The synergistic effect of concatenation in phylogenomics: the case in *Pantoea*

Marike Palmer[1], Stephanus N. Venter[1], Alistair R. McTaggart[1,2], Martin P.A. Coetzee[1], Stephanie Van Wyk[1], Juanita R. Avontuur[1], Chrizelle W. Beukes[1], Gerda Fourie[1], Quentin C. Santana[1], Magriet A. Van Der Nest[1], Jochen Blom[3] and Emma T. Steenkamp[1]

[1] Department of Biochemistry, Genetics and Microbiology, DST-NRF Centre of Excellence in Tree Health Biotechnology (CTHB) and Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, Gauteng, South Africa

[2] Queensland Alliance for Agriculture and Food Innovation, University of Queensland, Brisbane, Queensland, Australia

[3] Bioinformatics and Systems Biology, Justus Liebig Universität Gießen, Giessen, Germany

ABSTRACT

With the increased availability of genome sequences for bacteria, it has become routine practice to construct genome-based phylogenies. These phylogenies have formed the basis for various taxonomic decisions, especially for resolving problematic relationships between taxa. Despite the popularity of concatenating shared genes to obtain well-supported phylogenies, various issues regarding this combined-evidence approach have been raised. These include the introduction of phylogenetic error into datasets, as well as incongruence due to organism-level evolutionary processes, particularly horizontal gene transfer and incomplete lineage sorting. Because of the huge effect that this could have on phylogenies, we evaluated the impact of phylogenetic conflict caused by organism-level evolutionary processes on the established species phylogeny for *Pantoea*, a member of the *Enterobacterales*. We explored the presence and distribution of phylogenetic conflict at the gene partition and nucleotide levels, by identifying putative inter-lineage recombination events that might have contributed to such conflict. Furthermore, we determined whether smaller, randomly constructed datasets had sufficient signal to reconstruct the current species tree hypothesis or if they would be overshadowed by phylogenetic incongruence. We found that no individual gene tree was fully congruent with the species phylogeny of *Pantoea*, although many of the expected nodes were supported by various individual genes across the genome. Evidence of recombination was found across all lineages within *Pantoea*, and provides support for organism-level evolutionary processes as a potential source of phylogenetic conflict. The phylogenetic signal from at least 70 random genes recovered robust, well-supported phylogenies for the backbone and most species relationships of *Pantoea*, and was unaffected by phylogenetic conflict within the dataset. Furthermore, despite providing limited resolution among taxa at the level of single gene trees, concatenated analyses of genes that were identified as having no signal resulted in a phylogeny that resembled the species phylogeny of *Pantoea*. This distribution of signal and noise across the genome presents the ideal situation for phylogenetic inference, as the topology from a ≥70-gene concatenated species phylogeny is not driven by single genes, and our data suggests that this finding may also hold true for smaller datasets. We thus argue that, by using a

Submitted 23 October 2018
Accepted 26 February 2019
Published 16 April 2019

Corresponding author
Stephanus N. Venter,
fanus.venter@up.ac.za

Academic editor
Mikhail Gelfand

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj.6698**

How to cite this article Palmer M, Venter SN, McTaggart AR, Coetzee MPA, Van Wyk S, Avontuur JR, Beukes CW, Fourie G, Santana QC, Van Der Nest MA, Blom J, Steenkamp ET. 2019. The synergistic effect of concatenation in phylogenomics: the case in *Pantoea*. *PeerJ* **7**:e6698 http://doi.org/10.7717/peerj.6698

concatenation-based approach in phylogenomics, one can obtain robust phylogenies due to the synergistic effect of the combined signal obtained from multiple genes.

Subjects Bioinformatics, Genomics, Microbiology
Keywords Phylogenomics, Concatenate, Super trees, Phylogenetics, Phylogenetic signal, Phylogenetic conflict

INTRODUCTION

Whole genome sequences are now routinely used for phylogenetic inference, particularly in bacteria (*Abdul Rahman et al., 2016*; *Beukes et al., 2017*; *Callister et al., 2008*; *Gupta, Naushad & Baker, 2015*; *Meehan & Beiko, 2014*; *Palmer et al., 2017*; *Schwartz et al., 2015*; *Zhang et al., 2011*). Many approaches for investigating evolutionary relationships across different taxonomic ranks have been developed (*Daubin, Gouy & Perrière, 2001*; *Daubin, Gouy & Perriere, 2002*; *Jolley et al., 2012*; *Yokono, Satoh & Tanaka, 2018*). These range from alignment-free approaches (*Yokono, Satoh & Tanaka, 2018*) to alignment-based analyses of a small number of highly conserved genes across large numbers of taxa (e.g., different bacterial phyla or orders (*Abdul Rahman et al., 2016*; *Gupta, Naushad & Baker, 2015*; *Jolley et al., 2012*)), to using hundreds or thousands of genes, obtained from whole genome sequences and shared by all members of smaller groups (e.g., species, genus or family (*Meehan & Beiko, 2014*; *Palmer et al., 2017*; *Schwartz et al., 2015*; *Zhang et al., 2011*)).

The use of large numbers of shared genes for phylogenetic inference, referred to here as phylogenomics (*Daubin, Gouy & Perrière, 2001*; *Daubin, Gouy & Perriere, 2002*; *Eisen & Fraser, 2003*; *Kumar et al., 2011*), have been argued to be the most reliable option for recovering a species topology reflective of vertical descent (*Andam & Gogarten, 2011*; *Coenye et al., 2005*; *Daubin, Gouy & Perriere, 2002*; *Galtier & Daubin, 2008*). This is because the massive numbers of characters sampled is thought to dilute phylogenetic conflict within the dataset, to levels where a single robust evolutionary hypothesis is obtainable (*Andam & Gogarten, 2011*; *Coenye et al., 2005*; *Cohan, 2001*; *Daubin, Gouy & Perriere, 2002*; *Galtier & Daubin, 2008*; *Klenk & Göker, 2010*). It has been suggested, particularly in bacteria, that an overall genomic core (the set of genes shared by all members of a group) exists between closely related taxa that remains evolutionarily cohesive (*Coenye et al., 2005*; *Daubin, Gouy & Perriere, 2002*). The signal found within these core genes would thus be the signal for inheritance and would be appropriate for inferring the ancestral relationships (*Daubin, Gouy & Perriere, 2002*).

Despite some evidence for a genomic core (*Callister et al., 2008*; *Daubin, Gouy & Perriere, 2002*; *Grote et al., 2012*; *Sarkar & Guttman, 2004*), numerous studies have shown that the evolutionary trajectory of genes within this subgenomic compartment may be incongruent (*Bapteste et al., 2009*; *Dagan & Martin, 2006*; *Jeffroy et al., 2006*; *Rokas et al., 2003*; *Thiergart, Landan & Martin, 2014*). *Dagan & Martin (2006)* captured this conflict in their "tree of one percent" concept. They referred to research by Ciccarelli and colleagues (*2006*), who used 31 protein sequences to recover a robust phylogenetic hypothesis across a diverse set of bacterial taxa. This was after the removal of sequences harbouring

phylogenetic conflict from a conservative average bacterial genome of 3,000 genes. In other words, the resulting phylogenetic tree that was interpreted as the evolutionary history of the taxa, was based on roughly 1% of the average genome of these taxa (*Dagan & Martin, 2006*). Additionally, research has shown that species trees may in some cases be driven by only a handful of genes, particularly where contradictory species relationships are routinely observed from single gene trees (*Salichos & Rokas, 2013*; *Shen, Hittinger & Rokas, 2017*; *Thiergart, Landan & Martin, 2014*). It is thus still unclear whether employing genome data in a concatenation-based approach is truly an appropriate way of inferring evolutionary relationships, despite the popularity of this approach.

The incongruence often observed between gene and species trees can be attributed to two main factors: phylogenetic errors and organism-level evolutionary processes (*Doyle, 1992*; *Wendel & Doyle, 1998*). Phylogenetic errors (i.e., stochastic errors due to the use of too little information and systematic errors caused by non-phylogenetic signal) during tree inferences are mainly overcome by increased character and taxon sampling (*Hedtke, Townsend & Hillis, 2006*; *Jeffroy et al., 2006*; *Palmer et al., 2017*; *Philippe et al., 2011*; *Pollock et al., 2002*; *Yokono, Satoh & Tanaka, 2018*). Organism-level evolutionary processes can be difficult to account for if they result in different evolutionary histories for genes that cannot be integrated into a single bifurcating evolutionary hypothesis (*Wendel & Doyle, 1998*). When phylogenetic error is excluded, incomplete lineage sorting (ILS) and horizontal gene transfer (HGT) are frequently the primary organism-level processes responsible for phylogenetic incongruence (*Galtier & Daubin, 2008*; *Mallet, Besansky & Hahn, 2016*; *Retchless & Lawrence, 2010*). An ongoing debate in the scientific community is whether to concatenate and risk a well-supported but incorrect species tree that also captures all phylogenetic conflict in a dataset, or pool the phylogenetic signal from hundreds of gene trees in supertree or reconciliation approaches (*Daubin, Gouy & Perrière, 2001*; *Galtier & Daubin, 2008*; *Ren, Tanaka & Yang, 2009*; *Retchless & Lawrence, 2010*; *Sanderson & Driskell, 2003*; *Szöllősi et al., 2012*; *Williams et al., 2017*). Reconciliation approaches efficiently account for HGT because genome evolution is modelled and the data produced are used for quantifying gene transfer and for inferring species trees that accommodate this process (*Szöllősi et al., 2012*).

For this study, the bacterial genus *Pantoea* was used as a model to explore the impact of potentially conflicting signal caused by organism-level evolutionary processes on the current phylogenetic hypothesis for the group. This phylogeny was constructed previously using a concatenation-based approach that accounted for the majority of known phylogenetic errors through Maximum Likelihood analyses of partitioned datasets with appropriate evolutionary models (*Palmer et al., 2017*). *Pantoea* forms part of the family *Erwiniaceae* in the order *Enterobacterales* (*Adeolu et al., 2016*) and is closely related to the genera *Erwinia* and *Tatumella* (*Adeolu et al., 2016*; *Brady et al., 2010b*; *Glaeser & Kämpfer, 2015*; *Palmer et al., 2017*). This genus has been extensively studied and represents a diverse assemblage of organisms that employs an array of different and important lifestyles (*Brady et al., 2010a*; *Brady et al., 2009*; *Brady et al., 2010b*; *Lim et al., 2014*; *Ma et al., 2016*; *Palmer et al., 2016*; *Palmer et al., 2017*; *Walterson & Stavrinides, 2015*). Our three main objectives were to (i) determine whether or not the dataset used to infer hypotheses (based on concatenation

and a multi-species coalescent approach) included phylogenetic conflict, and if so, how this conflict is distributed across the genome; (ii) to determine whether the observed conflicts could be ascribed to organism-level evolutionary processes, such as HGT and ILS; and (iii) to determine whether limited sets of genes contain enough phylogenetic signal to overshadow potential conflict within the dataset in order to obtain phylogenies resembling the species phylogenetic hypotheses for *Pantoea*. To achieve these objectives we investigated conflict at the level of gene partitions and at specific nucleotide sites to detect recombination between the different lineages of the *Pantoea* species phylogeny and also to compare regions that differed significantly in their nucleotide composition to the rest of the alignments.

MATERIALS AND METHODS

Dataset preparation

Shared genes for the 27 taxa of interest (Table 1) were determined with the Efficient Database framework for comparative Genome Analyses using BLAST score Ratios (EDGAR) server (*Blom et al., 2016*). The nucleotide sequences for all shared genes were downloaded from the EDGAR server. Subsequently, the combined file of all sequences were split into individual gene files. Multiple sequence alignments of genes were generated with MUSCLE (*Edgar, 2004*) as part of CLC Main Workbench v 7.6 (CLC Bio, Aarhus, Denmark). This was followed by manual inspection and correction of alignments in BioEdit v. 7.0.9 (*Hall, 2011*) to ensure that the correct reading-frame was selected for all genes. Genes were then trimmed in BioEdit to eliminate gene length variation due to potential differences in gene prediction across the different genomes. To generate concatenated datasets, the respective nucleotide and protein sequences were combined with FASconCAT-G v. 1.02 (*Kück & Longo, 2014*).

Phylogenetic analyses

Approximate maximum likelihood (AML; *Price, Dehal & Arkin, 2010*) analyses were performed on all individual protein sequences, as well as, on the concatenated protein and nucleotide sequence data matrices. For individual gene trees, analyses were performed in a sequential manner, utilising an in-house python script (File S1). For computational efficiency, AML analyses were employed in this study instead of traditional maximum likelihood (ML) analyses in alternate software. Time estimates for the construction of a single gene tree based on ML is ca. 27 minutes/gene (RAxML v. 8.0.20 (*Stamatakis, 2014*)) versus ca. 4 minutes/gene for AML (FastTree v. 2.1), which is not surprising as up to 100 times speed increases were reported previously (*Price, Dehal & Arkin, 2010*). All AML phylogenies were constructed with FastTree v. 2.1 (*Price, Dehal & Arkin, 2010*) using default settings. When the relationships obtained from concatenated AML analyses were not robustly supported (SH-support >0.95), these relationships were verified using RAxML v. 8.0.20 (*Stamatakis, 2014*).

A multi-species coalescent (MSC) approach (*Mirarab et al., 2014*) was employed to construct a species tree from the individual gene phylogenies. This summary method was used to reconstruct a species tree, in the presence of potential ILS (*Mirarab et al., 2014*), by

Table 1 Genome sequences utilised in this study.

Genus	Species	Strain[a]	Accession number[b]	Reference
Pantoea	*Pantoea agglomerans*	R 190	JNGC00000000.1	*Lim et al. (2014)*
	Pantoea allii	LMG 24248^T	MLFE00000000.1	*Palmer et al. (2017)*
	Pantoea ananatis	LMG 2665^T	JMJJ00000000.1	*De Maayer et al. (2014)*
	Pantoea anthophila	11-2	JXXL00000000.1	*Wan et al. (2015)*
	Pantoea brenneri	LMG 5343^T	MIEI00000000.1	*Palmer et al. (2017)*
	Pantoea conspicua	LMG 24534^T	MLFN00000000.1	*Palmer et al. (2017)*
	Pantoea cypripedii	LMG 2657^T	MLJI00000000.1	*Palmer et al. (2017)*
	Pantoea deleyi	LMG 24200^T	MIPO00000000.1	*Palmer et al. (2017)*
	Pantoea dispersa	EGD-AAK13	AVSS00000000.1	–
	Pantoea eucalypti	aB	AEDL00000000.1	–
	Pantoea eucrina	LMG 2781^T	MIPP00000000.1	*Palmer et al. (2017)*
	Pantoea rodasii	LMG 26273^T	MLFP00000000.1	*Palmer et al. (2017)*
	Pantoea rwandensis	LMG 26275^T	MLFR00000000.1	*Palmer et al. (2017)*
	Pantoea septica	LMG 5345^T	MLJJ00000000.1	*Palmer et al. (2017)*
	Pantoea stewartii subsp. *stewartii*	DC 283	AHIE00000000.1	–
	Pantoea stewartii subsp. *indologenes*	LMG 2632^T	JPKO00000000.1	–
	Pantoea vagans	C9-1	CP001894.1, CP001893.1, CP001894.1	*Smits et al. (2010)*
	Pantoea wallisii	LMG 26277^T	MLFS00000000.1	*Palmer et al. (2017)*
	Pantoea sp.	At-9b	CP002433.1, CP002434.1, CP002435.1, CP002436.1, CP002437.1, CP002438.1	*Suen et al. (2010)*
	Pantoea sp.	A4	ALXE00000000.1	*Hong et al. (2012)*
	Pantoea sp.	GM01	AKUI00000000.1	*Brown et al. (2012)*
Tatumella	*Tatumella morbirosei*	LMG 23360^T	CM003276.1	–
	Tatumella ptyseos	ATCC 33301^T	ATMJ00000000.1	–
	Tatumella saanichensis	NML 06-3099^T	ATMI00000000.1	*Tracz et al. (2015)*
Erwinia	*Erwinia billingiae*	NCPPB 661^T	FP236843.1, FP236826.1, FP236830.1	*Kube et al. (2010)*
	Erwinia pyrifoliae	DSM 12163^T	FN392235.1, FN392236.1, FN392237.1	*Kube et al. (2010)*
	Erwinia tasmaniensis	Et 1-99^T	CU468135.1, CU468128.1, CU468130.1, CU468131.1, CU468132.1, CU468133.1	*Kube et al. (2008)*

Notes.

[a]Superscript^T indicates type strains for the species.

[b]All numbers refer to GenBank assembly accession numbers (http://www.ncbi.nlm.nih.gov/; accessed 28/2/2017).

subjecting the unrooted AML phylogenies to an MSC analysis in ASTRAL v. 5.6.3 (*Mirarab et al., 2014*). Outputs were indicated with branch lengths in coalescent units and support values for the four clusters around a specific branch (quartet score; *Sayyari & Mirarab, 2016*). Additionally, three other approaches were used to infer the *Pantoea* species tree. The first involved inference of a Neighbour-Joining (NJ) tree using distances based on Average Nucleotide Identity (ANI; *Richter & Rosselló-Móra, 2009*) values. These were available from a previous study (*Palmer et al., 2017*) and used to generate a pairwise distance matrix in Microsoft Excel™ from which a NJ tree was inferred using MEGA v. 6.06 (*Tamura et al., 2013*). Note that this precluded resampling of the data for evaluating branch support. Secondly, a Neighbor-Net network was inferred from the concatenated nucleotide data using default settings in SplitsTree v. 4 (*Huson & Bryant, 2005*). Thirdly, this software was also used to construct a consensus network from the single gene AML phylogenies with a zero threshold (exclude no trees) and edge weights set to count.

To determine the degree of congruence and distribution of signal across individual gene genealogies relative to the *Pantoea* species phylogenies, individual phylograms were manually inspected. During this process, gene genealogies supporting specific backbone nodes (that were consistently recovered using multiple inference approaches) within the *Pantoea* species phylogenies were identified. This was done by evaluating a set of twelve query hypotheses (representing all of the internal backbone nodes in the *Pantoea* species trees) against each of the individual gene genealogies to determine whether they contained and/or supported the expected nodes. Each genealogy was then marked as (1) fully supporting, (2) supporting, but with other taxa nested, (3) not supporting or (4) lacking signal for the specific node depicted in the query hypothesis. The signal obtained from each of the different gene trees were then related back to the physical order of the shared genes as they appear in the genome of *P. agglomerans* (*Lim et al., 2014*), to determine whether specific signal patterns could be associated to areas of the genome.

As an indication of how phylogenetic conflicts were distributed across the concatenated nucleotide alignment, incongruent signals for the *P. dispersa* and *P. ananatis* lineages were investigated. For these purposes, the nucleotide sites causing incongruence in the Neighbor-Net network was noted and related back to the gene identifier. These data were visualised across the concatenated alignment using Circos v. 0.69 (*Krzywinski et al., 2009*).

Recombination detection

To determine whether recombination, as an organism-level evolutionary process, could have contributed to phylogenetic conflict within the dataset, genes with possible signals for recombination were identified. This was done by subjecting the concatenated data matrix to the Recombination Detection Program (RDP) v. 4.84 (*Martin et al., 2015*), to test for recombination breakpoints using five genetic distance-based methods (RDP, GENECONV, MaxChi, Chimaera and 3Seq). RDP employs a sliding window to calculate pairwise distances between all unique taxon triplets for parsimony informative sites. Regions in contradiction to a UPGMA dendrogram, constructed from all sites, are identified as potentially recombinant (*Martin & Rybicki, 2000*). The GENECONV method entails pairwise comparisons of all polymorphic sites within the alignment to identify

higher than expected similarity over unusually long regions compared to the rest of the alignment (*Padidam, Sawyer & Fauquet, 1999*). MaxChi identifies potential recombination breakpoints by examining differences in the proportions of variable polymorphic sites using a sliding window to calculate pairwise χ^2 values (*Smith, 1992*). The Chimaera approach is in essence a modification of the MaxChi method, where triplets are screened using a sliding window for only polymorphic sites where recombinants match one of the parental sequences (*Posada & Crandall, 2001*). Lastly, the 3Seq method uses the same character set as Chimaera to query each sequence within each triplet combination to determine if it could be a possible recombinant of the other two sequences (*Boni, Posada & Feldman, 2007*). These data were also plotted on the concatenated alignment using Circos v. 0.69 (*Krzywinski et al., 2009*).

Randomised subset phylogenetic analyses

To determine whether limited sets of genes contained sufficient phylogenetic signal to overcome phylogenetic conflict within the dataset, randomised subsets of 20, 50, 60, 70, 80, 90, 100, 110 and 120 genes were constructed. For this purpose, genes were randomly identified in Microsoft Excel^{TM} [=RANDBETWEEN(1,1357)] without resampling. The concatenation and phylogenetic analyses were conducted in the same manner as described above. In all cases, ten individual data subsets were constructed, followed by obtaining a strict and majority rule consensus tree of the ten phylograms of each gene set (i.e., 20, 50, 60, 70, 80, 90, 100, 110 or 120 genes).

RESULTS

Detecting phylogenetic conflict

Using the AML approach, a robust and well-supported evolutionary hypothesis regarding the species relationships in *Pantoea* was obtained. The AML phylogeny was based on 337,780 amino acid columns corresponding to the protein sequences of 1,357 genes (Fig. 1). This phylogeny was also congruent with the phylogeny obtained with a larger taxon set for *Pantoea*, *Erwinia*, *Tatumella* and outgroup taxa by *Palmer et al. (2017)*, where ML inferences were performed with the appropriate evolutionary models for each gene partition. The only exceptions were the sister-grouping between *P. agglomerans* and *P. vagans* in the current tree, however due to their close relatedness this is not an uncommon problem, and the grouping of *P. deleyi* and *P. anthophila*. Similarly, a robust and equally well-supported phylogeny was obtained using the MSC approach where the species tree was inferred from the set of 1,357 individual gene trees (Fig. 1 and Fig. S1). Overall, the MSC topology was also congruent with the phylogeny obtained by *Palmer et al. (2017)*. Exceptions were only observed at nodes at tips or leave nodes (i.e., the groupings observed in the *P. agglomerans* and *P. dispersa* lineages).

The AML and MSC topologies were highly congruent (Fig. 1). The only differences were those regarding relationships within the *P. dispersa* lineage and the *P. agglomerans* lineage (both topologies also lacked support for the relationships within this lineage). In terms of the MSC topology, comparison of the quartet scores for the main, the first alternative and second alternative topologies possible at each node (*Sayyari & Mirarab, 2016*), showed that

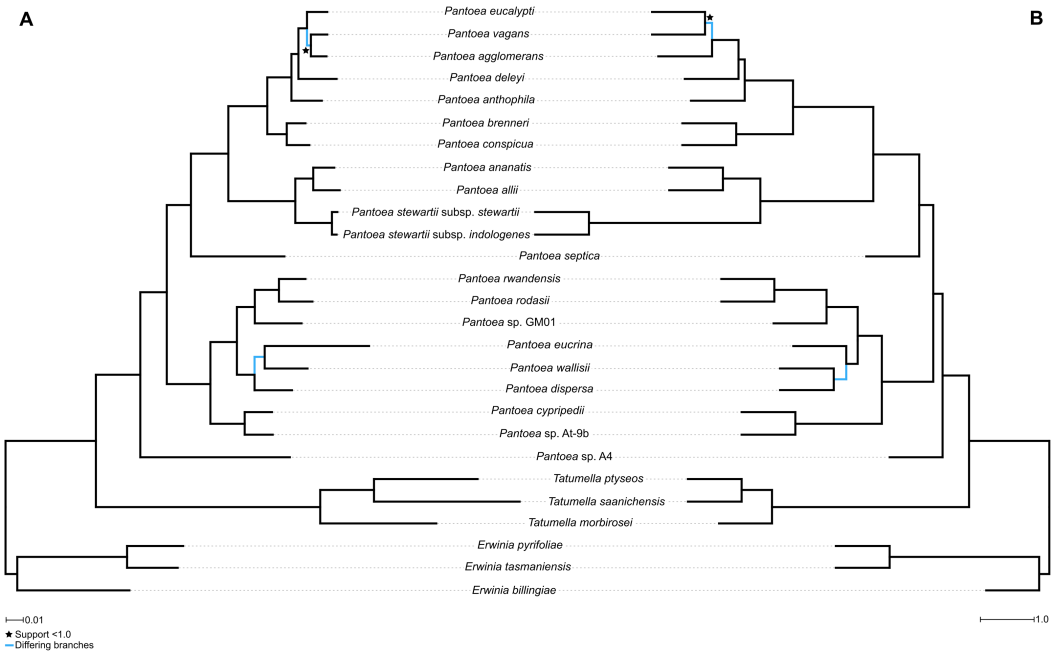

Figure 1 Comparison between the AML and MSCM phylogenies. Blue branches indicate differences in topology and branches with support of lower than 1.0 are indicated with a star. (A) The approximate maximum likelihood (AML) phylogeny constructed from the concatenated data matrix of the protein sequences of 1,357 genes, consisting of 337,780 amino acid columns. The phylogeny was constructed with FastTree v. 2.1 (*Price, Dehal & Arkin, 2010*) with the JTT (*Jones, Taylor & Thornton, 1992*) evolutionary model with CAT approximation. Simodaira-Hasegawa branch support values from 1,000 replicates were used. (B) A species phylogeny using the multispecies coalescent model as implemented in ASTRAL v.5.6.3 (*Mirarab et al., 2014*) based on the individual phylogenies constructed from the protein sequences of 1,357 genes. The scale bar indicates one coalescent unit (*Mirarab et al., 2014*). Terminal branches are indicated as one coalescent unit, as branch lengths for taxa corresponding to species can only be calculated where multiple individuals per species are analysed. Shorter branches correspond to higher levels of incongruence and are generally associated with high levels of incomplete lineage sorting (ILS). Support values are determined based on Bayesian posterior probability values computed from the single gene tree quartet frequencies (*Sayyari & Mirarab, 2016*).

Full-size DOI: 10.7717/peerj.6698/fig-1

the nodes where quartet scores between the topologies differed very little (where quartet scores for the three alternatives were almost equal) were generally those responsible for incongruence between the topologies inferred using different approaches (Fig. 1 and Fig. S1; particularly within the *P. agglomerans* and *P. dispersa* lineages). This suggests that none of these approaches are particularly robust when resolving closely related or undersampled lineages close to the leaves of the phylogenies.

The network approaches indicated a large amount of conflicting signal within the data. This was evident in the Neighbor-Net network (Fig. 2 and File S2) based on the concatenated nucleotide data matrix (1,010,946 bases), as well as the Consensus Network (Fig. S2 and File S3) of the individual gene trees (1,357 protein sequences). These conflicting signals were particularly prevalent at the deeper edges of the evolutionary hypotheses, e.g., the *P. dispersa* lineage compared to the *P. ananatis* lineage (denoted A and B in Fig. 1). However, despite

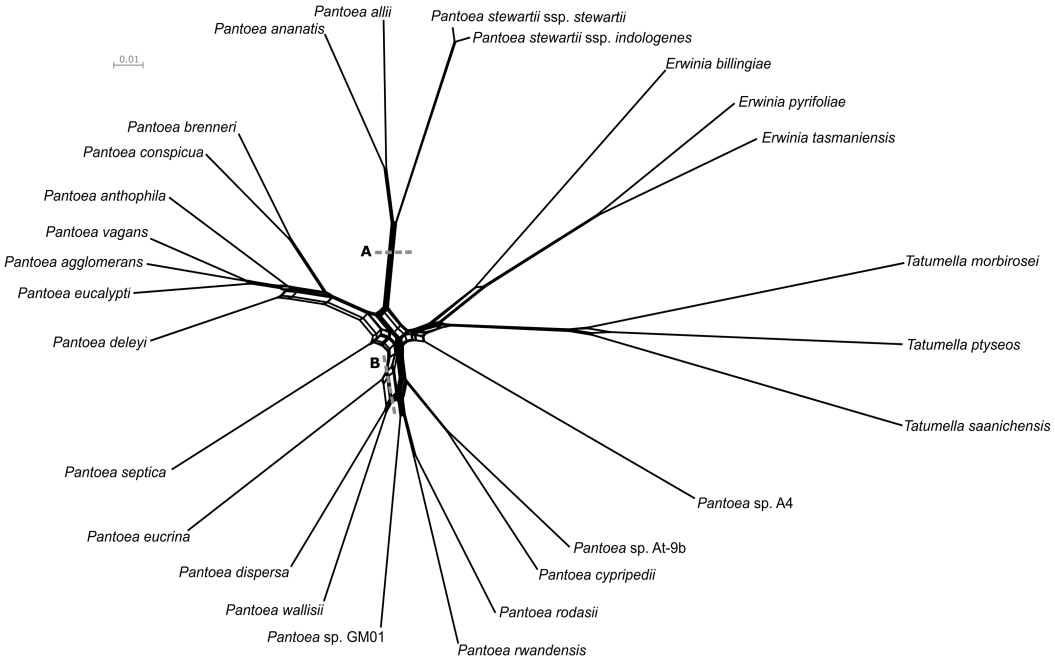

Figure 2 Neighbor-Net network from the concatenated nucleotide data. The Neighbor-Net network was constructed from *p*-distances with equal angles for the concatenated nucleotide dataset. Overall, the configuration of the network is congruent with the existing species phylogeny for the genus *Pantoea*. Clear separation between the *P. agglomerans* and *P. ananatis* lineages were obtained and were also clearly distinct from the *P. rodasii* and *P. dispersa* lineages. Point A denotes where signal in conflict to the grouping of the *P. ananatis* lineage was determined, while point B denotes where signal in conflict to the grouping of the *P. dispersa* lineage was determined (see text for details).

Full-size DOI: 10.7717/peerj.6698/fig-2

the presence of this conflict, the evolutionary hypotheses obtained with the networks, overall, reflected the relationships obtained for the AML and MSC phylogenies (Figs. 1 and 2 and Fig. S2). Furthermore, the topology obtained for the ANI-based distances was mostly congruent to the lineages recovered from the various species tree inference approaches (Fig. S3). All backbone nodes that were consistently recovered in the other approaches, were also recovered with the ANI-based distances with the exception of *P. eucrina* grouping as sister to the singleton, *P. septica*.

To determine the degree of incongruence caused by phylogenetic conflict, comparisons of all individual gene trees were evaluated against a set of twelve query phylogenetic hypotheses (Fig. 3A). These query hypotheses were constructed to evaluate monophyly of lineages or groups across the backbone of the *Pantoea* species phylogenies, thus shallower nodes near the tips of the trees (leaves) were not considered. None of the 1,357 gene trees were fully congruent with the respective phylogenetic hypotheses of *Pantoea*. Of the individual gene trees, only six genes supported all the nodes in the backbone for the groupings observed previously (File S4). Additionally, seven gene genealogies produced polytomies of taxa and thus were marked as containing no signal for any of the nodes observed in the phylogenetic hypotheses of *Pantoea* (Fig. 4 and File S4). The remaining gene trees supported at least one of the nodes in the backbone observed in the *Pantoea* species trees. Exclusion from

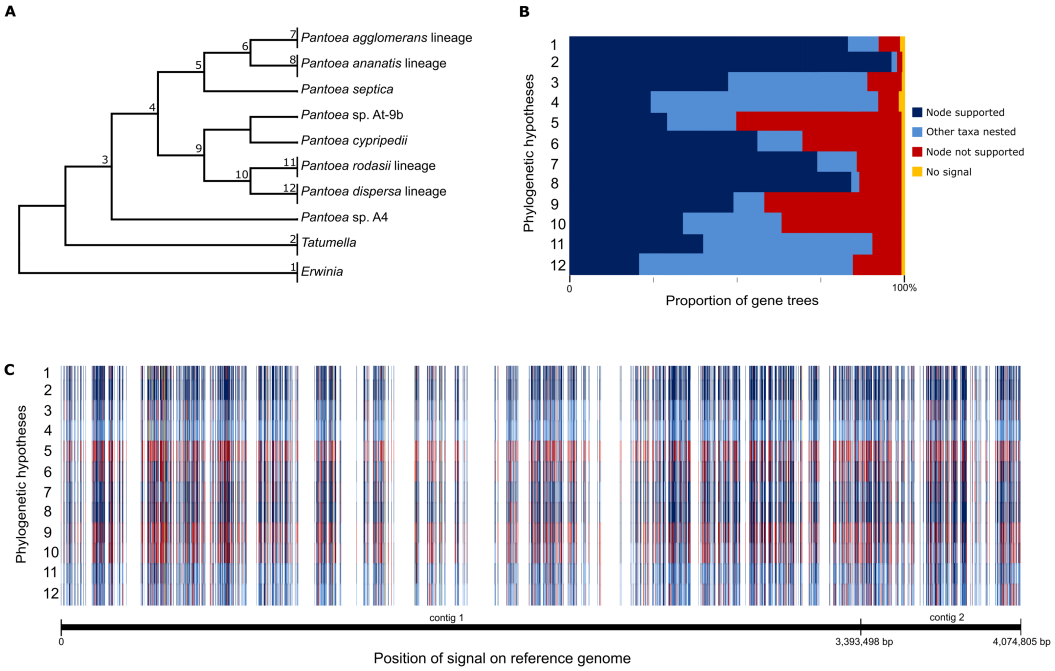

Figure 3 The summary of individual gene tree comparisons. (A) The phylogenetic hypotheses evaluated during topology comparisons. Each number represents a specific hypothesis, where the monophyly of the group at each node was evaluated. An example is hypothesis 3, where the overall monophyly of *Pantoea* was evaluated. (B) A relative frequency histogram depicting the proportion of individual gene trees that support the phylogenetic hypotheses evaluated. Dark blue indicates genes that fully supported the monophyly of the corresponding hypothesis, while light blue indicates support for the monophyly of the hypothesis with additional taxa nested within the group. Red indicates gene trees that were incongruent with the corresponding hypotheses and yellow denotes gene trees with no signal (polytomies). (C) The signal obtained for each gene genealogy compared to the phylogenetic hypotheses were plotted against the position of the genes on the chromosome of *Pantoea agglomerans*. The same colour scheme is applied as in the frequency histogram. All genes were located on the chromosome of *P. agglomerans* R190 and was distributed across the chromosome consisting of two contigs. Signal for the respective nodes within the species phylogeny were distributed across the chromosome and no patterns of shared signal were detected for groups of adjacent genes.

Full-size DOI: 10.7717/peerj.6698/fig-3

the concatenated analyses of either the six backbone-supporting genes or the seven genes providing no resolution among taxa, still provided the same overall topology, with the exception of the grouping of *P. agglomerans*, *P. vagans* and *P. eucalypti* (Figs. S4A and S4B) that lacked statistical support. The phylogenies constructed from the concatenated datasets with only the six backbone-supporting genes and only the seven genes showing no signal (confirmed with RAxML v. 8.0.20; Fig. 3H), also allowed the recovery of a mostly congruent phylogeny to that of the expected topology, but with very low or no support at a number of nodes and slight interspecies differences in the *P. agglomerans* lineage and the position of singleton taxa (Fig. 3H, Figs. S5A and S5B).

Based on these topology comparisons, it appeared that the signal supporting different nodes across the *Pantoea* species phylogenies were supported by different genes. As a means to investigate the distribution of phylogenetic conflict at the gene partition-level across the

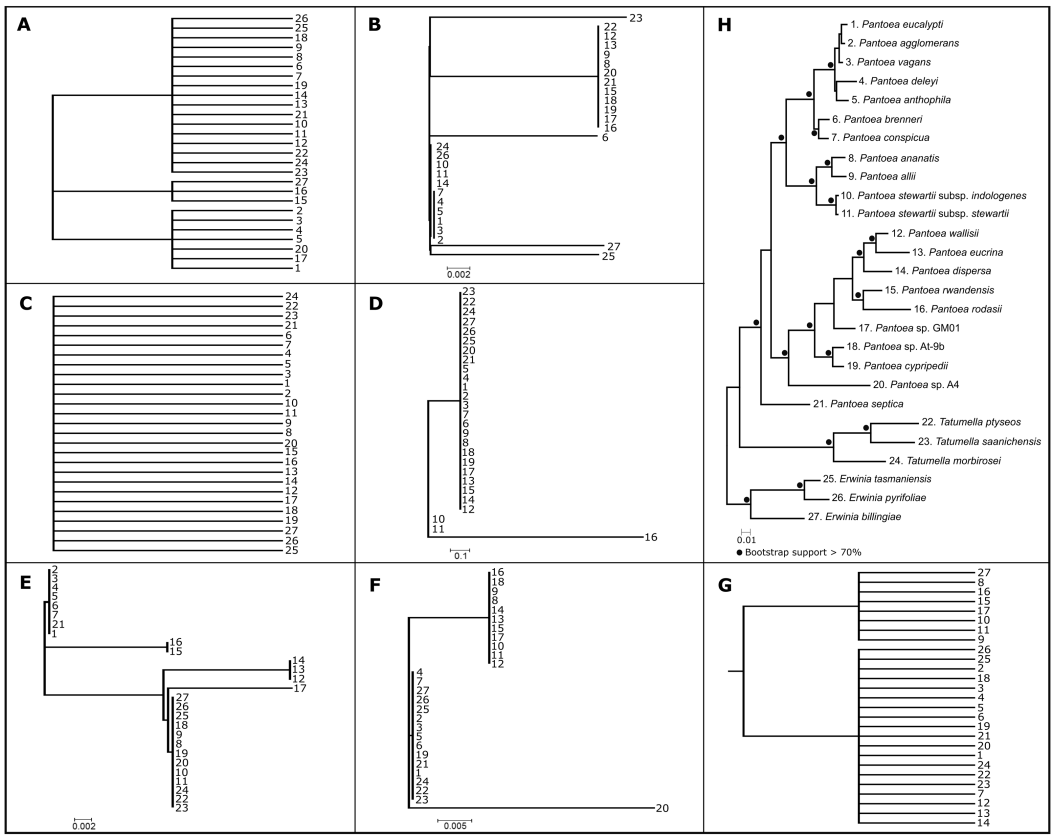

Figure 4 Summary of genes with limited to no signal. Seven single gene phylogenies determined with approximate maximum likelihood (AML) analyses for genes identified as containing no signal (see File S4) and the maximum likelihood (ML) phylogeny inferred from the combined sequence of these seven genes. (A–G) The AML phylogeny constructed from the protein sequences for the 30S ribosomal protein S18, UDP-diphospho-muramoylpentapeptide beta-N-acetylglucosaminyltransferase, Prolyl-tRNA synthetase, Glutamate 5-kinase, Cold shock-like protein *cspC*, 30S ribosomal protein S10 and 30S ribosomal protein S12, respectively. Taxa are numbered according to taxon descriptors in H. (H) The concatenated ML phylogeny constructed using RAxML v. 8.0.20 with the appropriate amino acid model inferred using ProtTest v. 3.4 for each partition. All bootstrap support values above 70% are indicated at nodes with dots. The phylogeny resembles the known species phylogeny for *Pantoea* with the exception of some species relationships within the *P. agglomerans* and *P. rodasii* lineages and the grouping of singleton taxa.

Full-size DOI: 10.7717/peerj.6698/fig-4

genome, the signal for each gene was plotted against the genome of *P. agglomerans* (Fig. 3C). All shared genes were localized to the chromosome of *P. agglomerans*. This analysis also revealed that signal for all nodes were randomly distributed across the chromosome of *P. agglomerans* and no apparent patterns of shared signal were detected for adjacent genes (Fig. 3C).

To interrogate the distribution of conflict across the dataset at the nucleotide-level, nucleotide positions in phylogenetic conflict with relationships observed within the *Pantoea* species phylogeny were identified. Of the 1,010,946 bases within the nucleotide alignment, 493,834 bases (48.7%) were identical across all taxa, with 517,112 nucleotide positions being variable between taxa. For these analyses the *P. ananatis* lineage, with the

least conflicting signal within the dataset (Fig. 2A), and the *P. dispersa* lineage, with the most conflicting signal within the dataset (Fig. 2B), were investigated as a best and worst case scenario, respectively. For the *P. ananatis* and *P. dispersa* lineages, a total of 1,764 and 3,856 nucleotide sites, respectively, supported relationships differing from the *Pantoea* species phylogenies (Fig. 2 points A and B; File S5). However, these sites were distributed across the concatenated alignment and were not localized to specific genomic regions (Fig. 5).

Recombination as a source of phylogenetic conflict

Using RDP, a total of 276 potential recombination events were detected (File S6), with 166 of these indicated as potentially caused by evolutionary processes other than recombination (*Martin et al., 2015*). This yielded 110 likely recombination events, occurring across 54 regions of the concatenated sequence, supported by at least three different analytical methods, of which 57 events were supported by all five methods (File S6). However, to avoid the inclusion of potentially artefactual recombination breakpoints associated with the concatenation process, we only considered recombination breakpoints occurring within the boundaries of single genes. This yielded a total of 15 recombination events, identified across 11 genes within the concatenated alignment (Fig. 5 and File S6). From these results, recombination break-points were detected in members of all lineages within *Pantoea*. None of these recombination breakpoints could, however, be linked to the nucleotide-level phylogenetic conflict identified (File S5).

Phylograms from limited sets of genes

The topology of the *Pantoea* species phylogenies could be recovered by some of the randomised subsets of 20, 50 and 60 genes, whereas all subsets containing the information for 70 or more genes recovered these nodes. Within each set of ten replicate data subsets, the length of individual alignments differed depending on the length of the specific genes used to construct them (Table 2 and File S7). For example, for the 20 gene subsets, the lengths of the alignments ranged from 5,560 to 7,152 amino acid columns, while the length of the alignments for the 120 gene subsets ranged from 36,614 to 42,793 amino acid columns (Table 2).

Overall, support for the backbone of the *Pantoea* species tree (Fig. 3A) deteriorated with a decrease in the number of genes concatenated and analysed (Table 2). When fewer genes were analysed in the replicates, various branches collapsed and branch support decreased in the strict consensus trees (Table 2, Fig. 6 and Fig. S6). Overall, strict consensus trees from the various replicates of 70 and more genes resulted in the recovery of a phylogeny congruent with the species phylogenies of *Pantoea*, however multiple individual replicates of the smaller datasets produced trees that were largely incongruent with these hypotheses. Only the trees from multi-gene subsets of 70 or more genes, consistently allowed robust and well-supported reconstruction of the expected *Pantoea* species trees, specifically with regards to branches in the backbone of the phylogeny (i.e., query hypotheses; Fig. 6).

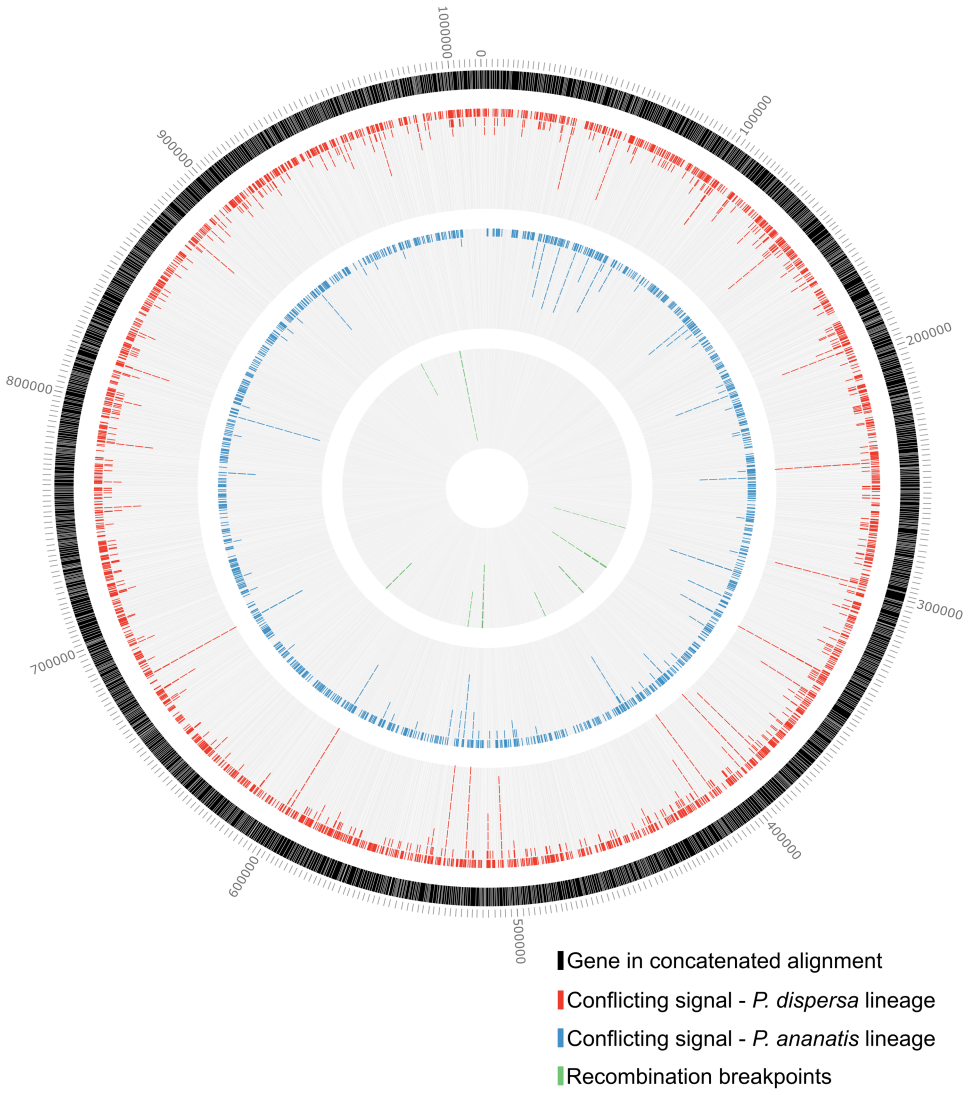

Figure 5 Conflicting signal and possible recombination breakpoints. A circular diagram depicting the nucleotide concatenated alignment of all shared genes. The outer track indicates the gene boundaries within the alignment, with tick marks representing the length in nucleotides at 2,000 bp intervals. The second track indicates the nucleotide positions within genes supporting conflicting topologies for the *P. dispersa* lineage to species groupings observed in the concatenated species phylogeny. A total of 3,856 nucleotide positions supported conflicting topologies for the *P. dispersa* lineage. The third track indicates nucleotide positions supporting conflicting topologies for the *P. ananatis* lineage compared to species groupings observed in the concatenated species phylogeny. For this lineage 1,764 nucleotide positions supported conflicting topologies for the *P. ananatis* lineage. The inner track represents recombination breakpoints detected within gene boundaries for the concatenated alignment. These breakpoints were supported by at least three of the five methods employed (RDP, GENECONV, Chimaera, MaxChi and 3Seq) for detecting recombination. Stacked tiles reflect the number of methods that were successful in detecting recombination events at those regions, as well as multiple recombination events within the same region in various species (See File S6).

Full-size DOI: 10.7717/peerj.6698/fig-5

Table 2 Summary of gene subset tests[a].

Number of genes in subset	Replicate	Length (bp)	Backbone nodes support range[b]	Leaf nodes support range
120 genes	1	40,069	1.00	0.50–1.00
	2	36,614	1.00	0.37–1.00
	3	39,663	1.00	0.93–1.00
	4	38,050	1.00	0.43–1.00
	5	37,931	1.00	0.46–1.00
	6	40,260	1.00	0.57–1.00
	7	39,776	1.00	0.82–1.00
	8	40,385	1.00	0.86–1.00
	9	42,793	1.00	0.67–1.00
	10	39,328	1.00	0.59–1.00
110 genes	1	35,298	1.00	0.95–1.00
	2	35,349	1.00	0.69–1.00
	3	36,798	1.00	0.86–1.00
	4	38,800	1.00	0.40–1.00
	5	35,445	1.00	0.73–1.00
	6	38,042	1.00	0.71–1.00
	7	40,172	0.99–1.00	0.38–1.00
	8	39,865	1.00	0.18–1.00
	9	40,737	1.00	0.78–1.00
	10	40,745	1.00	0.54–1.00
100 genes	1	33,340	1.00	0.78–1.00
	2	30,822	0.99–1.00	0.47–1.00
	3	33,433	1.00	0.65–1.00
	4	30,707	1.00	0.90–1.00
	5	31,340	1.00	0.58–1.00
	6	31,798	1.00	0.87–1.00
	7	29.562	1.00	0.64–1.00
	8	30,773	1.00	0.06–1.00
	9	34,064	1.00	0.88–1.00
	10	35,550	1.00	0.68–1.00
90 genes	1	31,353	1.00	0.68–1.00
	2	29,307	0.99–1.00	0.91–1.00
	3	31,941	1.00	0.94–1.00
	4	31,890	1.00	0.77–1.00
	5	29,695	1.00	0.84–1.00
	6	30,564	1.00	0.37–1.00
	7	25,162	1.00	0.78–1.00
	8	30,745	1.00	0.48–1.00
	9	28,146	1.00	0.55–1.00
	10	28,883	1.00	0.81–1.00

(*continued on next page*)

Table 2 (*continued*)

Number of genes in subset	Replicate	Length (bp)	Backbone nodes support range[b]	Leaf nodes support range
80 genes	1	24,020	1.00	0.35–1.00
	2	23,065	1.00	0.73–1.00
	3	25,922	0.99–1.00	0.86–1.00
	4	27,877	1.00	0.53–1.00
	5	25,288	0.99–1.00	0.70–1.00
	6	22,551	0.98–1.00	0.69–1.00
	7	26,417	1.00	0.59–1.00
	8	27,008	1.00	0.72–1.00
	9	25,156	1.00	0.30–1.00
	10	25,498	0.98–1.00	0.77–1.00
70 genes	1	22,011	0.99–1.00	0.16–1.00
	2	24,373	1.00	0.54–1.00
	3	24,420	1.00	0.45–1.00
	4	20,887	1.00	0.82–1.00
	5	22,286	0.99–1.00	0.11–1.00
	6	22,702	1.00	0.27–1.00
	7	23,787	0.99–1.00	0.83–1.00
	8	19,750	1.00	0.21–1.00
	9	23,770	0.99–1.00	0.68–1.00
	10	21,613	1.00	0.31–1.00
60 genes	1	18,755	0.99 - 1.00	0.92–1.00
	2	21,310	0.99–1.00	0.89–1.00
	3	21,745	0.99–1.00	0.77–1.00
	4	19,210	1.00	0.83–1.00
	5	19,495	1.00	0.83–1.00
	6	18,550	0.83 - 1.00	0.69–1.00
	7	20,389	1.00	0.58–1.00
	8	20,475	0.77–1.00	0.47–1.00
	9	17,331	1.00	0.01–1.00
	10	23,324	0.97–1.00	0.40–1.00
50 genes	1	14,890	1.00*	0.31–1.00
	2	18,079	1.00	0.27–1.00
	3	14,701	1.00	0.81–1.00
	4	13,983	0.71–1.00	0.00–1.00
	5	19,059	1.00	0.40–1.00
	6	18,412	0.99–1.00	0.86–1.00
	7	18,880	1.00	0.59–1.00
	8	14,411	0.85–1.00	0.33–1.00
	9	14,942	0.81–1.00	0.67–1.00
	10	14,531	1.00	0.28–1.00

(*continued on next page*)

Table 2 (*continued*)

Number of genes in subset	Replicate	Length (bp)	Backbone nodes support range[b]	Leaf nodes support range
20 genes	1	5,966	0.87–1.00	0.74–1.00
	2	5,834	0.99–1.00*	0.64–1.00
	3	6,859	0.98–1.00	0.58–1.00
	4	7,152	0.97–1.00*	0.00–1.00
	5	5,560	0.99–1.00	0.26–1.00
	6	6,316	0.48–1.00	0.71–1.00
	7	6,210	0.93–1.00	0.26–1.00
	8	6,517	0.99–1.00	0.22–1.00
	9	6,649	0.95–1.00	0.06–1.00
	10	6,436	0.90–1.00	0.53–1.00

Notes.

[a]See File S3.

[b]One backbone node is not recovered in the phylogenies marked with an asterisk.

DISCUSSION

This study employed a novel approach to investigate phylogenetic conflict within concatenated datasets. We interrogated the distribution and effect of both phylogenetic signal and conflict at the gene partition and nucleotide levels. This entailed the use of various phylogenetic analyses coupled with manual inspection and evaluation of individual gene trees. These data revealed the effects of phylogenetic conflict and signal in concatenated datasets, which are the input typically used for phylogenomic reconstruction. Our findings support the idea that all genes, even if they appear to be phylogenetically uninformative when analysed alone, contribute signal toward a phylogenomic evolutionary hypothesis and that the obtained topology is not driven by single genes. This is reminiscent of Aristotle's idea of synergism that "the whole is greater than the sum of its parts". In other words, by concatenating single genes a synergistic effect is achieved, where the combined data seems to be superior to that of the proverbial sum of the signal.

As demonstrated previously (*Palmer et al., 2017*), the full set of shared genes allowed reconstruction of a robustly supported phylogenetic hypothesis for *Pantoea* using AML. In fact, it is quite common to obtain a robust, highly supported phylogeny through concatenation of all shared genes, despite the incongruent nature of individual gene trees (*Hedtke, Townsend & Hillis, 2006*; *Jeffroy et al., 2006*; *Rokas et al., 2003*; *Salichos & Rokas, 2013*; *Thiergart, Landan & Martin, 2014*). However, none of our single gene genealogies were fully congruent with the phylogenomic species tree of *Pantoea*, while only six genes allowed recovery of the backbone of the species phylogeny. This was not surprising as various previous studies showed that very few or no genes typically support a particular species phylogeny fully (*Dagan & Martin, 2006*; *Hedtke, Townsend & Hillis, 2006*; *Jeffroy et al., 2006*; *Rokas et al., 2003*; *Salichos & Rokas, 2013*; *Thiergart, Landan & Martin, 2014*). In contrast to conclusions drawn previously (*Thiergart, Landan & Martin, 2014*), most of the *Pantoea* gene trees supported at least some of the nodes within the species phylogeny. In other words, support for the respective nodes was not necessarily obtained from the same genes but rather scattered across different genes.

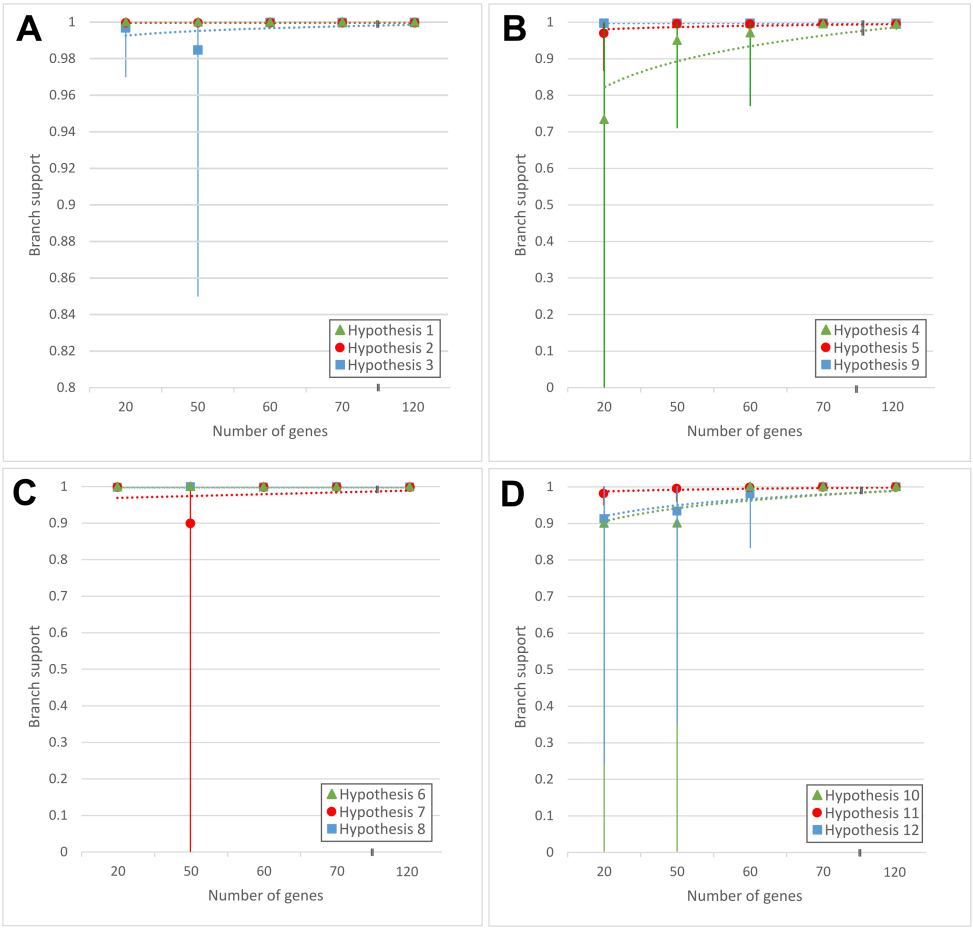

Figure 6 The SH branch support for specific hypotheses in the trees constructed from the subset datasets. Each hypothesis (Fig. 3A) was interrogated in each of the subset tree datasets, where 20, 50, 60, 70, 80, 90, 100, 110 and 120 genes were used to construct ten randomised datasets for each number of genes (File S7). The range indicated for each data point stretches from the lowest branch support (0 in the case where the nodes were not recovered) to the highest branch support (1 where the branch was fully supported) with the mean indicated with the data point. Regression analyses were performed in Microsoft ExcelTM to fit the best regression model to the data. (A) Support for the three hypotheses depicted represent the monophyly of the three genera *Erwinia* (hypothesis 1; green), *Tatumella* (hypothesis 2; red) and *Pantoea* (hypothesis 3; blue). All subsets datasets recovered the nodes representing the monophyly of the genera, but in the case of *Pantoea*, with less support in the replicates of the lower number of genes. (B) The support for the three test hypotheses (i) separating the remainder of *Pantoea* from *Pantoea* sp. A4 (hypothesis 4; green), (ii) grouping *P. septica*, and the *P. agglomerans* and *P. ananatis* lineages together (hypothesis 5; red) and (iii) the grouping of *P. cypripedii* and *Pantoea* sp. At-9b with the *P. rodasii* and *P. dispersa* lineages (hypothesis 9; blue). The node representing hypothesis 4 were not recovered in two repeats of the 20 gene subsets. (C) Hypothesis 6 depicts the sister grouping of the *P. agglomerans* and *P. ananatis* lineages with the support associated with the node depicted in green. The monophyletic grouping of the *P. agglomerans* lineage (hypothesis 7; red) were not recovered in one 50 gene repeat, but were further fully supported in all repeats. The grouping of the *P. ananatis* lineage was consistently recovered with full support (hypothesis 8; blue). (D) The support associated with the nodes depicting the sister relationship between the *P. rodasii* and *P. dispersa* lineages (hypothesis 10; green). This node was not recovered in one of the 20 gene repeats and one of the 50 gene repeats. hypothesis 11 represents the monophyletic grouping of the *P. rodasii* lineage, which was consistently recovered and well-supported with branch support >0.95 (red). Branch support associated with the monophyletic grouping of the *P. dispersa* lineage often ranged from very low (0.24) to fully supported (1) in the 20, 50 and 60 gene repeats (hypothesis 12; blue).

Full-size DOI: 10.7717/peerj.6698/fig-6

A fully resolved, well-supported phylogeny was obtained using the MSC approach, although it was not congruent with the AML tree regarding relationships at the tips of the trees. Concatenation is thought to be superior to MSC-based species trees if ILS is low, while MSC models perform better in the presence of moderate ILS (*Mirarab et al., 2014*). As our MSC and AML trees correspond perfectly regarding backbone nodes, these are strong hypotheses that may approach the real relationships among these taxa. However, because organism-level evolutionary processes were not quantified in this study, we cannot exclude the possibility that ILS were responsible for the incongruences observed. Our study therefore highlights that alternative approaches that model genome evolution, quantify organism-level evolutionary processes (*Szöllősi et al., 2012*; *Williams et al., 2017*), and that focusses on leave taxa are needed to fully resolve the species tree of a diverse assemblage such as *Pantoea*. This may be particularly true when taxa are very closely related or when the lineages in question are undersampled.

The random distribution of signal across the *Pantoea* genome is supported by the recovery of overall congruent subset phylogenies from random sub-samplings of gene sequences. Due to this random distribution, one should be able to obtain sufficient signal to reconstruct the species phylogeny by randomly sampling enough genes from the genome (*Dutilh et al., 2004*; *Gadagkar, Rosenberg & Kumar, 2005*). From our data, this idea was tested with consensus trees of 10 replicates with 20, 50, 60, 70, 80, 90, 100, 110 and 120 genes. We found that, with a decrease in the number of genes analysed, support for the backbone and the deeper branches decreased incrementally, as has also been observed previously (*Rokas et al., 2003*). Therefore, for these data and taxon set, it appear that at least 70 randomly selected genes from the genome is required to obtain a relatively robust, well-supported phylogeny, particularly to reconstruct the deeper relationships within and among the genera. Multi-gene phylogenies based on 70 genes may thus provide sufficiently robust hypotheses so that complete genome sequence data may not be required, our work suggests that sufficient data may be obtained from low level sequencing, although verification of this notion in other taxon sets is required.

The species tree hypotheses for *Pantoea* was generally also supported by the two network approaches employed here. Both accommodated non-vertical and non-phylogenetic signal (introduced through systematic error) as inferred from nucleotide data, as well as the individual gene trees (*Bryant & Moulton, 2002*; *Holland & Moulton, 2003*; *Holland, Jermiin & Moulton, 2005*). These methods produced networks in which the overall clustering patterns were generally congruent with that obtained through gene concatenation-based phylogenomic inferences. This would not have happened if insufficient signal (i.e., stochastic error *Jeffroy et al., 2006*; *Philippe et al., 2011*; *Rosenberg & Kumar, 2003*) or reconstruction artefacts (i.e., systematic error *Hedtke, Townsend & Hillis, 2006*; *Hillis, 1998*; *Pollock et al., 2002*; *Zwickl & Hillis, 2002*) were responsible for the observed relationships in the species trees. If conflict, particularly in the form of HGT and ILS, dominated the dataset, the splits graphs would not have such high overall congruence to the species trees (*Bryant & Moulton, 2002*), however, more in-depth future studies are required to fully elucidate the role and amount of organism-level evolutionary processes in the evolution of these taxa. Compared to previous analyses, often employing a limited gene set (*Chen et al.,*

2013; Kennedy et al., 2005), more box-like structures were observed in *Pantoea* networks, particularly in deeper edges. However, these boxes were generally smaller, where the increased number of boxes indicate more alternate or conflicting relationships, while the shorter edges correlate to the particular relationships being observed less frequently. If one considers that these conflicts are visualized for the full shared gene set, the level of conflict appears to be relatively low and comparable to that seen in other bacteria (Retchless & Lawrence, 2010), but considerably lower than taxa undergoing extensive HGT (Doroghazi & Buckley, 2010). The generally low level of conflict thus supports the idea that sufficient phylogenetic signal is present within the concatenated dataset to overshadow the limited conflict present.

Overall, the relationships obtained using the ANI-based distance approach were mostly congruent to the species trees. The incongruences that were present can be ascribed to the fact that ANI is notoriously unreliable as an indicator of relatedness, especially among more distantly related taxa (Konstantinidis & Tiedje, 2007; Palmer et al., 2017; Qin et al., 2014; Rosselló-Mora, 2005). This phenomenon is the reason why many prokaryotic taxonomist would rather purport the use of Average Amino Acid Identity (AAI) values at this level (Konstantinidis & Tiedje, 2007; Qin et al., 2014; Rosselló-Mora, 2005), as substitution saturation and other factors resulting from endogenous evolutionary processes may be responsible for the decline in informativeness of this metric, the more distantly related the taxa become (Palmer et al., 2017).

We investigated the phylogenetic conflict within the *Pantoea* dataset for the two lineages in which we observed the least and most phylogenetic conflict. Respectively, these were the *P. ananatis* lineage, which includes plant pathogenic species, and the *P. dispersa* lineage, which includes generalists (Palmer et al., 2018; Palmer et al., 2017; Walterson & Stavrinides, 2015). The number of nucleotide sites supporting alternate topologies to the species trees were limited, with only 0.75% of variable nucleotide sites (3,856 sites out of 517,112) supporting conflicting topologies in the *P. dispersa* lineage. Also, the remaining variable sites did not necessarily support the species relationships observed in the species trees, because different genes and nucleotide positions supported different nodes within the species phylogenies. Moreover, the conflicting signal within the dataset was not localised to specific genomic regions or genes, but rather, was randomly distributed. Taken together, these results thus suggest that (i) the use of network approaches for constructing phylogenies can be extremely valuable for identifying phylogenetic conflicts in datasets (Bryant & Moulton, 2002; Holland & Moulton, 2003; Holland, Jermiin & Moulton, 2005), and (ii) organism-level evolutionary processes like HGT and/or ILS impacts different lineages and taxa to varying degrees.

Conflicting phylogenetic signal in the *Pantoea* dataset could potentially result from recombination events that led to gene conversions between species (Daubin, Moran & Ochman, 2003; Fraser, Hanage & Spratt, 2007; Holmes, Urwin & Maiden, 1999; Posada & Crandall, 2001; Posada & Crandall, 2002). We found evidence for at least 15 recombination events in 11 shared genes in the dataset. We attributed these to recent instances of recombination, because older organism-level evolutionary events, particularly ancient HGT and ILS (Knowles, 2009; Meng & Kubatko, 2009; Retchless & Lawrence, 2010), become difficult to detect due to deterioration of signals by endogenous evolutionary

processes (*Daubin, Moran & Ochman, 2003*). It is also difficult to distinguish between these organism-level evolutionary processes as their signals may appear very similar (*Knowles, 2009*; *Nosil, 2008*; *Wendel & Doyle, 1998*) and future studies would be required to tease apart the roles of these processes in *Pantoea*. Nevertheless, identification of some of these organism-level evolutionary events in *Pantoea* provides possible mechanisms for how phylogenetic conflict could have been introduced into the data.

The use of all shared genomic information allowed for the recovery of robustly supported relationships, overcoming the weaknesses observed in individual gene datasets (*Andam & Gogarten, 2011*; *Daubin, Gouy & Perriere, 2002*; *Gadagkar, Rosenberg & Kumar, 2005*; *Galtier & Daubin, 2008*). In contrast to a previous similar study by *Thiergart, Landan & Martin (2014)*, comparison of the single gene trees were specifically performed with backbone nodes (excluding taxa closer to the tips of the trees) with the aid of the query hypotheses, which allowed us to interrogate each node and its associated signal manually. Although *Thiergart, Landan & Martin (2014)* also compared nodes between concatenated trees and the single gene phylogenies, their comparisons were focussed only on the recovery of identical nodes, which likely overestimated the effect of finer differences between the trees, leading them to their conclusion that the signal associated with the backbone or deeper nodes of their concatenated phylogenies are not preserved in single gene trees. Based on our data, three of the query hypotheses evaluated (see Fig. 3B and Fig. S1) had a large proportion of individual gene trees that did not support the expected monophyly of the taxon groups specified. These were hypothesis 5 in which *P. septica* is a singleton taxon placed as sister to the *P. agglomerans* and *P. ananatis* lineages (48.9% trees), hypothesis 9 in which *P. cypripedii* and *Pantoea* sp. At-9b are placed as sister to the *P. rodasii* and *P. dispersa* lineages (40.4% trees) and hypothesis 10 in which the *P. rodasii* and *P. dispersa* lineages are placed as sister groups (35.2% trees). In these instances, limited species have been sampled for the respective lineages. This undersampling of the diversity may contribute to the lack in robust recovery of the lineages due to large systematic error in the smaller datasets. In future, increased taxon sampling may resolve these problematic relationships more accurately in smaller datasets like those employed for the single gene trees (*Hedtke, Townsend & Hillis, 2006*; *Pollock et al., 2002*).

Comparison of single gene trees with phylogenies obtained from concatenated datasets, presents both a philosophical and logical quandary. It is widely accepted that single gene phylogenies, often with very limited or no statistical support, cannot be equated to a species phylogeny (*Degnan & Rosenberg, 2006*; *Doyle, 1992*; *Maddison, 1997*; *Pamilo & Nei, 1988*; *Rosenberg, 2002*). Despite the common practice of evaluating the robustness of a species phylogeny constructed from thousands or millions of characters, by its topological congruence to single gene trees (*Bapteste et al., 2009*; *Ciccarelli et al., 2006*; *Dagan & Martin, 2006*), these phylogenies are clearly not directly comparable and no conclusions regarding species evolution should be drawn from raw tree topology comparisons. This rationale is like comparing single molecules with chemical compounds and being disappointed that they do not share the same characteristics. Based on our data, the signal required for reconstructing a species phylogeny is dispersed and the only appropriate comparison of single gene trees to species trees would be when focus is placed on the evolution of a

particular gene or when species trees are inferred from single gene trees, as with the MSCM analyses.

Our findings confirm the robustness of phylogenies constructed from genomic data, based on the synergistic effect of combined genes, despite high levels of incongruence between individual gene trees. This is due to the phylogenetic signal for different nodes within the species phylogeny being distributed across the genome at higher levels than the randomly distributed conflicts within the dataset. These findings support previous conclusions suggested by several authors (*Andam & Gogarten, 2011*; *Daubin, Gouy & Perriere, 2002*; *Galtier & Daubin, 2008*; *Retchless & Lawrence, 2010*; *Rokas et al., 2003*), based on comparisons of single gene phylogenies with super trees and concatenated analyses using tree-to-tree distance approaches (*Daubin, Gouy & Perriere, 2002*; *Retchless & Lawrence, 2010*). Our results also suggest that the robustness of evolutionary hypotheses from whole genome data should be evaluated with phylogenetic network approaches that can depict conflicts, due to evolutionary processes or phylogenetic error, within the dataset (*Bryant & Moulton, 2002*; *Holland & Moulton, 2003*; *Huson & Bryant, 2005*). By employing such a total-evidence based approach, one would be able to recover a more realistic evolutionary hypothesis, particularly in terms of the deeper relationships, that also serves as an initial indication of the impact of organism-level evolutionary processes. Ultimately, such detailed evolutionary analyses would be invaluable for understanding the speciation process and for studying the development and distribution of important biological characteristics. Furthermore, our data also suggests that alternative approaches, focussing specifically on organism-level evolutionary processes, possibly at the population level, may be required to resolve relationships and elucidate the evolutionary history of younger taxa or leaves, where these processes may be rampant and phylogenetic incongruence highly prevalent.

CONCLUSIONS

We found that phylogenetic conflict, potentially caused by organism-level evolutionary processes, was present in our phylogenomic dataset at both the gene partition and nucleotide levels. Although this non-phylogenetic signal could result from organism-level evolutionary process, like HGT and ILS, more in-depth analyses are needed to differentiate between these processes and to quantify the overall impact of these processes on the evolutionary history of the taxa. From our results it appeared that both signal and noise are randomly distributed across the genome and that all genes included in a concatenation-based phylogenomic analysis contribute signal toward the possible species tree. In other words, for *Pantoea* at least, phylogenies constructed from concatenated datasets are not driven by single genes, but rather that the signal from individual genes work synergistically to provide robust phylogenies.

ACKNOWLEDGEMENTS

We would like to sincerely thank the reviewers, Dr. David Waite, Dr. Luis M. Rodriguez-R and an anonymous reviewer for valuable suggestions and significant contributions toward improving the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the National Research Foundation (South Africa), the DST-NRF Centre of Excellence in Tree Health Biotechnology (CTHB) and the University of Pretoria with regards to student funding. Informatics infrastructure was funded by the NRF National Bioinformatics Functional Genomics Grant (No: 93668). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
National Research Foundation (South Africa).
DST-NRF Centre of Excellence in Tree Health Biotechnology (CTHB).
University of Pretoria NRF National Bioinformatics Functional Genomics: 93668.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Marike Palmer conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Stephanus N. Venter, Emma T. Steenkamp and Martin P.A. Coetzee conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Alistair R. McTaggart performed the experiments, analyzed the data, authored or reviewed drafts of the paper, approved the final draft.
- Stephanie Van Wyk, Juanita R. Avontuur, Chrizelle W. Beukes, Gerda Fourie, Quentin C. Santana and Magriet A. Van Der Nest performed the experiments, analyzed the data, approved the final draft.
- Jochen Blom performed the experiments, contributed reagents/materials/analysis tools, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

A simple Python script was written to construct the individual phylogenies with FastTree v. 2.1 for the 1,357 protein sequences shared by all taxa analysed.

Two nexus files for the SplitsTree analyses are available as Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6698#supplemental-information.

REFERENCES

Abdul Rahman N, Parks DH, Vanwonterghem I, Morrison M, Tyson GW, Hugenholtz P. 2016. A phylogenomic analysis of the bacterial phylum fibrobacteres. *Frontiers in Microbiology* **6**:1469 DOI 10.3389/fmicb.2015.01469.

Adeolu M, Alnajar S, Naushad S, Gupta SR. 2016. 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., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. *International Journal of Systematic and Evolutionary Microbiology* **66(12)**:5575–5599 DOI 10.1099/ijsem.0.001485.

Andam CP, Gogarten JP. 2011. Biased gene transfer in microbial evolution. *Nature Reviews Microbiology* **9**:543–555 DOI 10.1038/nrmicro2593.

Bapteste E, O'Malley MA, Beiko RG, Ereshefsky M, Gogarten JP, Franklin-Hall L, Lapointe F-J, Dupré J, Dagan T, Boucher Y, Martin W. 2009. Prokaryotic evolution and the tree of life are two different things. *Biology Direct* **4**:Article 34 DOI 10.1186/1745-6150-4-34.

Beukes CW, Palmer M, Manyaka P, Chan WY, Avontuur JR, Van Zyl E, Huntemann M, Clum A, Pillay M, Palaniappan K. 2017. Genome data provides high support for generic boundaries in *Burkholderia sensu lato*. *Frontiers in Microbiology* **8**:Article 1154 DOI 10.3389/fmicb.2017.01154.

Blom J, Kreis J, Spänig S, Juhre T, Bertelli C, Ernst C, Goesmann A. 2016. EDGAR 2.0: an enhanced software platform for comparative gene content analyses. *Nucleic Acids Research* **44**:W22–W28 DOI 10.1093/nar/gkw255.

Boni MF, Posada D, Feldman MW. 2007. An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics* **176(2)**:1035–1047 DOI 10.1534/genetics.106.068874.

Brady CL, Cleenwerck I, Venter SN, Engelbeen K, De Vos P, Coutinho TA. 2010a. Emended description of the genus *Pantoea*, description of four species from human clinical samples, *Pantoea septica* sp. nov., *Pantoea eucrina* sp. nov., *Pantoea brenneri* sp. nov., and *Pantoea conspicua* sp. nov., and transfer of *Pectobacterium cypripedii* (Hori 1911) Brenner et al. 1973 emend. Hauben et al. 1998 to the genus as *Pantoea cypripedii* comb. nov. *International Journal of Systematic and Evolutionary Microbiology* **60**:2430–2440 DOI 10.1099/ijs.0.017301-0.

Brady CL, Venter SN, Cleenwerck I, Engelbeen K, Vancanneyt M, Swings J, Coutinho TA. 2009. *Pantoea vagans* sp. nov., *Pantoea eucalypti* sp. nov., *Pantoea deleyi* sp. nov., and *Pantoea anthophila* sp. nov. *International Journal of Systematic and Evolutionary Microbiology* **59**:2339–2345 DOI 10.1099/ijs.0.009241-0.

Brady CL, Venter SN, Cleenwerck I, Vandemeulebroecke K, De Vos P, Coutinho TA. 2010b. 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* **60**:484–494 DOI 10.1099/ijs.0.012070-0.

Brown SD, Sagar MU, Klingeman DM, Johnson CM, Stanton LM, Land MR, Lu T-YS, Schadt CW, Doktycz MJ, Pelletier DA. 2012. Twenty-one genome sequences from *Pseudomonas* species and 19 genome sequences from diverse bacteria isolated from the rhizosphere and endosphere of *Populus deltoides*. *Journal of Bacteriology* **194(21)**:5991–5993 DOI 10.1128/JB.01243-12.

Bryant D, Moulton V. 2002. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. In: *WABI*. New York: Springer, 375–391.

Callister SJ, McCue LA, Turse JE, Monroe ME, Auberry KJ, Smith RD, Adkins JN, Lipton MS. 2008. Comparative bacterial proteomics: analysis of the core genome concept. *PLOS ONE* **3(2)**:e1542 DOI 10.1371/journal.pone.0001542.

Chen S, Kim D-K, Chase MW, Kim J-H. 2013. Networks in a large-scale phylogenetic analysis: reconstructing evolutionary history of Asparagales (Lilianae) based on four plastid genes. *PLOS ONE* **8(3)**:e59472 DOI 10.1371/journal.pone.0059472.

Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311(5765)**:1283–1287 DOI 10.1126/science.1123061.

Coenye T, Gevers D, De Peer YV, Vandamme P, Swings J. 2005. Towards a prokaryotic genomic taxonomy. *FEMS Microbiology Reviews* **29(2)**:147–167 DOI 10.1016/j.fmrre.2004.11.004.

Cohan FM. 2001. Bacterial species and speciation. *Systematic Biology* **50(4)**:513–524 DOI 10.1080/10635150118398.

Dagan T, Martin W. 2006. The tree of one percent. *Genome Biology* **7**:Article 118 DOI 10.1186/gb-2006-7-10-118.

Daubin V, Gouy M, Perrière G. 2001. Bacterial molecular phylogeny using supertree approach. *Genome Informatics* **12**:155–164.

Daubin V, Gouy M, Perriere G. 2002. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research* **12**:1080–1090 DOI 10.1101/gr.187002.

Daubin V, Moran NA, Ochman H. 2003. Phylogenetics and the cohesion of bacterial genomes. *Science* **301(5634)**:829–832 DOI 10.1126/science.1086568.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLOS Genetics* **2(5)**:e68 DOI 10.1371/journal.pgen.0020068.

De Maayer P, Chan WY, Rubagotti E, Venter SN, Toth IK, Birch PR, Coutinho TA. 2014. Analysis of the *Pantoea ananatis* pan-genome reveals factors underlying its ability to colonize and interact with plant, insect and vertebrate hosts. *BMC Genomics* **15**:404 DOI 10.1186/1471-2164-15-404.

Doroghazi JR, Buckley DH. 2010. Widespread homologous recombination within and between *Streptomyces* species. *The Isme Journal* **4**:1136–1143 DOI 10.1038/ismej.2010.45.

Doyle JJ. 1992. Gene trees and species trees: molecular systematics as one-character taxonomy. *Systematic Botany* **17(1)**:144–163.

Dutilh BE, Huynen MA, Bruno WJ, Snel B. 2004. The consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. *Journal of Molecular Evolution* **58(5)**:527–539 DOI 10.1007/s00239-003-2575-6.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32(5)**:1792–1797 DOI 10.1093/nar/gkh340.

Eisen JA, Fraser CM. 2003. Phylogenomics: intersection of evolution and genomics. *Science* **300(5626)**:1706–1707 DOI 10.1126/science.1086292.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation. *Science* **315(5811)**:476–480 DOI 10.1126/science.1127573.

Gadagkar SR, Rosenberg MS, Kumar S. 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **304B(1)**:64–74.

Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **363(1512)**:4023–4029 DOI 10.1098/rstb.2008.0144.

Glaeser SP, Kämpfer P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. *Systematic and Applied Microbiology* **38(4)**:237–245 DOI 10.1016/j.syapm.2015.03.007.

Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, Rappé MS. 2012. Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *mBio* **3(5)**:e00252-12 DOI 10.1128/mBio.00252-12.

Gupta RS, Naushad S, Baker S. 2015. Phylogenomic analyses and molecular signatures for the class *Halobacteria* and its two major clades: a proposal for division of the class *Halobacteria* into an emended order *Halobacteriales* and two new orders, *Haloferacales* ord. nov., and *Natrialbales* ord. nov., containing the novel families *Haloferacaceae* fam. nov., and *Natrialbaceae* fam. nov. *International Journal of Systematic and Evolutionary Microbiology* **65**:1050–1069 DOI 10.1099/ijs.0.070136-0.

Hall T. 2011. BioEdit: an important software for molecular biology. *GERF Bulletin of Biosciences* **2(1)**:60–61.

Hedtke SM, Townsend TM, Hillis DM. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Systematic Biology* **55(3)**:522–529 DOI 10.1080/10635150600697358.

Hillis DM. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic Biology* **47(1)**:3–8 DOI 10.1080/106351598260987.

Holland BR, Jermiin LS, Moulton V. 2005. Improved consensus network techniques for genome-scale phylogeny. *Molecular Biology and Evolution* **23(5)**:848–855.

Holland B, Moulton V. 2003. Consensus networks: a method for visualising incompatibilities in collections of trees. In: *Algorithms in bioinformatics. WABI*. Berlin: Springer, 165–176 DOI 10.1007/978-3-540-39763-2_13.

Holmes EC, Urwin R, Maiden M. 1999. The influence of recombination on the population structure and evolution of the human pathogen Neisseria meningitidis. *Molecular Biology and Evolution* **16(6)**:741–749 DOI 10.1093/oxfordjournals.molbev.a026159.

Hong K-W, Han MG, Low S-M, Lee PKY, Chong Y-M, Yin W-F, Chan K-G. 2012. Draft genome sequence of *Pantoea* sp. strain A4, a *Rafflesia*-associated bacterium that produces N-acylhomoserine lactones as quorum-sensing molecules. *Journal of Bacteriology* **194(23)**:6610–6610 DOI 10.1128/JB.01619-12.

Huson DH, Bryant D. 2005. Estimating phylogenetic trees and networks using SplitsTree 4. *Available at http:// www.splitstree.org*.

Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends in Genetics* **22(4)**:225–231 DOI 10.1016/j.tig.2006.02.003.

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**:1005–1015 DOI 10.1099/mic.0.055459-0.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **8(3)**:275–282.

Kennedy M, Holland BR, Gray RD, Spencer HG. 2005. Untangling long branches: identifying conflicting phylogenetic signals using spectral analysis, neighbor-net, and consensus networks. *Systematic Biology* **54(4)**:620–633 DOI 10.1080/106351591007462.

Klenk H-P, Göker M. 2010. En route to a genome-based classification of Archaea and Bacteria? *Systematic and Applied Microbiology* **33(4)**:175–182 DOI 10.1016/j.syapm.2010.03.003.

Knowles LL. 2009. Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology* **58(5)**:463–467 DOI 10.1093/sysbio/syp061.

Konstantinidis KT, Tiedje JM. 2007. Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. *Current Opinion in Microbiology* **10(5)**:504–509 DOI 10.1016/j.mib.2007.08.006.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome Research* **19**:1639–1645 DOI 10.1101/gr.092759.109.

Kube M, Migdoll AM, Gehring I, Heitmann K, Mayer Y, Kuhl H, Knaust F, Geider K, Reinhardt R. 2010. Genome comparison of the epiphytic bacteria *Erwinia billingiae* and *E. tasmaniensis* with the pear pathogen *E. pyrifoliae*. *BMC Genomics* **11**:393 DOI 10.1186/1471-2164-11-393.

Kube M, Migdoll AM, Müller I, Kuhl H, Beck A, Reinhardt R, Geider K. 2008. The genome of *Erwinia tasmaniensis* strain Et1/99, a non-pathogenic bacterium in the genus *Erwinia*. *Environmental Microbiology* **10(9)**:2211–2222 DOI 10.1111/j.1462-2920.2008.01639.x.

Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. *Frontiers in Zoology* **11**:Article 81 DOI 10.1186/s12983-014-0081-x.

Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2011. Statistics and truth in phylogenomics. *Molecular Biology and Evolution* **29(2)**:457–472.

Lim J-A, Lee DH, Kim B-Y, Heu S. 2014. Draft genome sequence of *Pantoea agglomerans* R190, a producer of antibiotics against phytopathogens and foodborne pathogens. *Journal of Biotechnology* **188**:7–8 DOI 10.1016/j.jbiotec.2014.07.440.

Ma Y, Yin Y, Rong C, Chen S, Liu Y, Wang S, Xu F. 2016. *Pantoea pleuroti* sp. nov., isolated from the fruiting bodies of *Pleurotus eryngii*. *Current Microbiology* **72(2)**:207–212 DOI 10.1007/s00284-015-0940-5.

Maddison WP. 1997. Gene trees in species trees. *Systematic Biology* **46(3)**:523–536 DOI 10.1093/sysbio/46.3.523.

Mallet J, Besansky N, Hahn MW. 2016. How reticulated are species? *Bioessays* **38(2)**:140–149 DOI 10.1002/bies.201500149.

Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. 2015. RDP4: detection and analysis of recombination patterns in virus genomes. *Virus Evolution* **1(1)**:vev003 DOI 10.1093/ve/vev003.

Martin D, Rybicki E. 2000. RDP: detection of recombination amongst aligned sequences. *Bioinformatics* **16(6)**:562–563 DOI 10.1093/bioinformatics/16.6.562.

Meehan CJ, Beiko RG. 2014. A phylogenomic view of ecological specialization in the *Lachnospiraceae*, a family of digestive tract-associated bacteria. *Genome Biology and Evolution* **6(3)**:703–713 DOI 10.1093/gbe/evu050.

Meng C, Kubatko LS. 2009. Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. *Theoretical Population Biology* **75(1)**:35–45 DOI 10.1016/j.tpb.2008.10.004.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* **30(17)**:i541–i548 DOI 10.1093/bioinformatics/btu462.

Nosil P. 2008. Speciation with gene flow could be common. *Molecular Ecology* **17(9)**:2103–2106 DOI 10.1111/j.1365-294X.2008.03715.x.

Padidam M, Sawyer S, Fauquet CM. 1999. Possible emergence of new geminiviruses by frequent recombination. *Virology* **265(2)**:218–225 DOI 10.1006/viro.1999.0056.

Palmer M, De Maayer P, Poulsen M, Steenkamp ET, Van Zyl E, Coutinho TA, Venter SN. 2016. Draft genome sequences of *Pantoea agglomerans* and *Pantoea vagans* isolates associated with termites. *Standards in Genomic Sciences* **11**:Article 23 DOI 10.1186/s40793-016-0144-z.

Palmer M, Steenkamp ET, Coetzee MPA, Blom J, Venter SN. 2018. Genome-based characterization of biological processes that differentiate closely related bacteria. *Frontiers in Microbiology* **9**:Article 113 DOI 10.3389/fmicb.2018.00113.

Palmer M, Steenkamp ET, Coetzee MPA, Chan W-Y, Van Zyl E, De Maayer P, Coutinho TA, Blom J, Smits THM, Duffy B, Venter SN. 2017. Phylogenomic resolution of the bacterial genus Pantoea and its relationship with Erwinia and Tatumella. *Antonie van Leeuwenhoek* **110(10)**:1287–1309 DOI 10.1007/s10482-017-0852-4.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5(5)**:568–583.

Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLOS Biology* **9(3)**:e1000602 DOI 10.1371/journal.pbio.1000602.

Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. 2002. Increased taxon sampling is advantageous for phylogenetic inference. *Systematic Biology* **51(4)**:664–671 DOI 10.1080/10635150290102357.

Posada D, Crandall KA. 2001. Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proceedings of the National Academy of Sciences of the United States of America* **98(24)**:13757–13762 DOI 10.1073/pnas.241370698.

Posada D, Crandall KA. 2002. The effect of recombination on the accuracy of phylogeny estimation. *Journal of Molecular Evolution* **54(3)**:396–402 DOI 10.1007/s00239-001-0034-9.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5(3)**:e9490 DOI 10.1371/journal.pone.0009490.

Qin Q-L, Xie B-B, Zhang X-Y, Chen X-L, Zhou B-C, Zhou J, Oren A, Zhang Y-Z. 2014. A proposed genus boundary for the prokaryotes based on genomic insights. *Journal of Bacteriology* **196(12)**:2210–2215 DOI 10.1128/JB.01688-14.

Ren F, Tanaka H, Yang Z. 2009. A likelihood look at the supermatrix—supertree controversy. *Gene* **441(1–2)**:119–125 DOI 10.1016/j.gene.2008.04.002.

Retchless AC, Lawrence JG. 2010. Phylogenetic incongruence arising from fragmented speciation in enteric bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **107(25)**:11453–11458 DOI 10.1073/pnas.1001291107.

Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences of the United States of America* **106(45)**:19126–19131 DOI 10.1073/pnas.0906412106.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**:798–804 DOI 10.1038/nature02053.

Rosenberg MS, Kumar S. 2003. Taxon sampling, bioinformatics, and phylogenomics. *Systematic Biology* **52(1)**:119–124 DOI 10.1080/10635150390132894.

Rosenberg NA. 2002. The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology* **61(2)**:225–247 DOI 10.1006/tpbi.2001.1568.

Rosselló-Mora R. 2005. Updating prokaryotic taxonomy. *Journal of Bacteriology* **187(18)**:6255–6257 DOI 10.1128/JB.187.18.6255-6257.2005.

Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**:327–331 DOI 10.1038/nature12130.

Sanderson MJ, Driskell AC. 2003. The challenge of constructing large phylogenetic trees. *Trends in Plant Science* **8(8)**:374–379 DOI 10.1016/S1360-1385(03)00165-1.

Sarkar SF, Guttman DS. 2004. Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology* **70(4)**:1999–2012 DOI 10.1128/aem.70.4.1999-2012.2004.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Molecular Biology and Evolution* **33(7)**:1654–1668 DOI 10.1093/molbev/msw079.

Schwartz AR, Potnis N, Timilsina S, Wilson M, Patané J, Martins J, Minsavage GV, Dahlbeck D, Akhunova A, Almeida N, Vallad GE, Barak JD, White FF, Miller SA, Ritchie D, Goss E, Bart RS, Setubal JC, Jones JB, Staskawicz BJ. 2015. Phylogenomics of *Xanthomonas* field strains infecting pepper and tomato reveals diversity in effector repertoires and identifies determinants of host specificity. *Frontiers in Microbiology* **6**:Article 535 DOI 10.3389/fmicb.2015.00535.

Shen X-X, Hittinger CT, Rokas A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nature Ecology & Evolution* **1**:Article 0126 DOI 10.1038/s41559-017-0126.

Smith JM. 1992. Analyzing the mosaic structure of genes. *Journal of Molecular Evolution* **34(2)**:126–129 DOI 10.1007/BF00182389.

Smits TH, Rezzonico F, Kamber T, Goesmann A, Ishimaru CA, Stockwell VO, Frey JE, Duffy B. 2010. Genome sequence of the biocontrol agent *Pantoea vagans* strain C9-1. *Journal of Bacteriology* **192**:6486–6487 DOI 10.1128/JB.01122-10.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30(9)**:1312–1313 DOI 10.1093/bioinformatics/btu033.

Suen G, Scott JJ, Aylward FO, Adams SM, Tringe SG, Pinto-Tomàs A, Foster CE, Pauly M, Weimer PJ, Barry KW, Goodwin LA, Bouffard P, Li L, Osterberger J, Harkins TT, Slater SC, Donohue TJ, Currie CR. 2010. An insect herbivore microbiome with high plant biomass-degrading capacity. *PLOS Genetics* **6(9)**:e1001129 DOI 10.1371/journal.pgen.1001129.

Szöllősi GJ, Boussau B, Abby SS, Tannier E, Daubin V. 2012. Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proceedings of the National Academy of Sciences of the United States of America* **109(43)**:17513–17518 DOI 10.1073/pnas.1202997109.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30(12)**:2725–2729 DOI 10.1093/molbev/mst197.

Thiergart T, Landan G, Martin WF. 2014. Concatenated alignments and the case of the disappearing tree. *BMC Evolutionary Biology* **14**:266 DOI 10.1186/s12862-014-0266-0.

Tracz DM, Gilmour MW, Mabon P, Beniac DR, Hoang L, Kibsey P, Van Domselaar G, Tabor H, Westmacott GR, Corbett CR, Bernard KA. 2015. *Tatumella saanichensis*

sp. nov., isolated from a cystic fibrosis patient. *International Journal of Systematic and Evolutionary Microbiology* **65(Pt 6)**:1959–1966 DOI 10.1099/ijs.0.000207.

Walterson AM, Stavrinides J. 2015. Pantoea: insights into a highly versatile and diverse genus within the Enterobacteriaceae. *FEMS Microbiology Reviews* **39(6)**:968–984 DOI 10.1093/femsre/fuv027.

Wan X, Hou S, Phan N, Moss JS, Donachie SP, Alam M. 2015. Draft genome sequence of *Pantoea anthophila* strain 11-2 from hypersaline Lake Laysan, Hawaii. *Genome Announcements* **3(3)**:e00321-15 DOI 10.1128/genomeA.00321-15.

Wendel JF, Doyle JJ. 1998. *Phylogenetic incongruence: window into genome history and molecular evolution. Molecular systematics of plants II*. Boston: Springer, 265–296 DOI 10.1007/978-1-4615-5419-6_10.

Williams TA, Szöllősi GJ, Spang A, Foster PG, Heaps SE, Boussau B, Ettema TJ, Embley TM. 2017. Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences of the United States of America* **114(23)**:E4602–E4611 DOI 10.1073/pnas.1618463114.

Yokono M, Satoh S, Tanaka A. 2018. Comparative analyses of whole-genome protein sequences from multiple organisms. *Scientific Reports* **8**:6800 DOI 10.1038/s41598-018-25090-8.

Zhang Z-G, Ye Z-Q, Yu L, Shi P. 2011. Phylogenomic reconstruction of lactic acid bacteria: an update. *BMC Evolutionary Biology* **11**:1 DOI 10.1186/1471-2148-11-1.

Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Systematic Biology* **51(4)**:588–598 DOI 10.1080/10635150290102339.