally isolated from pome fruit trees. *International Journal of Systematic and Evolutionary Microbiology* 2016;66:1583–1592.

13. **Brady CL, Cleenwerck I, Westhuizen L van der, Venter SN, Coutinho TA *et al.*** Pantoea rodasii sp. nov., Pantoea rwandensis sp. nov. and Pantoea wallisii sp. nov., isolated from Eucalyptus. *International Journal of Systematic and Evolutionary Microbiology* 2012;62:1457–1464.

14. **Gueule D, Fourny G, Ageron E, Le Flèche-Matéos A, Vandenbogaert M *et al.*** Pantoea coffeiphila sp. nov., cause of the ‘potato taste’ of Arabica coffee from the African great lakes region. *International Journal of Systematic and Evolutionary Microbiology* 2015;65:23–29.

15. **Seemann T**. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

16. **Contreras-Moreira B, Vinuesa P**.. *Applied and Environmental Microbiology* 2013;79:7696–7701.

17. **R Core Team**. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. <https://www.r-project.org/> (2019).

18. **Pages H, Aboyoun P, Gentleman R, Debroy S**. Biostrings: Efficient manipulation of biological strings.

19. **Bodenhofer U, Bonatesta E, Horejs-Kainrath C, Hochreiter S**. msa: an R package for multiple sequence alignment. *Bioinformatics* 2015;31:3997–3999.

20. **Charif D, Lobry JR**. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman HE, Vendruscolo M (editors). *Structural approaches to sequence evolution: Molecules, networks, populations*. New York: Springer Verlag; 2007. pp. 207–232.

21. **Kumar S, Stecher G, Peterson D, Tamura K**. MEGA-CC: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* 2012;28:2685–2686.

22. **Xie Y**. knitr: A General-Purpose Package for Dynamic Report Generation in R.

23. **Wickham H**. *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer Verlag. <https://ggplot2.tidyverse.org> (2016).

24. **Wickham H**. stringr: Simple, Consistent Wrappers for Common String Operations. <https://cran.r-project.org/package=stringr> (2019).

25. **Kumar S, Stecher G, Li M, Knyaz C, Tamura K**. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution* 2018;35:1547–1549.

26. **Boundless**. Phylogenetic Analysis. <https://bio.libretexts.org/@go/page/9731> (2021).

27. **Ciufo S, Kannan S, Sharma S, Badretdin A, Clark K *et al.*** Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *International Journal of Systematic and Evolutionary Microbiology* 2018;68:2386–2392.