An In-depth Analysis of a Multilocus Phylogeny Identifies $leuS$ As a Reliable Phylogenetic Marker for the Genus *Pantoea*

James T. Tambong, Renlin Xu, Cynthia-Anne Kaneza and Jean-Claude Nshogozabahizi

Laboratory of Bacteriology, Agriculture and Agri-Food Canada, Ottawa, Ontario Canada.

**ABSTRACT:** Partial sequences of six core genes ($fusA$, $gyrB$, $leoS$, $pyrG$, $rplB$, and $rpoB$) of 37 strains of *Pantoea* species were analyzed in order to obtain a comprehensive view regarding the phylogenetic relationships within the *Pantoea* genus and compare tree topologies to identify gene(s) for reliable species and subspecies differentiation. All genes used in this study were effective at species-level delineation, but the internal nodes represented conflicting common ancestors in $fusA$- and $pyrG$-based phylogenies. Concatenated gene phylogeny gave the expected DNA relatedness, underscoring the significance of a multilocus sequence analysis. Pairwise comparison of topological distances and percent similarities indicated a significant differential influence of individual genes on the concatenated tree topology. $leoS$- and $fusA$-inferred phylogenies exhibited, respectively, the lowest (4) and highest (52) topological distances to the concatenated tree. These correlated well with high (96.3%) and low (64.4%) percent similarities of $leoS$- and $fusA$-inferred tree topologies to the concatenated tree, respectively. We conclude that the concatenated tree topology is strongly influenced by the gene with the highest number of polymorphic and non-synonymous sites in the absence of significant recombination events.

**KEYWORDS:** *Pantoea stewartii*, multilocus, phylogeny, $leoS$, topology

**CITATION:** Tambong et al. An In-depth Analysis of a Multilocus Phylogeny Identifies $leuS$ As a Reliable Phylogenetic Marker for the Genus Pantoea. Evolutionary Bioinformatics 2014:10 115–125 doi: 10.4137/EBi.S15738.

**RECEIVED:** March 31, 2014. **RESUBMITTED:** May 15, 2014. **ACCEPTED FOR PUBLICATION:** May 20, 2014.

**ACADEMIC EDITOR:** Jike Cui, Associate Editor

**TYPE:** Original Research

**FUNDING:** This research was funded by Agriculture and Agri-Food Canada through project #1800 and by the Defence R&D Canada-Centre for Security Science project # CRT1 09–462RD (leader, Dr. André Levesque).

**COMPETING INTERESTS:** Authors disclose no potential conflicts of interest.

**COPYRIGHT:** © the authors, publisher and licensee Libertas Academica Limited. This is an open-access article distributed under the terms of the Creative Commons CC-BY-NC 3.0 License.

**CORRESPONDENCE:** james.tambong@agr.gc.ca

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants.

**Introduction**

Species of the genus *Pantoea* are nonencapsulated, nonspore-forming gram negative bacteria of the *Enterobacteriaceae* family. They are widely distributed in nature and have been isolated from plants, animals, and humans. Plant-associated pantoeas are either epiphytes or pathogens and constitute the family. *Pantoea* species based on 16S rRNA showed that *P. agglomerans*, *Pantoea ananatis*, and *Pantoea stewartii* were closely related but have limitations at species- and subspecies-level identifications. An updated phylogeny using multiple loci (at least six) could provide some clarity to the DNA relatedness and the degree of uniqueness among current *Pantoea* species.

Multilocus sequence analysis (MLSA), with careful selection of representative sequences, provides an alternative approach to define species and improve pathogen identification. MLSA has been shown to be a powerful molecular method for microbial population genetics studies, delineation is the biochemical heterogeneity of *Pantoea agglomerans* and related strains and species that make routine identification difficult. Phylogenetic inferences and identification of *Pantoea* species based on 16S rRNA showed that *P. agglomerans*, *Pantoea ananatis*, and *Pantoea stewartii* were closely related but have limitations at species- and subspecies-level identifications. An updated phylogeny using multiple loci (at least six) could provide some clarity to the DNA relatedness and the degree of uniqueness among current *Pantoea* species.

Biochemical or nutritional characteristics used to differentiate *Pantoea* species/strains are inadequate and show considerable overlapping among species and subspecies. This might lead to potential errors in identification. An example
of species, and assignment of strains to defined bacterial species. The concept of MLSA involves systematic selection of several housekeeping/protein-coding genes to represent the chromosome. Inference of phylogeny on concatenated multiple protein-coding genes could allow for a clear delineation of species into sequence clusters that can provide valid indices for defining species, even though the borders can be fuzzy in highly recombinogenic bacterial species. Young and Park used three genes (\(\text{atpD, carA, and recA}\)) to study the relationships of plant pathogenic enterobacteria, while Delétoile et al. used six genes (\(\text{fusA, gyrB, leuS, pyrG, rplB, and rpoB}\)) to delineate the four core Pantoea species (\(\text{P. agglomerans, P. ananatis, P. stewartii, and P. dispersa}\)) with a focus on typing \(\text{P. agglomerans}\) because of its opportunistic-pathogen status in humans. Brady et al. used four housekeeping genes to assign new strains to the genus \(\text{Pantoea}\). MLSA derived from few loci could be less informative, with potential impact on accurate phylogeny inference and reliable molecular identification. None of the previous studies on \(\text{Pantoea}\) species reported the discriminatory potential of each gene or compared the tree topologies obtained from the different single genes with that of concatenated tree. This is required for a better understanding of the influence of the different loci on the concatenated tree topology and to identify a reliable phylogenetic marker gene with similar topology. The objectives of this study were to define a comprehensive view regarding the phylogenetic relationships within the \(\text{Pantoea}\) genus using six core housekeeping genes, to determine percent similarities of tree topologies of each individual genes compared with that of the concatenated tree, and to identify a gene with tree topology that is highly similar (at least 95.0%) to that of the concatenated tree for reliable species and subspecies differentiation. We hypothesized that a gene with the highest number of polymorphic sites and proportion of non-synonymous sites could have a strong influence on the concatenated tree topology. The robustness of the tree topology of the selected gene(s) was compared against whole genome-based tree of 15 species of \(\text{Pantoea, Erwinia, and Pectobacterium}\).

Materials and Methods

Bacterial strains and DNA extraction. Thirty-seven representative, reference, and/or type strains of \(\text{Pantoea}\) and \(\text{Tatumella}\) species were sequenced for multilocus phylogeny (Supplementary Table S1). Type strains of the genus \(\text{Pantoea}\) were obtained from the Belgian Coordinated Collections (BCCM/LMG; Ghent, Belgium). Strains were cultured in Luria-Bertani (LB) or nutrient broth (NB) and incubated at 28°C. Stock bacterial cultures were maintained on the same media supplemented with 25% w/v glycerol at −80 °C. Genomic DNA of the bacterial strains was purified using the Wizard SV Genomic DNA purification system Kit (Promega) following the manufacturer’s instructions.

Selection of housekeeping genes. The selected genes (\(\text{leuS, rpoB, gyrB, fusA, pyrG, and rplB}\)) were previously reported by Delétoile et al. in a study of \(\text{P. agglomerans}\) strains because of its opportunistic pathogen status in humans. Figure 1 shows the genomic location of the genes used in this study based on \(\text{P. ananatis}\) LMG 5342 (HE617160.1). These are single copy genes in most bacterial genera with important functional roles. \(\text{leuS}\) encodes leucyl-tRNA synthetase involved in translation. \(\text{gyrB}\) is implicated in glutamine hydrolysis through CTP synthase; \(\text{rpoB}\) encodes for the \(\beta\) subunit of RNA polymerase; \(\text{gyrB}\) is the structural gene for the DNA gyrase \(\beta\) subunit; and \(\text{fusA}\) encodes the protein synthesis elongation factor-G. The primers reported by Delétoile et al. were designed from conserved regions.

PCR amplification and nucleotide sequencing of core housekeeping genes. Partial sequences of six protein-coding genes (\(\text{leuS, rpoB, gyrB, fusA, pyrG, and rpoB}\)) were generated for all the reference/type strains not available in public databases. The primer pairs and conditions applied in this study are as described by Delétoile et al. with the exception of \(\text{rpoB}\). The primer pairs and conditions applied in this study with an initial denaturation of 95 °C for 3 minutes, followed by 40 cycles of 95 °C for 45 seconds, 65 °C for 45 seconds, 72 °C for 90 seconds, and a final extension at 72 °C for 10 minutes. Primers were synthesized by Invitrogen Inc. (Carlsbad, CA, USA), and PCR amplifications were performed in a thermal cycler (Bioterrex) using 10 ng of bacterial DNA (1 µL), 1 µL of 10× PCR buffer, 0.75 µL of 2 mM deoxynucleoside triphosphates (dNTPs), 0.08 µL of each primer (20 µM), 0.1 µL of Titanium Taq DNA polymerase (5 U/µL), BD/Biosciences/Clontech, and 6.99 mL of MilliQ water. Sequencing was done as previously described using ABI BigDye Terminator chemistry v3.1 (Applied Biosystems) and run on an ABI 3130xl automated sequencer (Applied Biosystems/Hitachi). Sequences were edited using SeqMan (DNASTAR).

Analysis of sequence data. The sequences were imported into MEGA5 for final editing taking into account corresponding amino acids and checked for correct in-frame reading using RevTrans version1.4. Sequences were aligned using a Linux OS version of MUSCLE, and the alignments were trimmed so that sequences of each given gene were of the same length.

Summary statistics for the sequences were computed using DNA SP version 5.1. G+C content, number of haplotypes (\(\delta\)), number of polymorphic (segregating) sites (\(S\)), synonymous sites (\(\pi_s\)), and non-synonymous sites (\(\pi_{ns}\)) with Jukes–Cantor correction statistics were calculated. The ratio of non-synonymous to synonymous substitutions (\(d_N/d_S\)) was calculated with Jukes–Cantor correction. Phi (\(\Phi_\theta\)) was computed for evidence of recombination at 95% confidence interval as previously described and implemented in SplitsTree4 software. The phi test is based on the compatibility of informative sites providing a P value, which when significant (\(P < 0.05\)) would indicate that events of recombination are highly likely.
Geneious 5.6.4 (http://www.geneious.com/) was used to concatenate corresponding gene sequences. Prior to concatenation, the test for concordance was performed using the CAMD.global function. Distance matrices of the six core housekeeping genes were computed in MEGA5 and analyzed for congruency using Congruence Among Distance Matrices (CADM), a statistical test for estimating the level of CADM as implemented in R-statistics. The null hypothesis of the CADM test ($H_0$) is the complete incongruence of two or more matrices, while the alternative hypothesis ($H_1$) indicates partial or complete congruency.

The six genes were compared as previously reported to determine the most discriminating gene or genes. Briefly, a matrix of the phylogenetic distances between all the 37 strains was constructed for each single gene and pairwise least-square tendency lines were generated and compared. The discriminatory potential of the concatenated sequences was, also, compared with those of single genes.

**Phylogenetic analyses of single gene and concatenated sequences using maximum likelihood.** Aligned individual gene and concatenated sequences were used to infer maximum-likelihood (ML) phylogenies using PhyML version 3.0 and MEGA5 with the general time reversible (GTR) substitution model, selected on the basis of the Akaike information criterion implemented in jMODELTEST version 2.1.1. ML was executed with the Subtree pruning and regrafting (SPR) and nearest-neighbor interchange (NNI) tree improvement algorithms as implemented in PhyML and MEGA5 with 1000 bootstrap replicates for single gene phylogenies.

**Comparison of tree topologies.** Pairwise comparisons of the tree topologies derived from single gene and concatenated sequences were performed using the dist.topo ($x, y$, method = "PH85") command as implemented in R-statistics. This function is based on Penny and Hendy (1985) rates (PH85) and describes the topological distance between two trees as twice the number of internal branches defining different bipartitions of the tips with a single numeric value as output. In addition, the Compare2Trees software was also used to compare phylogenies obtained from individual genes and concatenated sequences. This algorithm pairs up each branch in one phylogeny with a matching branch in the second phylogeny, and finds the optimum one-to-one map between branches in the two trees in terms of a topological score. This software identifies those parts of the trees that differ both in terms of topology and branch length.

**Genome- and leuS-based phylogenies.** The degree of similarity/dissimilarity of leuS phylogeny to that based on whole genome was determined using 15 publicly available genomes of Pantoea and closely related genera (Erwinia and Pectobacterium). Complete leuS gene sequences were extracted from the respective genomes and used to infer phylogenies using the composition vector (CV) approach. The CV approach is based on the frequency of appearance of overlapping oligonucleotides of length $K$ in the genome or gene. CVTree (http://tlife.fudan.edu.cn/cvtr/) was used to construct genome- and leuS-based phylogenies with a $K$-value of 6 and Sulfolobus acidocaldarius Ron12/I (NC_020247) as outgroup. $K = 6$ was selected as the best based on our preliminary evaluation of different values. leuS- and genome-based tree topologies were compared as indicated above.

**Sequence accession numbers.** Two hundred and twenty-two sequences generated in this study were deposited in the GenBank database with accession numbers KF482534–KF482570, KF482571–KF482607, KF482608–KF482644, KF482645–KF482681, KF482682–KF482718, and KF482719–KF482755 for fuaA, gyrB, leuS, pyrG, rtpB, and rpoB, respectively.

**Results.** Summary statistics of sequenced genes. Table 1 summarizes the nucleotide sequence data for the six genes. Sequence length (bp) ranged from 306 (pyrG) to 722 (gyrB).
with comparable G + C contents (51.5–56.9%). The number of segregating (polymorphic) sites as a percentage of sequence length was significantly highest in leuS (45.57%) followed by gyrB (36.84%) and pyrG (31.7%), while the rpoB gene exhibited the lowest (18.62%). Nucleotide diversity of non-synonymous sites was highest for leuS (data not shown). All the ratios of dN/dS were <1 indicating that the six genes were subject to purifying selection. The phi test suggested that the gene fusA had significant high recombination events while pyrG and rpoB exhibited moderate recombination events (Table 1). The genes, gyrB, leuS, and rpoB, did not show significant events of recombination.

**Discriminatory potential of the different genes.** The six genes were compared in order to determine the most discriminating gene or genes. A matrix of the phylogenetic distances between all the 37 strains was constructed for each single gene and the distances (3996 values) of pairs of strains were plotted.19 Plots were generated between leuS (exhibiting the highest number of polymorphic sites) or rpoB (lowest polymorphic sites) and the other five core genes. The ratio between the leuS slope and the slopes of the other genes was calculated as a measure of the discriminating potential of each gene19: leuS/rpoB (3.75 times); leuS/pyrG (1.88 times); leuS/rpoB (1.88 times); leuS/fusA (1.55 times); leuS/gyrB (1.25 times). The slopes of least square tendency lines indicate that Pantoea strains have the most distinguished genetic distance for one another in gene leuS, and to suggest that leuS is the most discriminating gene, which is followed by gyrB. Similarly, the matrix constructed for the concatenated nucleotide sequences of the six protein-coding genes (2997nt) was compared in a pairwise manner to assess the correlation and the relative discriminatory power (Fig. 2). The matrix distance of the concatenated sequence least correlated with rpoB, while maximum correlations were obtained with leuS and gyrB (Fig. 2). leuS and gyrB were 50 and 20% more discriminatory than the concatenated gene sequences, while fusA was comparable to that of the concatenated sequence. rpoB, pyrG, and rpoB were less discriminatory (Fig. 2) than the concatenated sequences.

### Table 1. Summary statistics for the six house-keeping gene and concatenated sequences used in this study.

| LOCUS  | SEQUENCE LENGTH (bp) | GC CONTENT (%) | S | H | dN/dS | ω (P) |
|--------|----------------------|----------------|---|---|-------|-------|
| fusA   | 588                  | 51.5           | 19 | 163 | 0.12038 | 0.00002*** |
| gyrB   | 722                  | 52.5           | 22 | 266 | 0.03531 | 0.09800  |
| leuS   | 643                  | 56.9           | 21 | 293 | 0.09654 | 0.98600  |
| pyrG   | 306                  | 52.7           | 18 | 97  | 0.01515 | 0.03900*  |
| rpoB   | 409                  | 52.9           | 20 | 114 | 0.05951 | 0.02600*  |
| rplB   | 333                  | 55.7           | 17 | 62  | 0.04008 | 0.67000  |
| Concatenated | 3001 | 53.7           | 23 | 994 | 0.17937 | 0.00000*** |

**Notes:** ω, phi test for evidence of recombination. P values < 0.05 (* or ***) indicate significant recombination events. **Abbreviations:** h, number of haplotypes; S, number of polymorphic (segregating) sites; dN/dS, ratio of non-synonymous to synonymous substitutions.

**Single gene and concatenated phylogenies.** Single gene (Fig. 3A–F) and concatenated (Fig. 4) phylogenetic trees from leuS, gyrB, rpoB, fusA, pyrG, and rplB were computed using ML to determine the phylogenetic relationships between Pantoea species. All single gene trees showed the main lineages (Fig. 3A–F) but some of the internal nodes represented conflicting common ancestors for some species in fusA- and pyrG-based phylogenies. The *P. stewartii* subgroup regrouped the two described subspecies (subspecies stewartii and indoligenes) while the *P. ananatis* subgroup has all the strains of *P. ananatis* and *Pantoea allii*. The *P. agglomerans* subgroup included strains from *P. agglomerans*, *Pantoea vagans*, *Pantoea deleyi*, and *Pantoea eucalypti*. However, there were significant differences among genes with respect to the phylogenetic associations of *P. dispersa*, *Pantoea septica*, and *Pantoea cypripedii*. This discrepancy was most evident in fusA phylogeny with these three species sharing a direct ancestor with *P. stewartii* subgroup (Fig. 3D).

The expected phylogenetic associations were observed in the tree derived from concatenated gene sequences (Fig. 4). The tree inferred with concatenated sequences confirmed the phylogenetic affiliations of the core *Pantoea* subgroups (*P. stewartii*, *P. ananatis*, *P. agglomerans*, and *P. dispersa*) and resolved the conflicting common ancestors represented by some internal nodes especially for *P. septica*, *P. dispersa*, *P. cypripedii*, and *P. deleyi*. The *P. stewartii* subgroup did not include any of the newly described species while the *P. ananatis* subgroup now includes strains of *P. allii*. Strains of four new species (*P. anthophila*, *P. eucalypti*, *P. vagans*, and *P. deleyi*) could be associated with the *P. agglomerans* subgroup. *P. cypripedii*, a species recently transferred from *Pectobacterium* to *Pantoea* based on DNA relatedness, is affiliated with the *P. dispersa* subgroup. A new *P. septica* subgroup consisting of *P. septica*, *P. calida*, and *P. gaviniae* was identified based on genetic distance from the closest subgroups. Both strains of *P. calida* and *P. gaviniae* were isolated from powder infant milk and are highly (96.1%) similar genetically, but distantly similar to *P. septica*, a strain isolated from...
human stool. The inferred concatenated tree topology was highly congruent to \textit{leu}S-derived tree.

**Comparison of tree topologies.** Table 2 shows the topological distances between the different phylogenetic trees generated based on PH85 rates and percent similarity of tree topologies. The highest PH85 topological distances (48 to 52) were obtained between \textit{fusA}-inferred phylogeny and the other genes, an indication of high differences in tree topologies. \textit{leu}S-derived tree showed the lowest PH85 topological distance of 20 with \textit{gyrB}-inferred tree, suggesting high similarity. This is consistent with the highest percent similarity (88.8%) computed between \textit{leu}S and \textit{gyrB} tree topologies (Table 2). \textit{leu}S-derived tree topology was only 64.4% similar to that of \textit{fusA}. The lowest percent similarity (59.1%) was recorded between \textit{rpoB} and \textit{pyrG} phylogenies (Table 2). Tree topology (Fig. 4) derived from concatenated sequences was 96.7, 89.7, or 70.8% similar to \textit{leu}S, \textit{gyrB}, or \textit{rpoB} phylogeny, respectively. High similarity of \textit{leu}S tree topology correlates well with the low topological distance based on PH85 rates (Table 2). The \textit{fusA} phylogeny showed the highest topological distance of 52 when compared to tree topology derived from concatenated sequences.

The tree edges within each single gene tree that differ from the corresponding edges on the concatenated tree are identified in Figure 3A–F. Three edges in \textit{leu}S-derived trees differed from their corresponding tree edges on the concatenated tree with similarities of only 25, 62.5, and 75.0% (Fig. 3A), while \textit{fusA} and \textit{pyrG}, respectively, each showed seven differences with similarities ranging from 8.8 to 73.3% (Fig. 3D) and 33.3 to 86.7% (Fig. 3E), respectively. \textit{rpoB} tree topology had six edges (Fig. 3C) while \textit{gyrB} showed four edges (37.5, 55.6, 62.5, or 75.0%; Figure 3B) that differed from the corresponding edges on the concatenated tree. \textit{rpoB} showed five edges (33.0, 40.0, 50.0, 50.0, and 83.0%; Figure 3F) that are not identical to those of the corresponding edges on the concatenated tree.

**Genome- and \textit{leu}S-based phylogenies.** The tree topologies based on \textit{leu}S and that derived from 15 complete genomes of \textit{Pantoea}, \textit{Erwinia}, and \textit{Pectobacterium} showed highly similar clustering patterns (Fig. 5). Three clusters representing the different genera were apparent (Fig. 5). The \textit{leu}S-derived tree topology (Fig. 5A) was 94.0% identical to that of genome-based phylogeny (Fig. 5B). In the \textit{leu}S tree topology, the edge bifurcating to \textit{Pectobacterium carotovorum} PC1 and \textit{P. carotovorum} PCC21 seems to differ from the corresponding edge on the genome-based tree. On the genome-based tree topology, these two strains branched independently (Fig. 5B). A low similarity (73.8%) was observed between \textit{leu}S-derived tree topology and that of a tree derived from the corresponding 40 complete genomes of 15 distantly related genera of the \textit{Enterobacteriaceae} (data not shown).

**Discussion**

This study describes an in-depth analysis of a multilocus phylogeny of \textit{Pantoea} species to determine the DNA relatedness of the \textit{Pantoea} species using six genes and to compare single gene tree topologies of the different protein-coding genes to that from concatenated data. This allowed for the identification of \textit{leu}S as a reliable phylogenetic marker for the genus \textit{Pantoea}.

The MLSA approach has been used in several studies to reliably infer molecular relatedness among bacteria species.\cite{cole2014,cole2014119} According to Cole et al.,\cite{cole2014} MLSA is "intermediary" of the 16S rRNA and the genome-based approach for phylogeny. It has been used to improve resolution at lower taxonomic ranks such as species and subspecies or to evaluate and eliminate phylogenetic inconsistencies due to lateral gene transfer.\cite{cole2014119} It is considered as a novel standard in microbial molecular systematics for species delineation.\cite{cole2014} Several studies have proposed MLSA as an alternative to DNA–DNA hybridization\cite{cole2014119,cole2014119} but a consensus has not been reached due to some inherent problems.\cite{cole2014} Whole genome-based phylogenies have potential to reveal more on similarities/dissimilarity between \textit{Pantoea} species; however, at this time only six genomes (5 \textit{P. ananatis} and 1 \textit{P. vagani}) are available. In addition, computational
Figure 3. Single gene phylogenetic trees inferred by maximum-likelihood (ML) of 37 strains of Pantoea and Tatumella using the general time reversible substitution model. A, leuS; B, gyrB; C, rpoB; D, fusA; E, pyrG; and F, rplB. ML trees were reconstructed for each gene sequences with 1000 bootstrap replicates. Bootstrap values higher than 50 are shown in black on nodes. Tatumella spp. were used as outgroup. Taxa and tree edges, and numbers in color denote differences, and percent edge similarity to that of the concatenated tree.
and bioinformatic challenges need to be resolved before a comparative analysis can be routinely done on whole genomes of many bacteria. This indicates that MLSA will remain a useful tool in bacterial phylogeny and systematics.

It is clear that systematic biases are reduced by combining the phylogenetic signal from all loci in a concatenated dataset (multi-locus analyses). This has been the main objective of previous applications of MLSA in delineating *Pantoea* species, even though on a limited number of species.\(^1\,^2\,^3\) This study did not only achieve this goal but also did a comparison of single gene tree topologies to that of concatenated data-set as an indirect assessment of relative contribution of a gene to the final tree topology. In the absence of recombination events, we hypothesized that a single gene with the highest number of polymorphic sites could have a strong influence on the concatenated tree topology. This analysis allowed for the identification of *leu*S as a reliable phylogenetic marker for the genus *Pantoea*. A topological difference of only 3.3% was observed between *leu*S-derived and the concatenated tree topologies while those of the other genes ranged from 10.3% (*gyrB*) to 38.1% (*pyrG*). This suggests a differential influence of the loci studied on the outcome of the concatenated tree topology.

Also, *leu*S-based tree topologies compared favorably (94%) with whole genome-based tree of all publicly available *Pantoea* (5 genomes) and 10 closely related genera (*Erwinia* and *Pectobacterium*). However, relatively low similarity suggests that the robustness and reliability of the *leu*S phylogeny shows potential limitations with data consisting of different genera. This hypothesis was tested using 40 whole genomes of 15 enterobacterial genera and their corresponding complete *leu*S resulting to a lower (73.8%) topological similarity (data not shown).

Our results illustrate the importance of using multiple genes to buffer phylogenetic conflicting signals and to inform better on the validity of assigning strains to species and subspecies. Conflicting signals could be due to horizontal
Three of the six genes used in this study showed high recombination events, with fusA locus showing the highest probability. This could explain why the fusA-derived phylogeny was significantly different from the phylogenies based on topological distances or percent similarities. The fusA phylogeny revealed a potential genetic exchange between the strains of *P. stewartii*, *P. septica*, *P. dispersa* and *P. cypripedii*. Genetic exchange between bacteria are significant and common evolutionary processes mediated by conjugation, transduction, and transformation. These processes promote the acquisition of novel genetic material and often lead to the emergence of new phenotypes/genotypes.

Table 2. Pairwise topological distances based on Penny and Hendy’s rate (above diagonal) and percent similarity (below diagonal) of tree topologies using Compare2Trees.

|       | leuS   | gyrB   | fusA   | pyrG   | rpoB   | rlpB   | CONCATENATED |
|-------|--------|--------|--------|--------|--------|--------|-------------|
| leuS  | 0 (100%) | 20     | 52     | 44     | 32     | 40     | 4           |
| gyrB  | 88.80%  | 0 (100%) | 50     | 42     | 36     | 46     | 12          |
| fusA  | 64.40%  | 66.20%  | 0 (100%) | 48     | 52     | 52     | 52          |
| pyrG  | 61.90%  | 65.70%  | 80.10%  | 0 (100%) | 44     | 42     | 42          |
| rpoB  | 71.40%  | 72.20%  | 65.40%  | 59.10%  | 0 (100%) | 34     | 30          |
| rlpB  | 66.10%  | 66.20%  | 75.00%  | 69.50%  | 73.20%  | 0 (100%) | 36          |
| concatenated | 96.70%  | 89.70%  | 63.60%  | 61.90%  | 70.80%  | 66.70%  | 0 (100%)   |

Notes: PH85 topological distance is defined as twice the number of internal branches defining different bipartitions of the tips with a single numeric value. Percent similarity of tree topologies obtained using Compare2Trees.

---

Figure 5. Similarities between leuS-inferred (A) and whole genome-based (B) tree topologies of 15 publicly available *Pantoea*, *Erwinia*, and *Pectobacterium* genomes. Taxa and tree edges, and numbers in color denote differences, and percent edge similarity of leuS-based tree topology to that from whole genomes. *Sulfolobus acidophilus* Ron121 (NC_020247) was used as outgroup.
but little chance of recombination. This could be as a result of errors in single base substitutions during DNA replication or repair. Further investigation is required to elucidate the mechanisms involved.

Conclusion
An analysis was performed on a multilocus phylogeny of the genus Pantoea to determine the phylogenetic relationships of Pantoea species. It is clear that by comparing the phylogenetic signal from all loci in a concatenated dataset (multi-locus analyses) systematic biases are reduced. An in-depth comparison of tree topologies derived from single individual genes suggests that the gene *leu* has a strong influence on the concatenated tree topology, while *pyrG* and *fusA* had lesser effects. *leu* satisfies the requirements for a suitable phylogenetic marker as previously reported. It is a single-copy housekeeping, protein-coding gene in *Pantoea*, it is widely distributed and has a high discriminative power of all the genes studied. Also, *leu* has the lowest level of recombination, suggesting that it is not prone to lateral transfer. Its sequence length and high number of polymorphic sites allows sufficient phylogenetic resolution and the development of a reliable molecular diagnostic assay for genotyping strains of *P. stewartii*. We are currently validating *leu*-based assays for reliable detection of six plant pathogenic *Pantoea* species in a single reaction.

Acknowledgments
The authors thank Shea Miller and John Bissett for reviewing the manuscript.

Author Contributions
JTT conceived and designed the project/experiment, analyzed data, and wrote the manuscript. RX performed all the PCR amplifications and gel electrophoresis. CAK sequenced the core housekeeping genes. JCN edited the manuscript. All authors reviewed and approved of the final manuscript.

REFERENCES
1. Delétoile A, Decret D, Courant S, et al. Phylogeny and identification of *Pantoea* species and typing of *Pantoea agglomerans* strains by multilocus gene sequencing. *J Clin Microbiol*. 2009;47(2):300–10.
2. Brady CL, Cleenwerck I, Venter SN, Engelbeen K, De Vos P, Coutinho TA. Emended description of the genus *Pantoea*, description of four species from human clinical samples, *Pantoea opticosa* sp. nov., *Pantoea enricisa* sp. nov., *Pantoea brenerri* sp. nov. and *Pantoea complicis* sp. nov., and transfer of *Pectobacterium* sp. *P. opticosa* (Hori 1913) Brener et al. 1973 emend. Hashen et al. 1998 to the genus as *Pantoea opticosa* comb. nov. *Int J Syst Evol Microbiol*. 2010;60(pt 10):2430–40.
3. Naum M, Brown EW, Mason-Gamer RJ. Is 16s rDNA a reliable phylogenetic marker to characterize relationships below the family level in the *enterobacteriaceae*? *J Mol Evol*. 2008;66(6):630–42.
4. Young JM, Park DC. Relationships of plant pathogenic enterobacteriaceae based on partial atpD, *car*A, and recA as individual and concatenated nucleotide and peptide sequences. *Syst Appl Microbiol*. 2007;30(5):343–54.
5. Bishop CJ, Aanesen DM, Jordan GE, Kilian M, Hanage WP, Spratt BG. Assigning strains to bacterial species via the internet. * BMC Biol*. 2009;7:3.
Supplementary Data

Table S1. Bacterial strains used in this study.

| BACTERIAL SPECIES/STRAIN | SOURCE/HOST        | COUNTRY         |
|--------------------------|--------------------|-----------------|
| **Pantoea stewartii subsp. stewartii:** |                    |                 |
| LMG 2715<sup>T</sup>    | Zea mays           | USA             |
| DC162                   | Z. mays            | USA             |
| DOAB 022                | Z. mays            | Canada          |
| DC116                   | Z. mays            | USA             |
| DC146                   | Z. mays            | USA             |
| SS104                   | Z. mays            | USA             |
| DC145                   | Z. mays            | USA             |
| SW13                    | Chaetocnema pulicaria | USA         |
| DC283                   | Z. mays            | USA             |
| DC147                   | Z. mays            | USA             |
| **Pantoea stewartii subsp. indoligenes:** |                    |                 |
| LMG 2632<sup>T</sup>    | Setaria italica    | India           |
| LMG 2630                | Cyamopsis tetragonolobus | Unknown       |
| LMG 2631                | Pennisetum americanum | India       |
| DOAB 213               | Unknown            | Unknown         |
| **Pantoea ananatis:**   |                    |                 |
| LMG 2665<sup>T</sup>    | Ananas comosus     | Brazil          |
| LMG 2667                | Ananas comosus     | Hawaii, USA     |
| LMG 2668                | Ananas comosus     | Hawaii, USA     |
| LMG 2675                | Puccinia graminis f. sp. tritici | Hungary |
| **Pantoea agglomerans:** |                    |                 |
| LMG 1286<sup>T</sup>    | knee laceration    | Zimbabwe        |
| LMG 2102               | Deer               | USA             |
| **Pantoea anthophila:** |                    |                 |
| LMG 2558<sup>T</sup>    | Impatiens balsamina | India       |
| LMG 2560               | Tagetes erecta     | United Kingdom  |
| **Pantoea eucalypti:**  |                    |                 |
| LMG 24197<sup>T</sup>   | Eucalyptus sp.     | Uruguay         |
| LMG 24198               | Eucalyptus sp.     | Uruguay         |
| **Pantoea allii:**      |                    |                 |
| LMG24248<sup>T</sup>    | Allium cepa        | South Africa    |
| Bacterial species/strain | Source/host        | Country         |
| **Pantoea vagans:**     |                    |                 |
| LMG24199<sup>T</sup>    | Eucalyptus sp.     | Uganda          |
| **Pantoea calida:**     |                    |                 |
| LMG 25383<sup>T</sup>   | Powdered infant formula | Switzerland |
| **Pantoea gaviniae:**   |                    |                 |
| LMG 25382<sup>T</sup>   | Powdered infant formula | Switzerland |
| **Pantoea septica:**    |                    |                 |
| LMG 5345<sup>T</sup>    | Human, stool       | USA             |
| **Pantoea dispersa:**   |                    |                 |
| LMG2603<sup>T</sup>     | Soil               | Japan           |

(Continued)
Table S1. (Continued)

| BACTERIAL SPECIES/STRAIN | SOURCE/HOST | COUNTRY |
|--------------------------|-------------|---------|
| *Pantoea deleyi*:        |             |         |
| LMG 24200<sup>T</sup>    | *Eucalyptus* sp. | Uganda |
| *Tatumella punctata*     |             |         |
| LMG 23562<sup>T</sup>    | *Citrus sinensis* | Japan |

Note: *The type strains were obtained from BCCM/LMG bacterial collection, University of Ghent, Belgium. Strains with prefixes SW or DC were kindly provided by Dr. David Coplin.*