Untitled

Kimberly Hinz

2020/12/08

# Abstract

*Mixta* is a recently described genus within the bacterial family *Erwiniaceae* that 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, and identified genes that may have resulted from horizontal gene transfer. 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 for model testing and computing phylogenetic trees for each gene for both nucleotide and amino acid sequences. Distance matrices were then extracted from these trees. Nucleotide and amino acid identity analyses were also done in R programming language. *Pantoea* was the closest relative to the *Mixta* species in most of the analyses; however, results were not consistent across analyses. Some genes were also found to be more similar to other, non-*Pantoea* species. Diligence must be given to the selection of genes for phylogenetic analysis and to the method chosen to prevent xenologous signal from distorting the true relationships. Furthermore, future research should consider that different phylogenetic analyses may give different results.

# Introduction

The bacterial family Erwiniaceae includes diverse Gram-negative [1, 2] organisms including the genera *Mixta*, *Pantoea*, *Erwinia*, *Tatumella*, *Buchnera*, *Wigglesworthia*, *Phaseolibacter*, *Kalamiella* (LPSN, <http://www.bacterio.net/>). Three of these, *Mixta*, *Pantoea*, and *Erwinia*, have been the focus of numerous studies given their relevance to plant and human pathogens. For example, the genus *Erwinia* contains plant pathogens (e.g. E. amylovora) [3–5], epiphytes (e.g. E. tasmaniensis) [6, 7], species that may contribute to plant disease [8], and others that are insect-associated [9–11]. Similarly, *Pantoea*, a genus reported by Gavini et al.[12], includes species that inhabit a variety of environments [12–14], including human and plant pathogens [2, 15–20] as well as species associated with plant growth [21, 22], insects [23], and fungi [24, 25]. *Mixta*, originally *Pantoea*, is a novel genus first suggested by Palmer et al. [26] and then later proposed by the same researchers in 2018 [27]. Two species, M. calida and M. gaviniae (previously species of *Pantoea*), were originally found in infant formula and its production environment [13]. Fritz et al. [28] demonstrated that some strains of M. calida were linked with post-surgical meningitis in humans. Currently, there are only five *Mixta* species validly published (LPSN, <http://www.bacterio.net/>).

Understanding the phylogeny of bacterial species and genes allows researchers to investigate the evolutionary history of the species, their relatedness, and the individual genes that allow the species to survive under various conditions [26, 29–33]. The most common methods in phylogenetic analysis are those including 16S ribosomal RNA (rRNA) sequences [34], multi-locus sequence analysis (MLSA) with several housekeeping genes [34–36], and ribosomal MLSA with 53 structural ribosomal proteins [26, 29, 37]. However, the phylogenetic trees generated from these methods may not be very robust [34, 36, 38] because of the lower phylogenetic signal from the highly conserved housekeeping genes [34, 35, 39–42]. This is particularly seen within Erwiniaceae wherein most of our understanding of the phylogeny and interrelationships are based on the 16S rRNA genes [43–47] because of its ubiquity. However, it has insufficient resolution at the species level [36] and the genera *Pantoea*, *Erwinia*, *Tatumella*, and *Mixta* are often not monophyletic [48]. Another issue that may arise with MLSA is unknowingly including paralogues (i.e. genes originating from an intragenomic duplication event) or xenologues (i.e. genes originating from horizontal gene transfer) [26, 49].

The DNA topoisomerase (gyrase) subunit (gyrB), DNA-directed RNA polymerase subunit (rpoB), ATP synthase subunit (atpD), and initiation translation factor (infB) have been shown to be more reliable than 16S rRNA gene sequences for determining intra- and intergeneric relationships [50–53] and a robust method for the Erwiniaceae family [54]. The genera within Erwiniaceae are monophyletic when concatenated partial gyrB, rpoB, atpD, and infB gene sequences are used [38, 55, 56].

Nevertheless, despite efforts to find a suitable method for identifying the species tree for Erwiniaceae genera, *Mixta*’s position with relation to *Pantoea*, *Erwinia*, and *Tatumella* vastly differs between studies [2, 26, 27, 47, 48, 57–59]. Not only do different gene combinations provide different results, but different algorithms and models for alignments and phylogenetic trees also contribute to the contentious results [45, 60]. For example, Prakash et al. [2], Palmer et al. [26], Rezzonico et al. [48], and Gueule et al. [59] all used gyrB, rpoB, atpD, and infB nucleotide gene sequences in their phylogenetic analyses. Prakash et al. conducted a neighbour-joining statistical method while the others used a maximum-likelihood (ML) method. Three of the studies – Rezzonico et al., Prakash et al., and Gueule et al. – placed the *Mixta* species within the *Pantoea* genera whereas the analysis in Palmer et al. placed *Mixta* outside of the other three genera. Palmer et al. [27], Rezzonico et al. [48], and Brady et al. [58] also ran analyses with the four MLSA genes as amino acid sequences. These three studies show contentious results as Rezzonico et al. found that *Mixta* is related closest with *Tatumella*, Palmer et al. found that *Mixta* is related closest with *Pantoea*, and Brady et al. found *Mixta* within the *Pantoea* genus. In a separate study conducted by Palmer et al. [26], two ML trees were computed, one with 1039 core amino acid sequence and the other with 52 ribosomal MLSA nucleotide genes. Both ML trees contained species from *Mixta*, *Pantoea*, *Erwinia*, and *Tatumella*; however, the analysis with the core genes placed *Pantoea* and *Tatumella* as sister taxa followed by two *Mixta* species whereas the analysis with ribosomal MLSA genes placed *Pantoea* and *Tatumella* as sister taxa followed by *Erwinia* and then by the *Mixta* species.

There are two possible causes behind contentious results in phylogenetic analyses. One is the use of different methods, including the use of a different phylogenetic tree model from another study despite both studies conducting maximum likelihood trees and the same sequence type. Additionally, there is no common phylogenetic analysis model between nucleotide and amino acid sequences; therefore, evolutionary history is estimated differently. The other reason may be horizontal gene transfer (HGT) events within the *Erwiniaceae* or from a non-closely-related species. HGT events occur frequently within bacteria and often the genes provide the bacteria the ability to adapt to new environments and/or survive selective pressures.

The objective of this paper was to determine the cause behind the contentious phylogenetic results for the *Mixta* species. To accomplish this, homologous genes were extracted from ten complete or partial genomes - two for each of *Mixta*, *Pantoea*, *Erwinia*, and *Tatumella*, and two outgroups - that were retrieved from public databases. In this study, both the nucleotide and amino acid sequences for the same genes were separately analyzed. Model testing was done for each gene in the dataset and phylogenetic trees were computed using the recommended model. 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 in order to compare the results between evolutionary history and sequence similarity. Phylogenetic trees were then done with either one gene or multiple genes to confirm results.

# Materials and Methods

## Genomes and Gene Sets

Publicly available full genomes were retrieved from NCBI comprising of ten type strains from *Mixta*, *Pantoea*, *Erwinia*, *Tatumella*, *Enterobacter*, and *Pseudomonas* (Table 1). *Enterobacter cloacae* is in the same order of *Enterobacterales* as *Mixta*, *Pantoea*, *Erwinia*, and *Tatumella* whereas *Pseudomonas syringiae* is in the same class of *Gammaproteobacteria*. The genomes were annotated with Prokka v1.14.1 [61] using default parameters and homologous genes were extracted using GET\_HOMOLOGUES software package [62] 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 genome |
| Mixta gaviniae | DSM\_22758 | GCA\_002953195.1 | complete genome |
| 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 genome |
| Tatumella ptyseos | NCTC\_11468 | GCA\_900478715.1 | complete genome |
| Tatumella saanichensis | NML\_06-3099 | GCA\_000439375.1 | contigs |
| Enterobacter cloacae subsp cloacae | ATCC\_13047 | GCA\_000025565.1 | complete genome |
| Pseudomonas syringae pv syringae | ICMP\_3023 | GCA\_001401075.1 | scaffold |

In R [63], two filter functions were written to pass through gene files of the nucleotide sequences. Since both nucleotide and amino acid sequences were analyzed, the filters were only applied to the nucleotide sequences. A list of the passed nucleotide gene files was used to isolate the corresponding amino acid gene files to be used in later analyses.

The first filter passed through gene files that contained ten, single copy sequences. A minimum relative length was set where all sequences in a gene file had to be at least 90% the length of the longest sequence in the same file. For example, if the longest sequence in a file is 1000 bp, then the rest of the sequences had to be at least 900 bp. Of 954 genes, 38 genes were excluded for having at least extra sequence and an additional 175 genes were excluded because at least one sequence did not meet the length requirements.

However, there may be an evolutionary reason why at least one of the sequences did not meet the length requirements. The second filter took the 175 gene files that were denied by the first filter one at a time and split the sequences into two groups; the first group contained the gene sequences that were over the 90% length cutoff and the second contained the rest that were shorter than the cutoff length. The first requirement for this filter was that all sequences in the file were longer than 80% of the length of the longest genes in the file. The second requirement was that both representative of a genus were in the same group, either above or below the 90% cutoff. The two outgroup species were not required to be in the same group since they are not from the same genus. An additional 58 genes were added to the working dataset, resulting in 799 genes. A list of the 799 gene names was used to grab the corresponding amino acid files.

## Gene Alignment

Gene alignments were in R using a ClustalW algorithm. The gene fasta files were read into R using the readDNAStringSet() function (package: Biostrings) [64]. Alignments were done using the msaClustalW() function (package: msa) [65] using default parameters and setting maximum number of iterations to 100. The aligned functions were then converted into a readable format using msaConvert() (package: msa)[65] and written into a new fasta file using the write.fasta() function (package: seqinr) [66]

## Distance Matrices

After alignments, model testing was done for each gene through MEGA-CC [67] for both nucleotide and amino acid sequences. In a command line terminal, MEGA-CC is called along with a text file listing the file pathways to all 799 genes and a MAO file detailing the analysis preferences. Default parameters were used for model testing. This resulted in a comma-separated values (CSV) file for each gene. These CSV files were read into R and the model for phylogenetic analysis was extracted based on lowest Bayesian Information Criterion (BIC). A list of the unique models and the number of genes that required each model are given in Table 2 for nucleotide sequences and in Table 3 for amino acid sequences.

Table 2. The number of genes that required each phylogenetic tree model according to model testing and the lowest BIC for nucleotide sequences.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | X. | X..1 | Model.1 | X..2 | X..3 | Model.2 | X..4 | X..5 |
| GTR\_G | 89 | 11.14 | K2 | 1 | 0.13 | T92\_G | 313 | 39.17 |
| GTR\_G\_I | 42 | 5.26 | K2\_G | 94 | 11.76 | T92\_G\_I | 33 | 4.13 |
| HKY\_G | 22 | 2.75 | K2\_G\_I | 1 | 0.13 | TN93\_G | 149 | 18.65 |
| HKY\_G\_I | 1 | 0.13 | K2\_I | 2 | 0.25 | TN93\_G\_I | 52 | 6.51 |

Table 3. The number of genes that required each phylogenetic tree model according to model testing and the lowest BIC for amino acid sequences.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | X. | X..1 | Model.1 | X..2 | X..3 | Model.2 | X..4 | X..5 |
| cpREV | 3 | 0.38 | JTT\_G\_I | 2 | 0.25 | mtREV24\_G\_I | 1 | 0.13 |
| cpREV\_G | 6 | 0.75 | JTT\_I | 1 | 0.13 | rtREV\_G | 3 | 0.38 |
| Dayhoff | 2 | 0.25 | LG | 19 | 2.38 | WAG | 1 | 0.13 |
| Dayhoff\_G | 11 | 1.38 | LG\_G | 563 | 70.46 | WAG\_G | 118 | 14.77 |
| Dayhoff\_I | 2 | 0.25 | LG\_G\_F | 27 | 3.38 | WAG\_G\_F | 2 | 0.25 |
| JTT | 1 | 0.13 | LG\_G\_I | 3 | 0.38 | WAG\_G\_I | 1 | 0.13 |
| JTT\_G | 25 | 3.13 | LG\_I | 5 | 0.63 | WAG\_I | 1 | 0.13 |
| JTT\_G\_F | 1 | 0.13 | mtREV24 | 1 | 0.13 | NA | – | – |

For each unique model, a text file was written listing the gene file pathways of each gene that required that model. In a command line terminal, MEGA-CC [67] was called with a text file listing gene file pathways for a certain model and the corresponding MAO file. This was done for each unique model. These MAO files specified the use of a maximum likelihood statistical method, 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 default.

Distance matrices were extracted from each of the phylogenetic trees and saved as Excel files. These files were read into R and the genetic distances between both of the *Mixta* species and all of the sequences were extracted for further analysis in R. Tables were created in R Markdown using the kable() function (package: knitr) [68] and figures were created using the ggplot() function (package: ggplot2) [69].

# Results and Discussion

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

## Genetic Distance

### Closest Relative

Genetic distance gives a numeric estimate of evolutionary change between species. The species that has the shortest genetic distance to a species of interest is considered to be closest relative. For the following analysis, model testing was done first to estimate the best model for each gene. Next, individual phylogenetic gene trees were created and it was from these that distance matrices were exported. *P. septica* was the closest relative to 66.3% and 66.5% of the *M. calida* and *M. gaviniae* genes, respectively, for the nucleotide sequences (Figure 1). In comparison, only 42.7% and 42.8% of *M. calida* and *M. gaviniae* genes, respectively, are closely related to *P. septica* for the amino acid sequences (Figure 1), but this is still the majority. These results suggest that *P. septica* is the closest relative to *Mixta* followed by *P. agglomerans*. For the amino acid sequences, instead of the *Mixta* genes being more closely related to *P. septica*, they are more closely related to the other species in this study (Figure 1). For example, the percentage of genes related to *P. agglomerans* increased from 13.9% for both of the *Mixta* species to 22.3 and 22.4% to *M. calida* and *M. gaviniae*, respectively. This may be due to a loss of phylogenetic signal when using amino acid sequences as a result of codon degeneracy and having a third of available sites.

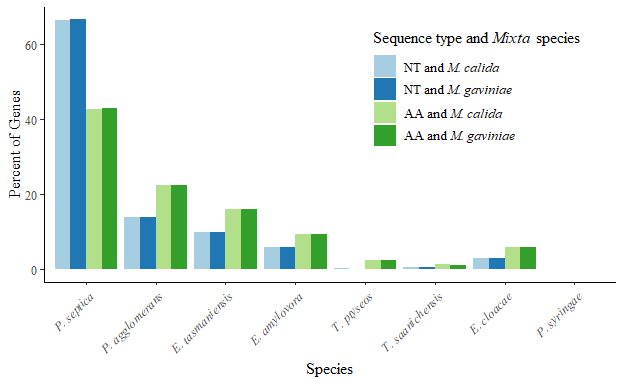


Figure 1. The percentage of genes (out of 799) that were most closely related to any of the eight non-*Mixta* species using genetic distances. The species with the shortest genetic distance from both of the *Mixta* species is considered to be the closest relative. Blue bars are the nucleotide sequences of the genes with light blue bars representing the percentage of *M. calida* genes and dark blue bars representing *M. gaviniae*. Green bars are the corresponding amino acid sequences with light green representing *M. calida* genes and dark blue bars representing *M. gaviniae*. Approximately 66% of *M. calida* and *M. gaviniae* genes were most closely related to *P. septica* when using nucleotide sequences in comparison to only 43% when using amino acid sequences.

Unexpectedly, the number of *Mixta* genes that are more closely related increased from 2.9% to 6.0% when using the amino acid sequences instead of the nucleotide sequences (Figure 1). This, along with the other differences in relatedness for *Mixta* genes between nucleotide and amino acid sequences, suggests that the sequence type used in the analysis will give different results, even when using the same genes.

If genetic distance infers the same closest relative for a *Mixta* gene for both nucleotide and amino acid analyses, then these results may be more reliable. Thus, for genes where the closest relative is an *Erwinia*, *Tatumella*, or *Enterobacter* species both nucleotide and amino acid analyses, they may have resulted from a horizontal gene transfer (HGT) event rather than vertical evolution from a *Pantoea* species. No *Mixta* genes were found to be closely related to *P. syringiae* for either the nucleotide or amino acid sequences (Figure 1). This confirms that *P. syringiae* is a suitable outgroup for this analysis and suggests that any HGT events may be restricted to the order *Enterobacterales*.

### Best Model vs. Best Available Model

In MEGAX [70], the user is restricted to what models can be used when estimating genetic distance; however, the user is not restricted when estimating phylogenetic trees. Model testing in MEGAX tests all of the models that are available for phylogenetic tree analysis. Therefore, the recommended model based on lowest BIC value may not be available for genetic distance analysis. GTR (General Time Reversible) and HKY (Hasegawa-Kishino-Yano) models, as well as the invariant sites (+I) parameter, are not available for genetic distance analysis. In order to determine if the limited options would affect the percentage of each species being the closest relative to the *Mixta* genes, genetic distances were estimated for each gene for the nucleotide sequences in MEGA-CC. The model testing results were filtered to remove the models that are not available for genetic distances in MEGAX and then distance matrices were estimated for each gene using the best available model based on lowest BIC value. As a result, only four of twelve models were available with more than 87% of the genes now using either the Tamura 3-parameter with a Gamma parameter (T92\_G) or the Tamura-Nei with a Gamma parameter (TN93\_G) (Table 4). This is in comparison to the original 58% of genes that required these models when not restricted by availability. Additionally, the best model was not available for approximately 43% of the genes (242 genes). The mean BIC difference between the best model and the best available model was 2.65 with a range of 0 to 122.6 where 0 means that the best available model is the same as the best model. When ignoring cases where the best available model was also the best recommended model, the mean BIC difference was 13.8 with a range of 0.1 to 122.6. These difference may result in a less accurate estimation of genetic distance between species.

Table 4. The number of genes requiring each model when using phylogentic tree analysis vs genetic distance analysis in MEGAX for nucleotide sequences.

|  |  |  |
| --- | --- | --- |
| Model | X..of.genes.requiring.each.model | X..of.genes.requiring.each.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 *Mixta* species were extracted from the distance matrices. *P. septica* was estimated to be the closest relative to 12.6 and 14.4% more genes for *M. calida* and *M. gaviniae*, respectively, when the best available models were used (Figure 2). The next largest difference was for genes that were calculated to be most closely related to *P. agglomerans* (Figure 2). Approximately 5.5 and 7.1% less genes were most closely related to *P. agglomerans* when the best available model was used (Figure 2). Additionally, there was more disparity in the closest relatives between *> calida* and *M. gaviniae* when only the best available model was used (Figure 2). When the best model was used, the two *Mixta* typically had the same closest relative for each gene; however, this was the not the case when model options were limited to those available. Due to differences in evolutionary change estimation between models and the addition or exclusion of additional parameters, genetic distances may be inaccurately calculated when the best model is not used and such results may even suggest that sibling taxa are not as closely related as originally hypothesized.

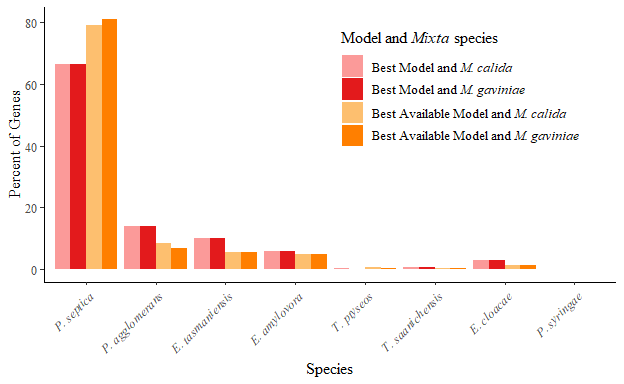


Figure 2. The percentage of genes (out of 799) that were most closely related to any of the eight non-*Mixta* species when using the recommended (best) model from model testing vs the restricted model (best available) options available for genetic distance analysis in MEGAX for nucleotide sequences. The species with the shortest genetic distance from both of the *Mixta* species is considered to be the closest relative. Evolutionary patterns may vary by gene; therefore, it is appropriate to conduct model testing and use the model that explains the most variation. Nevertheless, this model may not be available in some software programs or analysis methods within a software program. Pink bars represent the percentage of *Mixta* genes that are most closely related to any of the non-*Mixta* species when using the recommended model. Orange bars represent the percentage of *Mixta* genes that are most closely related to any of the non-*Mixta* species when using the best available model for genetic distance analysis in MEGAX in which GTR and HKY models and invariant sites parameter are not offered. The lighter shades represent *M. calida* and the darker shades represent *M. gavinaie*.

When the closest relative results were compared between when using the best model vs when using the best available, about 20.0 and 22.3% of *M. calida* and *M. gaviniae* genes were estimated to have a different closest relative. Thus, for 82 *M. calida* genes and 64 *M. gaviniae* genes, the same closest relative was called despite using a different model. This may because the best model and the best available for those genes were similar and explained approximately the same amount of variance.

Each model estimates genetic distance differently. The best model explains most of the variance between sequences and gives a more accurate estimation of genetic distance. Using a model that does not explain most of the variance between sequences will estimate genetic distance less accurately which may result in related taxa (such as those from the same species) having a longer genetic distance between them than in actuality. This may also result in the opposite for taxa that are not closely related (i.e. *Pantoea* spp. and *E. cloacae*). Even though only two of the nucleotide models and one parameter were unavailable for genetic distance analysis in MEGAX, the results for closest relative to *Mixta* genes was substantially impacted. This was particularly the case between the two *Mixta* as one would expect them to have the same closest relative for most of the homologous genes. Therefore, using the best model is recommended. However, this may not be possible when estimating genetic distance using more than one gene that may require different models. Thus, it must be considered that one or more the genes in the concatenation may have a different evolutionary pattern than the others and results should be investigated further to ensure accuracy.

Even though genetic distance estimation in MEGAX did not offer all of the models and parameters as suggested by model testing, all of the models and parameters were available for phylogentic tree analysis and genetic distance can be exported from these trees. However, this has to be done manually by opening the tree file in MEGAX and exporting the distance matrix as either an Excel or CSV file and is therefore not feasible when hundreds or thousands of genetic distance matrices are needed.

### Spatial Patterns of Closest Relatives

To assess whether there were any positional effects of the genes and spatial patterns of the closest relatives, circular plots were generated using the *Mixta* gene IDs for nucleotide sequences (Figure 3) and amino acid sequences (Figure 4). The gene IDs give the approximate location of each gene relative to the *Mixta* genome. The length and colour of a bar represents the species that was the closest relative to the *Mixta* gene at that position. Since the closest relatives were the same both *Mixta* species except for 1 and 2 genes for nucleotide and amino acid sequences, respectively, only the results for *M. calida* are shown. These genes will be discussed later on.

The genes showing greatest identity (lowest distance) to P. septica are distributed across the genome for both M. calida and M. gaviniae (Figures 1 and 2), further suggesting Mixta and Pantoea species have a recent common ancestor. If many genes were horizontally transferred in a single event, then large blocks of genes from a single species would be expected. However, the Erwinia lines (green) are fairly even throughout the genome and the outgroup lines (red and orange) are separated with genes in between. This suggests that Mixta underwent many HGT after speciation from Pantoea.

Positional effects of all of the species and range of the distances were also evaluated (Figures 3 and 4). For most genes, the genetic distance for each of the eight species ranged from 0 to 1.5 with only a few greater than 2. The cluster of genes around ID 425 in both M. calida and M. gaviniae have a very small range and so are highly conserved among all of the species. These genes were annotated as being ribosomal genes. Furthermore, Pantoea species (blue dots) are typically the closest relative to Mixta, followed by Erwinia (green). The Tatumella species (purple and pink) and the outgroup species (red and orange) vary as to which are the furthest relative to Mixta.

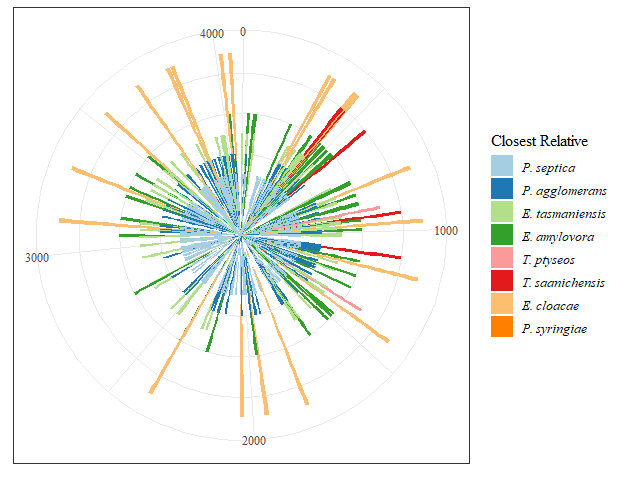


Figure 3.

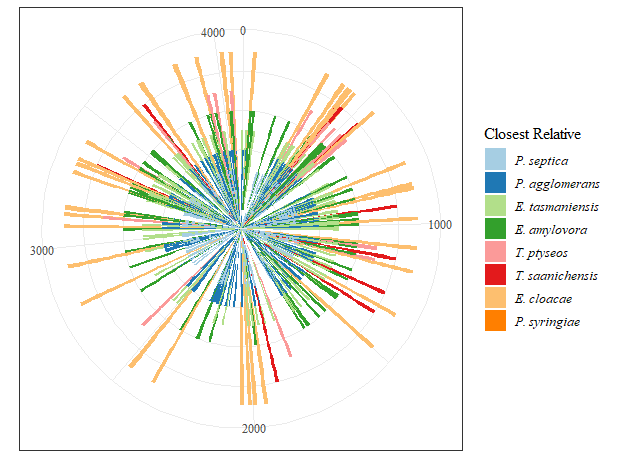


Figure 4.

Nucleotides: ttuB where M. calida is with T. ptyseos and M. gaviniae is with P. septica Amino acids: ttuB where M. calida is with T. saanichensis and M. gaviniae is with P. septica secY where M. calida is with E. tasmaniensis and M. gaviniae is with P. agglomerans # Conclusion

# Literature Cited

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. **Hauben L, Swings J**. *Genus XII. Erwinia*. 2nd ed. New York: Springer; 2005.

4. **Lopez MM, Rosello M, Llop P, Ferrer S, Christen R *et al.*** Erwinia piriflorinigrans sp. nov., a novel pathogen that causes necrosis of pear blossoms. *International Journal of Systematic and Evolutionary Microbiology* 2011;61:561–567.

5. **Matsuura T, Mizuno A, Tsukamoto T, Shimizu Y, Saito N *et al.*** Erwinia uzenensis sp. nov., a novel pathogen that affects European pear trees (Pyrus communis L.). *International Journal of Systematic and Evolutionary Microbiology* 2012;62:1799–1803.

6. **Geider K, Auling G, Du Z, Jakovljevic V, Jock S *et al.*** Erwinia tasmaniensis sp. nov., a non-phytopathogenic bacterium from apple and pear trees. *International Journal of Systematic and Evolutionary Microbiology* 2006;56:2937–2943.

7. **Mergaert J, Hauben L, Cnockaert MC, Swings J**. Reclassification of non-pigmented Erwinia herbicola strains from trees as Erwinia billingiae sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 1999;49:377–383.

8. **Buonaurio R, Moretti C, Silva DP da, Cortese C, Ramos C *et al.*** The olive knot disease as a model to study the role of interspecies bacterial communities in plant disease. *Frontiers in plant science* 2015;6:434.

9. **Campillo T, Luna E, Portier P, Fischer-Le Saux M, Lapitan N *et al.*** Erwinia iniecta sp. nov., isolated from Russian wheat aphid (Diuraphis noxia). *International Journal of Systematic and Evolutionary Microbiology* 2015;65:3625–3633.

10. **Harada H, Oyaizu H, Kosako Y, Ishikawa H**. Erwinia aphidicola, a new species isolated from pea aphid, Acyrthosiphon pisum. *The Journal of General and Applied Microbiology* 1997;43:349–354.

11. **Skrodenyte-Arbaciauskiene V, Radziute S, Stunzenas V, Buda V**. Erwinia typographi sp. nov., isolated from bark beetle (Ips typographus) gut. *International Journal of Systematic and Evolutionary Microbiology* 2012;62:942–948.

12. **Gavini F, Mergaert J, Beji A, Mielcarek C, Izard D *et al.*** Transfer of Enterobacter agglomerans (Beijerinck 1888) Ewing and Fife 1972 to Pantoea gen. nov. as Pantoea agglomerans comb. nov. and Description of Pantoea dispersa sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 1989;39:337–345.

13. **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.

14. **Liu Y, Wang S, Zhang D, Wei S, Zhao S *et al.*** Pantoea beijingensis sp. nov., isolated from the fruiting body of Pleurotus eryngii. *Antonie van Leeuwenhoek* 2013;104:1039–1047.

15. **De Baere T, Verhelst R, Labit C, Verschraegen G, Wauters G *et al.*** Bacteremic Infection with Pantoea ananatis. *Journal of Clinical Microbiology* 2004;42:4393.

16. **Fullerton D, Lwin AA, Lal S**. Pantoea agglomerans liver abscess presenting with a painful thigh. *European journal of gastroenterology & hepatology* 2007;19:433–435.

17. **Kratz A, Greenberg D, Barki Y, Cohen E, Lifshitz M**. Pantoea agglomerans as a cause of septic arthritis after palm tree thorn injury; case report and literature review. *Archives of disease in childhood* 2003;88:542–544.

18. **Lim P, Chen S-L, Tsai C-Y, Pai M-A**. Pantoea peritonitis in a patient receiving chronic ambulatory peritoneal dialysis (Case Report). *Nephrology (Carlton, Vic)* 2006;11:97–99.

19. **Schmid H, Schubert S, Weber C, Bogner J**. Isolation of a Pantoea dispersa-Like Strain from a 71-Year-Old Woman with Acute Myeloid Leukemia and Multiple Myeloma. *Infection* 2003;31:66–67.

20. **Rostenberghe H van, Noraida R, Wan Ibrahim WP, Hasan H, Mohamed Z *et al.*** The clinical picture of neonatal infection with Pantoea species. *Japanese journal of infectious diseases* 2006;59:120–121.

21. **Smits THM, Rezzonico F, Kamber T, Blom J, Goesmann A *et al.*** Metabolic versatility and antibacterial metabolite biosynthesis are distinguishing genomic features of the fire blight antagonist Pantoea vagans C9-1. *PLoS ONE* 2011;6:e22247–e22247.

22. **Kim HJ, Lee JH, Kang BR, Rong X, McSpadden Gardener BB *et al.*** Draft genome sequence of Pantoea ananatis B1-9, a nonpathogenic plant growth-promoting bacterium. *Journal of Bacteriology* 2012;194:729.

23. **Palmer M, Maayer P de, Poulsen M, Steenkamp ET, Zyl E van *et al.*** Draft genome sequences of Pantoea agglomerans and Pantoea vagans isolates associated with termites. *Standards in genomic sciences* 2016;11:23.

24. **Ma Y, Yin Y, Rong C, Chen S, Liu Y *et al.*** Pantoea pleuroti sp. nov., Isolated from the Fruiting Bodies of Pleurotus eryngii. *Current Microbiology* 2016;72:207–212.

25. **Rong C, Ma Y, Wang S, Liu Y, Chen S *et al.*** Pantoea hericii sp. nov., Isolated from the Fruiting Bodies of Hericium erinaceus. *Current Microbiology* 2016;72:738–743.

26. **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.

27. **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.

28. **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.

29. **Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD *et al.*** A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus Neisseria. *Microbiology* 2012;158:1570–1580.

30. **Prasanna A, Mehra S**. Comparative Phylogenomics of Pathogenic and Non-Pathogenic Mycobacterium. *PLoS ONE* 2013;8:e71248.

31. **Angus A, Agapakis C, Fong S, Yerrapragada S, Estrada-de los Santos P *et al.*** Plant-Associated Symbiotic Burkholderia Species Lack Hallmark Strategies Required in Mammalian Pathogenesis. *PloS one* 2014;9:e83779.

32. **De Maayer P, Chan WY, Rubagotti E, Venter SN, Toth IK *et al.*** Analysis of the Pantoea ananatis pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. *BMC Genomics* 2014;15:404.

33. **Fouts D, Matthias M, Adhikarla H, Adler B, Santos L *et al.*** What Makes a Bacterial Species Pathogenic?:Comparative Genomic Analysis of the Genus Leptospira. *PLoS Neglected Tropical Diseases* 2016;10:e0004403.

34. **Konstantinidis KT, Tiedje JM**. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology* 2007;10:504–509.

35. **Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T *et al.*** Re-evaluating prokaryotic species. *Nature Reviews Microbiology* 2005;3:733–739.

36. **Glaeser SP, Kämpfer P**. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology* 2015;38:237–245.

37. **Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C *et al.*** Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 2012;158:1005–1015.

38. **Brady C, Cleenwerck I, Venter S, Vancanneyt M, Swings J *et al.*** Phylogeny and identification of Pantoea species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). *Systematic and Applied Microbiology* 2008;31:447–460.

39. **Fox GE, Wisotzkey JD, Jurtshuk P**. How Close Is Close: 16S rRNA Sequence Identity May Not Be Sufficient To Guarantee Species Identity. *International Journal of Systematic Bacteriology* 1992;42:166–170.

40. **Konstantinidis KT, Tiedje JM**. Towards a genome-based taxonomy for prokaryotes. *Journal of bacteriology* 2005;187:6258–6264.

41. **Staley JT**. The bacterial species dilemma and the genomic-phylogenetic species concept. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2006;361:1899–1909.

42. **Richter M, Rosselló-Móra R**. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106:19126–19131.

43. **Francino M, Santos S, Ochman H**. *Phylogenetic relationships of bacteria with special reference to endosymbionts and enteric species*. New York: Springer; 2006.

44. **Hauben L, Moore ERB, Vauterin L, Steenackers M, Mergaert J *et al.*** Phylogenetic Position of Phytopathogens within the Enterobacteriaceae. *Systematic and Applied Microbiology* 1998;21:384–397.

45. **Naum M, Brown EW, Mason-Gamer RJ**. Is 16S rDNA a Reliable Phylogenetic Marker to Characterize Relationships Below the Family Level in the Enterobacteriaceae? *Journal of Molecular Evolution* 2008;66:630–642.

46. **Spröer C, Mendrock U, Swiderski J, Lang E, Stackebrandt E**. The phylogenetic position of Serratia, Buttiauxella and some other genera of the family Enterobacteriaceae. *International Journal of Systematic Bacteriology* 1999;49 Pt 4:1433–1438.

47. **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.

48. **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.

49. **Koonin EV**. Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics* 2005;39:309–338.

50. **Hedegaard J, Steffensen S, Nørskov-Lauritsen N, Mortensen K, Sperling-Petersen H**. Identification of Enterobacteriaceae by partial sequencing of the gene encoding translation initiation factor 2. *International Journal of Systematic Bacteriology* 1999;49 Pt 4:1531–1538.

51. **Dauga C**. Evolution of the gyrB gene and the molecular phylogeny of Enterobacteriaceae: A model molecule for molecular systematic studies. *International Journal of Systematic and Evolutionary Microbiology* 2002;52:531–547.

52. **Mollet C, Drancourt M, Raoult D**. rpoB sequence analysis as a novel basis for bacterial identification. *Molecular Microbiology* 1997;26:1005–1011.

53. **Paradis S, Boissinot M, Paquette N, Bélanger S, Martel E *et al.*** Phylogeny of the Enterobacteriaceae based on genes encoding elongation factor Tu and F-ATPase -subunit. *International journal of systematic and evolutionary microbiology* 2005;55:2013–2025.

54. **Brady CL, Venter SN, Cleenwerck I, Vandemeulebroecke K, De Vos P *et al.*** Transfer of Pantoea citrea, Pantoea punctata and Pantoea terrea to the genus Tatumella emend. as Tatumella citrea comb. nov., Tatumella punctata comb. nov. and Tatumella terrea comb. nov. and description of Tatumella morbirosei sp. nov. *International Journal of Systematic and Evolutionary Microbiology* 2010;60:484–494.

55. **Brady C, Venter S, Cleenwerck I, Engelbeen K, Vancanneyt M *et al.*** Pantoea vagans sp nov., Pantoea eucalypti sp nov., Pantoea deleyi sp nov and Pantoea anthophila sp nov. *International Journal of Systematic and Evolutionary Microbiology* 2009;59:2339–2345.

56. **Moretti C, Hosni T, Vandemeulebroecke K, Brady C, De Vos P *et al.*** Erwinia oleae sp. nov., isolated from olive knots caused by Pseudomonas savastanoi pv. savastanoi. *International Journal of Systematic and Evolutionary Microbiology* 2011;61:2745–2752.

57. **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.

58. **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.

59. **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.

60. **Octavia S, Lan R**. The Family Enterobacteriaceae. *The Prokaryotes: Gammaproteobacteria* 2014;225–286.

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

62. **Contreras-Moreira B, Vinuesa P**. GET\_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. *Applied and Environmental Microbiology* 2013;79:7696–7701.

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

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

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

66. **Charif D, Lobry JR**. Seqin{R} 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.

67. **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.

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

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

70. **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.

# Appendix