Antonie van Leeuwenhoek (2017) 110:1287–1309
DOI 10.1007/s10482-017-0852-4

ORIGINAL PAPER

Phylogenomic resolution of the bacterial genus *Pantoea* and its relationship with *Erwinia* and *Tatumella*

Marike Palmer · Emma T. Steenkamp · Martin P. A. Coetzee · Wai-Yin Chan · Elritha van Zyl · Pieter De Maayer · Teresa A. Coutinho · Jochen Blom · Theo H. M. Smits · Brion Duffy · Stephanus N. Venter

Received: 25 November 2016 / Accepted: 23 February 2017 / Published online: 2 March 2017

Abstract Investigation of the evolutionary relationships between related bacterial species and genera with a variety of lifestyles have gained popularity in recent years. For analysing the evolution of specific traits, however, a robust phylogeny is essential. In this study we examined the evolutionary relationships among the closely related genera *Erwinia*, *Tatumella* and *Pantoea*, and also attempted to resolve the species relationships within *Pantoea*. To accomplish this, we used the whole genome sequence data for 35 different strains belonging to these three genera, as well as nine outgroup taxa. Multigene datasets consisting of the 1039 genes shared by these 44 strains were then generated and subjected to maximum likelihood phylogenetic analyses, after which the results were compared to those using conventional multi-locus sequence analysis (MLSA) and ribosomal MLSA (rMLSA) approaches. The robustness of the respective phylogenies was then explored by considering the factors typically responsible for destabilizing phylogenetic trees. We found that the nucleotide datasets employed in the MLSA, rMLSA and 1039-gene datasets contained significant levels of homoplasy, substitution saturation and differential codon usage, all of which likely gave rise to the observed lineage specific rate heterogeneity. The effects of these factors were much less pronounced in the amino acid dataset

Electronic supplementary material The online version of this article (doi:10.1007/s10482-017-0852-4) contains supplementary material, which is available to authorized users.

M. Palmer · E. T. Steenkamp · W.-Y. Chan · E. van Zyl · T. A. Coutinho · S. N. Venter (✉)
Department of Microbiology and Plant Pathology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa
e-mail: fanus.venter@up.ac.za

M. P. A. Coetzee
Department of Genetics, Forestry and Agricultural Biotechnology Institute (FABI),, University of Pretoria, Pretoria, South Africa

P. De Maayer
School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa

J. Blom
Computational Genomics, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany

T. H. M. Smits · B. Duffy
Environmental Genomics and Systems Biology Research Group, Institute of Natural Resource Sciences, Zürich University of Applied Sciences (ZHAW), Wädenswil, Switzerland

for the 1039 genes, which allowed reconstruction of a fully supported and resolved phylogeny. The robustness of this amino acid tree was also supported by different subsets of the 1039 genes. In contrast to the smaller datasets (MLSA and rMLSA), the 1039 amino acid tree was also not as sensitive to long-branch attraction. The robust and well-supported evolutionary hypothesis for the three genera, which confidently resolved their various inter- and intrageneric relationships, represents a valuable resource for future studies. It will form the basis for studies aiming to understand the forces driving the divergence and maintenance of lineages, species and biological traits in this important group of bacteria.

Keywords Phylogenetics · Non-phylogenetic signal · MLSA · Core genome · *Enterobacteriaceae*

Introduction

In recent years, a number of studies have investigated the evolution of different lifestyles, pathogenicity features and survival strategies within or between bacterial genera based on genomic data (Bennett et al. 2012; Prasanna and Mehra 2013; Angus et al. 2014; De Maayer et al. 2014; Fouts et al. 2016). The main focus of these studies has often been on understanding how the species or groups of species have evolved and what promoted their biological differentiation. In such a study, an essential first step is to obtain a robust phylogeny for resolving relationships and inferring evolutionary histories (Bennett et al. 2012; Prasanna and Mehra 42013). These phylogenies are then used for studying the emergence and development of biological traits and for determining the possible causes of divergence within and between genera.

The phylogenetic tree that depicts the relationships among the species of a genus is typically referred to as a "species tree" (Klenk and Göker 2010; Andam and Gogarten 2011). For bacteria, species trees have been traditionally inferred using the sequence information from housekeeping genes. These include phylogenetic analysis of 16S ribosomal RNA (rRNA) sequences (Konstantinidis and Tiedje 2007), multi-locus sequence analysis (MLSA) with 4–7 housekeeping genes (Gevers et al. 2005; Konstantinidis and Tiedje 2007; Glaeser and Kämpfer 2015), and more recently, ribosomal MLSA (rMLSA) based on 53 structural ribosomal proteins (Bennett et al. 2012; Jolley et al. 2012). However, the phylogenies generated with these data are often not particularly robust (Konstantinidis and Tiedje 2007; Brady et al. 2008; Glaeser and Kämpfer 2015). This is primarily due to a lack of phylogenetic signal in the highly conserved gene datasets used (Fox et al. 1992; Gevers et al. 2005; Konstantinidis and Tiedje 2005, 2007; Staley 2006; Richter and Rosselló-Móra 2009). Other issues that may also detract from the overall stability of a species tree pertains to the use of paralogues (i.e., homologues originating from an intragenomic duplication event) or xenologues (i.e., homologues originating from horizontal gene transfer) (Koonin 2005). In fact, a number of the markers commonly used in bacterial systematics have been shown to be present in multiple copies in the genomes of certain taxa (Boucher et al. 2004; Conville and Witebsky 2007) or are even found on plasmid elements (Anda et al. 2015). Some of these genes have also been shown to be acquired horizontally (Rivera et al. 1998; Boucher et al. 2004).

The availability of whole genome sequence (WGS) information has revolutionised the fields of evolutionary biology and bacterial systematics. Despite the fact that horizontal gene transfer (HGT) significantly impacts the evolution of most, if not all, bacterial groups (Woese 2000; Gogarten et al. 2002; Jain et al. 2002; Boto 2010; Cohen et al. 2011), it is now possible to infer trees that trace the shared ancestry among all the species of a genus using WGS information. Here the assumption is that the dominant phylogenetic signal in the genome of an individual is reflective of its parental lineage and that this would "overshadow" the signals associated with HGT (Andam and Gogarten 2011). As a result, the overall evolution of the genus under examination will likely be depicted in the form of a bifurcating tree. For example, robust species trees have been inferred using this approach for *Acinetobacter* (Chan et al. 2012) and *Neisseria* (Bennett et al. 2012).

The WGS-based approach for building species trees involves the use of all (or a large number) of the gene sequences common to the members of the focal genus and its outgroups. This approach is currently regarded as the most reliable approach for inferring species trees (Chan et al. 2012; Lang et al. 2013) because it takes advantage of all of the phylogenetically informative characters included in the genomes of the taxa under investigation (Chan et al. 2012). WGS-based datasets are, therefore, large and their use for inferring species trees outperforms those

consisting of single gene or small sets of housekeeping gene sequences (Daubin et al. 2002; Coenye et al. 2005; Galtier and Daubin 2008; Bennett et al. 2012; Chan et al. 2012). Compared to smaller datasets, the phylogenetic signal associated with vertical descent in WGS-based datasets far outweighs the noise (Andam and Gogarten 2011; Chan et al. 2012; Lang et al. 2013). In other words, even if paralogues, xenologues or highly conserved sequences are mistakenly included in the WGS-based dataset, the phylogenetic signal associated with their aberrant evolutionary histories will be diluted by the total signal of vertical descent embedded in these large datasets. This is not the case for the smaller datasets that are conventionally used for inferring bacterial species trees (Rivera et al. 1998; Boucher et al. 2004).

Another benefit of using the shared gene content for inferring species trees is that most of the sequences included in the dataset form part of the so-called core genomes (Daubin et al. 2002; Coenye et al. 2005) of the taxa under investigation. The core genome consists of the genetic material common to the taxon and includes those genes present in nearly all of its members (Lan and Reeves 2000; Coenye et al. 2005). Accordingly, the core genome usually represents only a small subset of the taxon's pan genome (Makarova et al. 2006; Lukjancenko et al. 2012) and its genes are considered to be essential for survival and often encode products involved in crucial cellular processes (Hacker et al. 2012). The latter, combined with the mainly vertical inheritance of the core genome component (Daubin et al. 2002; Coenye et al. 2005), therefore, highlights the value of using core gene datasets for studying evolutionary trajectories that have shaped the biology and ecology of the taxa under investigation (Daubin et al. 2002; Coenye et al. 2005).

In this study we were interested in reconstructing the species tree for the genus *Pantoea*. This genus currently comprises 23 species and subspecies (Gavini et al. 1989b; Mergaert et al. 1993; Brady et al. 2007, 2008, 2009, 2010a, 2011, 2012; Popp et al. 2010; Gueule et al. 2015; Prakash et al. 2015; Tanaka et al. 2015), with a further two species (*P. pleuroti* and *P. hericii*) recently described but not yet validated (Ma et al. 2016; Rong et al. 2016). Members of this taxon exhibit a diverse range of phenotypic characteristics, especially in terms of physiological attributes and niche occupation (Brady et al. 2008, Walterson and Stavrinides 2015). For example, *Pantoea* includes various human and plant pathogens (De Baere et al. 2004; Cruz et al. 2007; Brady et al. 2010a), as well as species with plant growth promoting abilities (Smits et al. 2011; Kim et al. 2012), and species associated with insects (Palmer et al. 2016) and fungi (Ma et al. 2016, Rong et al. 2016) to name but a few. Although *Pantoea* is usually recovered as a monophyletic group in phylogenetic trees, interspecific relationships are not well resolved (Rezzonico et al. 2009; Brady et al. 2012; Tambong et al. 2014; Gueule et al. 2015). Furthermore, the overall position of *Pantoea* within the *Enterobacteriaceae* has not been conclusively established. The genus *Tatumella* is commonly regarded as its sister taxon (Brady et al. 2008, 2010a, 2012), although other intergeneric relationships have also been reported (Brady et al. 2008; 2010b, 2012; Kamber et al. 2012; Smits et al. 2013; Glaeser and Kämpfer 2015; Gueule et al. 2015). We hypothesize that these inconsistent inter- and intrageneric relationships are mainly due to the small datasets often being used for phylogenetic inference (Brady et al. 2008, 2010b; Glaeser and Kämpfer 2015). Another contributing factor pertains to incomplete taxon selection where datasets often exclude one or more of the relevant taxa from analyses (Naum et al. 2008; Tambong et al. 2014; Zhang and Qiu 2015).

The overall goal of this study was therefore to use a WGS-based approach to determine the generic relationships of *Pantoea* within the *Enterobacteriaceae* and then to infer a species tree for *Pantoea.* To achieve these goals, our aims were four-fold. Firstly, to allow for meaningful inter- and intrageneric comparisons, the WGSs of twelve *Pantoea* species were determined, which complemented those of fourteen strains already available in the public domain (Table 1). Secondly, a maximum likelihood phylogeny depicting the relationships among *Pantoea* species, as well as among *Pantoea* and other genera, were inferred using the aligned shared gene sequences extracted from the WGS data. Thirdly, the robustness of this tree was evaluated by considering the various factors known to negatively affect phylogenetic analyses (Xia and Xie 2001; Zwickl and Hillis 2002; Jeffroy et al. 2006; Heath et al. 2008; Philippe et al. 2011). Finally, to determine the possible causes for the incongruent intra- and intergeneric relationships previously reported for *Pantoea* and its relatives (Brady et al. 2010a; Glaeser and Kämpfer 2015; Gueule et al. 2015), we evaluated the conventional methods (i.e.,

Table 1 Isolates with available genome sequences and those determined in this study*

Species	Strain	Origin	Accession number
Erwinia amylovora	LA 636	Apple, Mexico	CBVT00000000.1
Erwinia billingiae	NCPPB 661 T	Pear, UK	FP236843.1, FP236826.1, FP236830.1
Erwinia mallotivora	BT-MARDI	Papaya, Malaysia	JFHN00000000.1
Erwinia pyrifoliae	DSM 12163 T	Asian pear, Korea	FN392235.1, FN392236.1, FN392237.1
Erwinia tasmaniensis	Et 1-99 T	Apple flowers, Australia	CU468135.1, CU468128.1, CU468130.1, CU468131.1, CU468132.1, CU468133.1
Erwinia toletana	DAPP-PG 735	Olive knot, Italy	AOCZ00000000.1
Erwinia tracheiphila	PSU-1	Wild gourd, USA	APJK00000000.1
Erwinia sp.	9145	Information missing	JQNE00000000.1
Erwinia sp.	Ejp 617	Asian pear, Japan	CP002124.1, CP002125.1, CP002126.1
Pantoea agglomerans	R 190	Apple, Korea	JNGC00000000.1
*Pantoea allii**	LMG 24248 T	Onion seed, South Africa	MLFE00000000.1
Pantoea ananatis	LMG 2665 T	Pineapple, Brazil	JFZU00000000.1
Pantoea anthophila	11-2	Hypersaline lake, Hawaii	JXXL00000000.1
*Pantoea brenneri**	LMG 5343 T	Human, USA	MIEI00000000.1
*Pantoea calida**	LMG 25383 T	Infant formula, –	MLFO00000000.1
*Pantoea conspicua**	LMG 24534 T	Human, France	MLFN00000000.1
*Pantoea cypripedii**	LMG 2657 T	Orchid, USA	MLJI00000000.1
*Pantoea deleyi**	LMG 24200 T	Eucalyptus, Uganda	MIPO00000000.1
Pantoea dispersa	EGD-AAK13	Soil, India	AVSS00000000.1
Pantoea eucalypti	aB	Bark beetle, USA	AEDL00000000.1
*Pantoea eucrina**	LMG 2781 T	Human, USA	MIPP00000000.1
*Pantoea gaviniae**	LMG 25382 T	Infant formula, –	MLFQ00000000.1
*Pantoea rodasii**	LMG 26273T	Eucalyptus, Colombia	MLFP00000000.1
*Pantoea rwandensis**	LMG 26275 T	Eucalyptus, Rwanda	MLFR00000000.1
*Pantoea septica**	LMG 5345 T	Human, USA	MLJJ00000000.1
Pantoea stewartii ssp. *stewartii*	DC 283	Maize, USA	AHIE00000000.1
Pantoea stewartii ssp. *indologenes*	LMG 2632 T	Fox millet, India	JPKO00000000.1
Pantoea vagans	C9-1	Apple, USA	CP001894.1, CP001893.1, CP001894.1
*Pantoea wallisii**	LMG 26277 T	Eucalyptus, South Africa	MLFS00000000.1
Pantoea sp.	At-9b	Leaf cutter ant, USA	CP002433.1, CP002434.1, CP002435.1, CP002436.1, CP002437.1, CP002438.1
Pantoea sp.	A4	Rafflesia flower, Malaysia	ALXE00000000.1
Pantoea sp.	IMH	Soil, Mongolia	JFGT00000000.1
Pantoea sp.	GM01	Poplar, USA	AKUI00000000.1
Pantoea sp.	PSNIH1	Shelf, USA	CP009880.2, CP009881.1, CP010325.1, CP0009882.1, CP010326.1, CP009883.1, CP009884.1
Pantoea sp.	PSNIH2	Hand rail, USA	CP009866.1, CP009867.1, CP009868.1, CP009869.1, CP009870.1, CP009871.1
Tatumella morbirosei	LMG 23360 T	Pineapple, Philippines	CM003276.1

Table 1 continued

Species	Strain	Origin	Accession number
Tatumella ptyseos	ATCC 33301 T	Human, USA	ATMJ00000000.1
Tatumella saanichensis	NML 06-3099 T	Human, Canada	ATMI00000000.1
Tatumella sp.	UCD-D suzukii	Fruit fly, USA	JFJX00000000.1
Brenneria goodwinii	OBR 1	Information missing	CGIG00000000.1
Cronobacter sakazakii	ATCC 29544 T	Human, USA	CP011047.1, CP011048.1, CP011049.1, CP011050.1
Enterobacter cloacae spp. *cloacae*	ATCC 13047 T	Human, USA	CP001918.1, CP001919.1, CP001920.1
Franconibacter helveticus	LMG 23732 T	Fruit powder, Switzerland	AWFX00000000.1
Klebsiella pneumoniae ssp. *pneumoniae*	ATCC 13883 T	Human, –	JOOW00000000.1
Kluyvera ascorbata	ATCC 33433 T	Human, USA	JMPL00000000.1
Pectobacterium carotovorum ssp. *carotovorum*	NCPPB 312 T	Potato, Denmark	JQHJ00000000.1
Serratia marcescens ssp. marcescens	ATCC 13880 T	Pond water, –	JMPQ00000000.1
Yokenella regensburgei	ATCC 49455 T	Insect gut, Germany	JMPS00000000.1

MLSA and rMLSA) for investigating relatedness amongst taxa by making use of datasets containing representatives of all relevant genera. A robust *Pantoea* species tree will form an essential foundation for future studies focusing on the evolution of characteristics and traits related to the different survival strategies within the genus. This study will also provide the basis for taxonomic clarity in terms of available genome data and the phylogenetic position of *Pantoea* relative to its sister genera within the *Enterobacteriaceae*.

Materials and methods

Genome sequencing of twelve *Pantoea* species

The genome sequences of twelve *Pantoea* species (*P. allii*, *P. brenneri*, *P. calida*, *P. conspicua*, *P. cypripedii, P. deleyi*, *P. eucrina*, *P. gaviniae*, *P. rodasii*, *P. rwandensis*, *P. septica* and *P. wallisii*) (Table 1, Supplementary Table S1) were determined in this study. For this purpose, the type strains of these twelve species were grown on nutrient agar for 48 h at 28 °C. High quality DNA was extracted using a CTAB method (Cleenwerck et al. 2002). The genomic DNA was then subjected to whole genome shotgun sequencing using the Ion Torrent™ Personal Genome Machine® (PGM) System (ThermoFisher Scientific) at the University of Pretoria Sequencing Facility or the Roche 454 GS-Junior sequencer at Agroscope Research Station in Wädenswil, Switzerland. The raw sequence reads were trimmed and filtered using FASTX Tools (Gordon and Hannon 2010), where those with sequence quality scores < 20 were discarded. The trimmed and filtered data were assembled with the Roche Newbler 2.6 or 2.7 programs (Margulies et al. 2005).

Taxon selection

The taxa included in our WGS-based datasets were chosen to span the known diversity of the genus *Pantoea* (hence the generation of additional WGS data

here). We also endeavoured to utilize a wide selection of species (with available WGSs) within each of *Erwinia* and *Tatumella*, which are known to be closely related to *Pantoea* (Brady et al. 2008; Brady et al. 2010a; Glaeser and Kämpfer 2015). These included formally described species as well as potentially novel species, based on average nucleotide identity (ANI) values (Gevers et al. 2005; Konstantinidis and Tiedje 2005, Richter and Rosselló-Móra 2009). This was done by obtaining all the relevant WGSs from the National Centre for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/; accessed 7/5/2015) and then subjecting them to ANI analyses in JSpecies 1.2.1 (Richter and Rosselló-Móra 2009). These analyses involved pair-wise comparisons of shared regions of the genomes to obtain a similarity value across the genome (Gevers et al. 2005; Konstantinidis and Tiedje 2005; Richter and Rosselló-Móra 2009). Where multiple genome sequences for a species were available, the WGS data of its type strain or a suitable conspecific isolate (based on similarity of available housekeeping genes) was used.

Identification of shared genes and construction of datasets

Sets of shared genes were determined with the EDGAR (Efficient Database framework for comparative Genome Analyses using BLAST score Ratios) server (https://edgar.computational.bio.uni-giessen.de) (Blom et al. 2016). The combined fasta files obtained from EDGAR were split into individual gene files from which five multigene datasets were constructed. The *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset consisted of the genes shared among all of the species used in this study (Table 1), while the *Erwinia* + *Pantoea* + *Tatumella* dataset included those shared among the examined species of the three genera. For the nucleotide substitution and codon bias analyses (see below) a third smaller multigene dataset, *Erwinia* + *Pantoea* + *Tatumella*_reduced was constructed which included 11 taxa, specifically selected to represent the diversity within *Erwinia*, *Pantoea* and *Tatumella*. The last two multigene datasets were the conventional MLSA dataset consisting of four genes (*atpD*, *gyrB*, *infB* and *rpoB*) previously used to investigate relationships in these genera (Brady et al. 2008; Glaeser and Kämpfer 2015), and the rMLSA dataset that consists of 52 of the known 53 genes encoding the structural ribosomal proteins (Bennett et al. 2012; Jolley et al. 2012).

For the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset, five subsets were constructed by grouping the genes included in this dataset into broad functional groupings. This was done by subjecting the genes to functional annotation using the Rapid Annotation using Subsystem Technology (RAST) server (Aziz et al. 2008). Five subsets ('Cellular functioning', 'Metabolism', 'Informational', 'External factors' and 'Unclassified') containing the amino acid sequences were generated based on the subsystem classification of the genes.

For the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset, three subsets were constructed based on the type of selection experienced by the genes. For this purpose, individual gene alignments (see below) were subjected to selection analysis using HyPhy (Pond and Muse 2005) as implemented in MEGA 6.06 (Tamura et al. 2013). Gene-wide dN/dS values were determined for each individual gene. These values were then plotted as a line graph in Microsoft Excel 2013. Of the 1039 shared genes, those with a dN/dS below 1 were regarded as being under purifying selection, while those with dN/dS higher than 1 were considered as being under diversifying selection. Genes with dN/dS values ranging from 0.9 to 1.1 were viewed as potentially experiencing neutral or nearly neutral evolution. Three datasets containing the amino acids sequences of the genes under different selection pressure were thus constructed and referred to as 'Purifying' ($dN/dS < 1$), 'Diversifying' ($dN/dS > 1$) and 'Neutral' ($0.9 < dN/dS > 1.1$).

Sequence alignments

Except for the *Erwinia* + *Pantoea* + *Tatumella*_reduced dataset, which required codon-based alignment, all datasets were treated as follows. Individual gene files for all datasets were batch-aligned using MUSCLE (Edgar 2004) as part of the CLC Main Workbench 7.6 package (CLC Bio). The alignments were then subjected to GBLOCKS 0.91 b (Castresana 2000) to discard any parts of alignments with missing data. For all multigene datasets, the relevant aligned gene sequences were concatenated and partitioned using FASconCAT-G v. 1.02 (Kuck and Longo 2014). In all cases, amino acid alignments were generated in addition to the nucleotide datasets. Amino acid

datasets were partitioned with the appropriate amino acid model determined by ProtTest 3.4 (Abascal et al. 2005) as implemented in FASconCAT-G. All nucleotide multigene datasets were also concatenated with the third codon positions excluded from the datasets.

Both the nucleotide and amino acid sequences for the *Erwinia* + *Pantoea* + *Tatumella*_reduced datasets were treated in the same manner. The *Erwinia* + *Pantoea* + *Tatumella*_reduced dataset was batch-aligned with MUSCLE. We then manually curated the individual nucleotide gene files in BioEdit (Hall 2011) to ensure that all gene alignments were in the correct reading frame, as well as to discard any regions with a large amount of missing data. These aligned sequences were then concatenated with FASconCAT-G to obtain a supermatrix for both the nucleotide and amino acid sequences, as well as a data matrix with the third codon positions excluded from the nucleotide datasets.

Phylogenetic analyses

For phylogenetic analysis of the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups, *Erwinia* + *Pantoea* + *Tatumella*, MLSA and rMLSA amino acid and nucleotide datasets, RAxML 8.2.1 (Stamatakis 2014) was used to construct maximum likelihood (ML) trees. Suitable partitioning files for use in this software were produced by FASconCAT-G (Kuck and Longo 2014). For the amino acid dataset, each gene utilized the best-fit substitution model as indicated by ProtTest 3.4 (Abascal et al. 2005) with independent model parameters. For the nucleotide dataset, each gene utilized the General Time Reversible (GTR) model of substitution (Tavaré 1986) with independent model parameters. In the analyses, parameters for the GTR model were independently estimated and optimized for each of the respective gene sequences. Thus, the appropriate substitution model (based on the substitution rates and α shape parameter) was inferred for each gene in the dataset, which allowed for ML analyses to be conducted with the model parameters that best fit each gene. Because of computational demands, RAxML was only used to obtain trees with the best likelihood, and branch support was estimated separately. This involved approximate likelihood analyses of the unpartitioned datasets using FastTree 2.1.8 (Price et al. 2010) from which non-parametric, Shimodaira-Hasegawa-like branch support values (Guindon et al. 2010) were estimated. We also used Seqboot (Felsenstein 2005) to construct 1000 bootstrap replicate data matrices for the datasets, which were then analysed with FastTree, from which bootstrap support values were estimated using the publicly available perl script CompareToBootstrap.pl (http://www.microbesonline.org/fasttree/treecmp.html).

Analysis of the five functional data subsets (i.e., 'Cellular functioning', 'Metabolism', 'Informational', 'External factors' and 'Unclassified'), as well as the selection datasets (i.e., 'Purifying', 'Diversifying' and 'Neutral') were performed with FastTree 2.1.8 (Price et al. 2010) to obtain approximate likelihood phylograms. Non-parametric, Shimodaira- Hasegawa-like branch support values (Guindon et al. 2010), as well as bootstrap support obtained using Seqboot (Felsenstein 2005) and CompareToBootstrap.pl (Price et al. 2010), were also estimated for the topologies obtained.

Homoplasy index

The possible impact of homoplasy (convergent mutations or similarities among taxa that are not due to common ancestry and that can affect tree reconstruction) (Philippe et al. 2011; West-Eberhard 2003) on the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups, MLSA and rMLSA datasets (both amino acid and nucleotide data in all three cases and, in the case of the nucleotide datasets, both with and without the third codon positions), was estimated. This was done by calculating the homoplasy index (HI) for each dataset using PAUP* 4.0 (Swofford 2002). The HI was determined for all parsimony informative sites by making use of the amino acid-based ML topology obtained for the 1039 shared genes.

Nucleotide substitution saturation analysis

Detailed nucleotide substitution patterns were determined for the *Erwinia* + *Pantoea* + *Tatumella*_reduced dataset and the MLSA dataset. This was done by correlating the actual substitutions in the dataset with those inferred under an appropriate model of nucleotide substitution (Jeffroy et al. 2006; Philippe et al. 2011). For this purpose, we used pair-wise uncorrected p-distances (i.e., the proportion, p, of nucleotide sites at which the two sequences being compared are different) and pair-wise nucleotide-based distances under the General Time Reversible (GTR) model

(Tavaré 1986) with the minimum evolution distance algorithm (Desper and Gascuel 2002) for the nucleotide sequences. These two estimates were both determined in DAMBE 6.0.1 (Xia and Xie 2001) and were calculated for the first, second and third codon positions. The same was done for the amino acid *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset by using pair-wise amino acid-based distances under the Jones-Taylor-Thornton (JTT) model (Jones et al. 1992) using MEGA 6.0.6 (Tamura et al. 2013). Microsoft Excel 2013 was then used to graphically plot the respective distances and to perform linear regression analyses for determining the slope of the regression line fitting the data.

Codon usage bias

The relative synonymous codon usage (RSCU) for *Erwinia*, *Pantoea* and *Tatumella* was determined from the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset using DAMBE 6.0.1 (Xia and Xie 2001). The data obtained for all species of each genus analysed, were used to calculate the mean for the genus, with the minimum and maximum within the group serving as the negative error value and the positive error value. These values for each genus were then plotted per codon and sorted by amino acids in Microsoft Excel 2013. Two-tailed, unpaired t-tests were performed in Microsoft Excel 2013 in a pair-wise manner, to determine whether mean values between genera differed significantly (H_0: $\overline{\text{Genus 1}} = \overline{\text{Genus 2}}$; $\alpha = 0.05$).

Lineage specific rate heterogeneity

To determine the presence of lineage specific rate heterogeneity, Tajima's relative rate tests (Tajima 1993) were performed in MEGA 6.0.6 (Tamura et al. 2013). Molecular sequences of three taxa were tested at a time. The amino acid *Erwinia* + *Pantoea* + *Tatumella* dataset was used for rate tests. The null hypothesis tested was equal rates across all taxa.

Long branch attraction

The possible involvement of *P. calida*, *P. gaviniae* and *Tatumella* in long branch attraction (LBA) was investigated in the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid and nucleotide datasets. To determine the effect of the inclusion of these taxa, phylogenetic trees were constructed (as described previously) from the respective datasets *Erwinia* + *Pantoea* + *Tatumella* + Outgroups with the respective inclusion and exclusion of these taxa (Bergsten 2005).

The same process was then applied to the rMLSA and MLSA datasets, as well as the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid dataset, with focus on the outgroup taxa included. This involved including various combinations as well as single outgroups for rooting of the trees. These phylogenetic analyses utilized FastTree 2.1.8 (Price et al. 2010) for inferring the tree with SH-support values for branch support.

Results

Genome sequences of twelve *Pantoea* species

The genome assemblies of the twelve species consisted of 3.9–5.8 million bases at sequencing depths ranging from 13× to 155× (Supplementary Table S1). The overall assembly statistics for these new WGSs were comparable to those for most previously reported *Pantoea* species (Smits et al. 2010; Wang et al. 2011; Brown et al. 2012; Hong et al. 2012; Conlan et al. 2014; De Maayer et al. 2014; Lim et al. 2014; Tian and Jing 2014; Wan et al. 2015). All twelve assemblies have been deposited in the relevant nucleotide database at NCBI (see Table 1 for accession numbers).

ANI-based taxon selection

Adequate taxon sampling is crucial for the accuracy of phylogenetic analyses by allowing better model and parameter estimation (Zwickl and Hillis 2002; Heath et al. 2008; Nabhan and Sarkar 2012) and avoiding artefacts associated with divergent taxa (Kim 1996; Hillis 1998; Mitchell et al. 2000; Philippe et al. 2011). We therefore evaluated and improved the taxon selection for this study using the whole genome similarity metric ANI (Gevers et al. 2005; Konstantinidis and Tiedje 2005; Richter and Rosselló-Móra 2009). This tool was used to ensure broad and appropriate sampling within each of the three genera, where it enabled identification of undescribed isolates

of species for which WGSs are available. These included *Pantoea* sp. A4, At-9b, GM01, PSNIH1, PSNIH2, IMH, *Erwinia* sp. Ejp617 and 9145 and *Tatumella* sp. UCD-D suzukii. This approach also allowed the exclusion from our datasets of what can be considered conspecifics (i.e., those with ANI > 96%) (Gevers et al. 2005; Konstantinidis and Tiedje 2007; Richter and Rosselló-Móra 2009) for which WGS information is available (*Pantoea* sp. PSNIH1, PSNIH2, *Erwinia* sp. Ejp617 and *Tatumella* sp. UCD-D suzukii).

ANI analysis was also used to investigate the similarity of the different taxa within these genera. The members of the respective genera all had ANI values ca. >75% (Fig. 1 and Supplementary Tale S2). However, despite having an ANI value of ca. > 88% between the pair, *P. gaviniae* LMG 25382^{T} and *P. calida* LMG 25383^{T} showed much lower ANI values with other *Pantoea* species at 73.44-78.48%. For the comparisons of these two species with *Erwinia* and *Tatumella*, ANI values of 73.59-77.54% and 70.9-72.43%, respectively, were obtained. Also, the isolate labelled as "*Pantoea* sp. IMH" likely represents a member of *Erwinia* (Rezzonico et al. 2016) due to the high ANI values it shares with other strains belonging to this genus. The final dataset thus consisted of 44 taxa, which included 21 strains of *Pantoea*, three species of *Tatumella*, nine strains of *Erwinia* (including *Pantoea* sp. IMH), in addition to the unusual taxa *P. gaviniae* and *P. calida* (Table 1). The dataset also

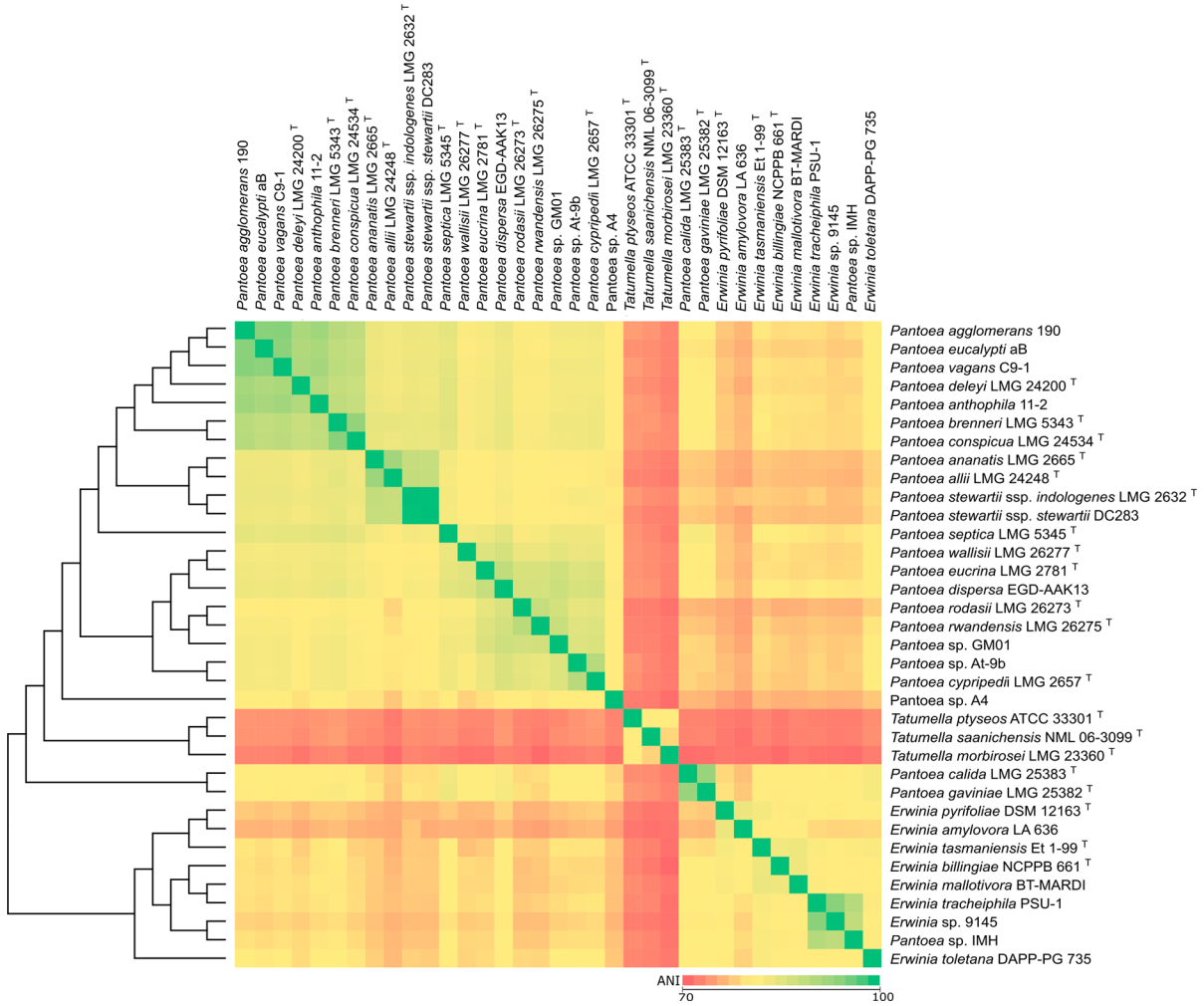


Fig. 1 A cladogram inferred from the amino acid topology for the genes shared by *Erwinia*, *Pantoea* and *Tatumella*. Pairwise Average Nucleotide Identities (ANIb) calculated using BLAST in JSpecies (Richter and Rosselló-Móra 2009) are indicated as a heat map

contained nine taxa from other genera in the *Enterobacteriaceae* that were included to serve as outgroup taxa (Table 1).

WGS-based phylogeny for *Pantoea* and its relatives *Erwinia* and *Tatumella*

Although the various *Pantoea*, *Erwinia*, and *Tatumella* genomes examined had 1112 genes in common, the dataset including nine outgroup taxa (i.e., *Erwinia* + *Pantoea* + *Tatumella* + Outgroups) consisted of 44 taxa and 1039 genes. These genes were identified using the strict orthology estimation implemented in EDGAR (Blom et al. 2016), resulting in a mean % identity of ~69% (median ~73%) and a mean Expect-value of 1e-09 (median 1e-118) for accepted BLAST hits. The nucleotide alignment for the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset contained 679,685 characters, while the amino acid alignment consisted of 224,707 characters. The overall ML topologies obtained for these datasets were similar in terms of relationships among the ingroup taxa. The only differences between the trees related to *Pantoea* sp. A4 and the clade containing *Pantoea eucalypti* (De Maayer et al. 2012), *P. vagans* and *P. agglomerans* (Supplementary Fig. S1). However, the results of SH-like tests implemented in RAxML (Fig. 2) showed that the amino acid topology does not score significantly worse in terms of likelihood than that of the nucleotide topology for the nucleotide data matrix, while the nucleotide topology scored significantly worse than the amino acid topology for the amino acid data matrix. Based on this information and the various estimates regarding its robustness (see below), the tree inferred from the amino acid *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset

Topologies \ Data matrix	Core AA	Core nt	Ribosomal nt	MLSA nt	Informational AA	Cellular functioning AA	External factors AA	Diversifying AA	Neutral AA	Purifying AA
Core AA	Best	No	Yes	Yes	No	No	No	Yes	No	No
Core nt	Yes	Best	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ribosomal nt	Yes	Yes	Best	No	Yes	Yes	Yes	Yes	Yes	Yes
MLSA nt	Yes	Yes	Yes	Best	Yes	Yes	Yes	Yes	Yes	Yes
Informational AA	No	No	Yes	Yes	Best	No	No	Yes	No	No
Cellular functioning AA	No	No	Yes	Yes	Yes	Best	No	Yes	No	Yes
External factors AA	Yes	Yes	Yes	Yes	Yes	Yes	Best	Yes	No	Yes
Diversifying AA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Best	No	Yes
Neutral AA	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Best	Yes
Purifying AA	No	No	Yes	Yes	No	No	No	Yes	No	Best

* Best = Topology obtained with original maximum likelihood analysis
Yes = Likelihood value significantly lower at a confidence level of 1%
No = Likelihood value not significantly lower at a confidence level of 1%
AA = Amino acid sequence based topology/ data matrix
nt = Nucleotide sequence based topology/ data matrix

Fig. 2 Shimodaira-Hasegawa (SH) topology tests (Stamatakis 2014) performed with all topologies showing differences in the relationships among ingroup taxa compared to the topology obtained from the protein sequences of all shared genes. The data matrices are indicated at the top of the figure, with the corresponding topology obtained indicated on the left. The type of data used (nucleotide—nt or amino acid—aa) are indicated for each data matrix. Alternate topologies were scored as either significantly worse or not significantly worse at a confidence interval of 1% based on the likelihood scores obtained for the topologies given the data matrix

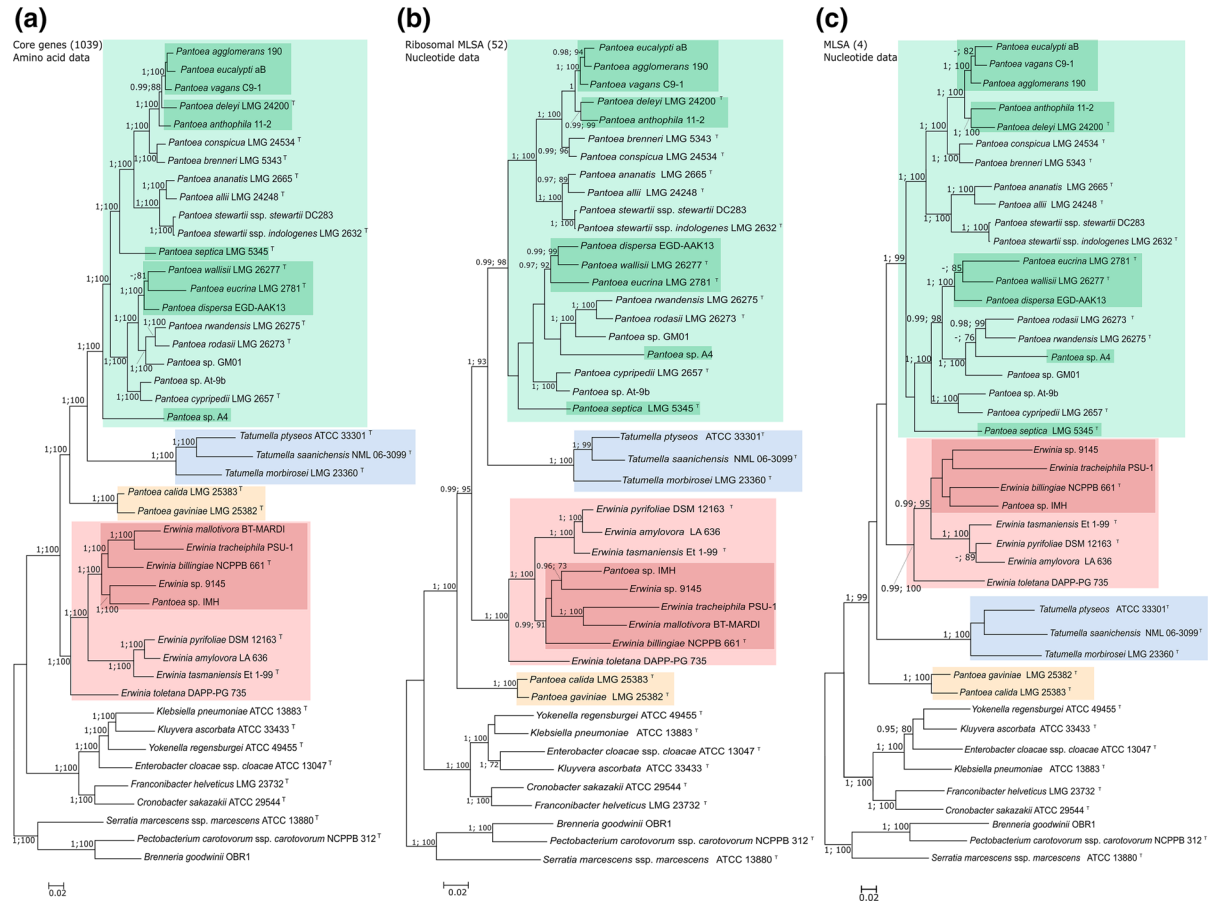
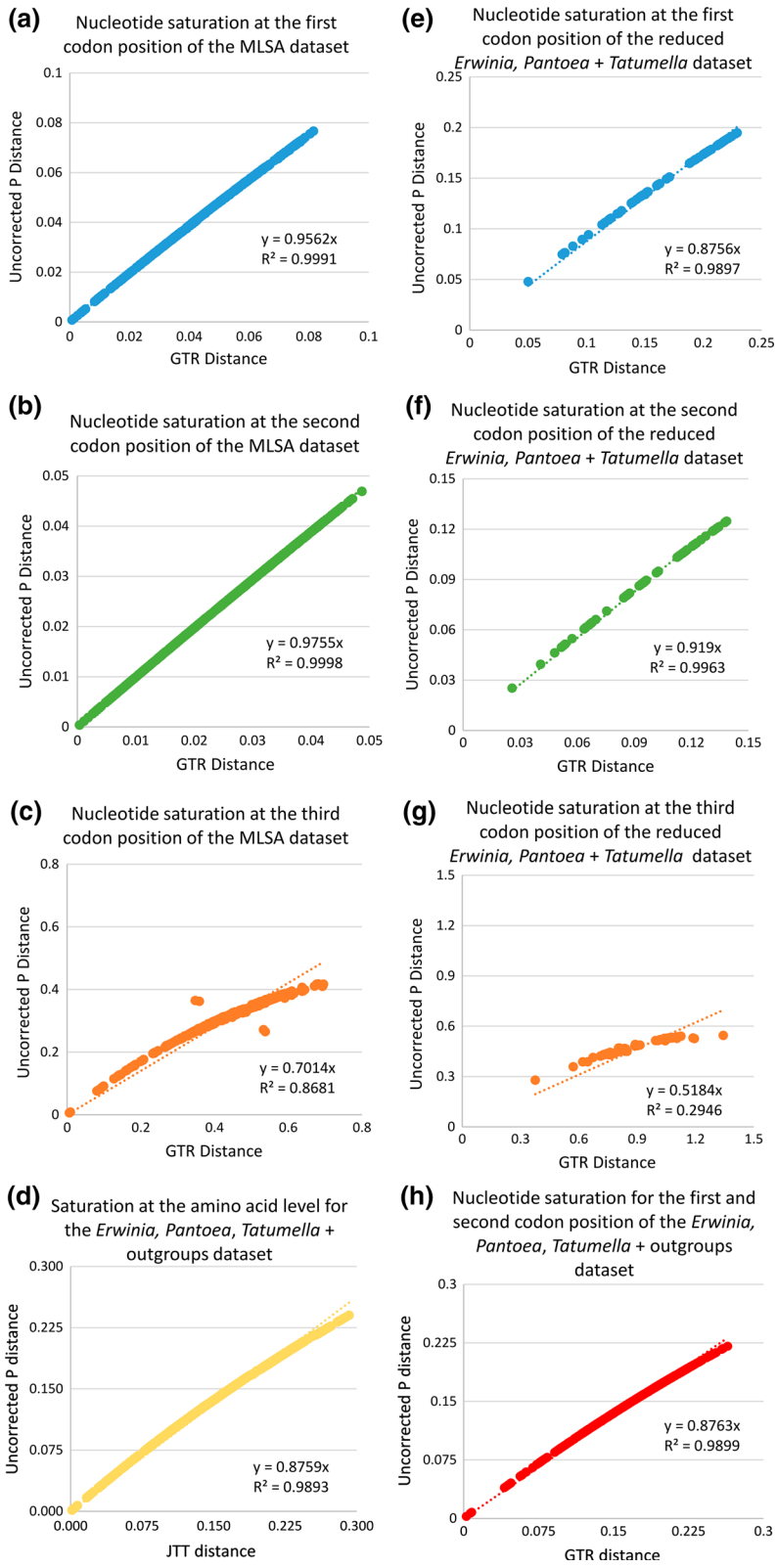


Fig. 3 Maximum-Likelihood (ML) phylogenies of **a** the amino acid *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset, **b** ribosomal MLSA and **c** the conventional MLSA (*atpD*, *gyrB*, *infB* and *rpoB*). For the rMLSA dataset a "Swiss-cheese" dataset was constructed due to the absence of mostly single genes in a number of taxa being potentially due to sequencing quality and assembly of genomes. *E. mallotivora* was also excluded from the MLSA dataset as one of the MLSA genes (*gyrB*) was absent from this genome, possibly also due to sequencing quality. All ML trees were constructed from partitioned datasets using RAxML (Stamatakis 2014) with branch support inferred from FastTree (Price et al. 2010) with SH-support and bootstrap values inferred from 1000 replicates indicated at nodes. *Darkened blocks* indicate differences in the relationships among ingroup taxa across the three topologies

was regarded as the more accurate hypothesis for describing the inter- and intrageneric relationships among the ingroup taxa.

The three genera were recovered as monophyletic groups with high support (Fig. 3a). These analyses also showed that the species *P. calida* and *P. gaviniae* probably represent a distinct genus potentially including some newly described species of these genera, while *Pantoea* sp. IMH represents a member of the genus *Erwinia*. Overall, *Pantoea* and *Tatumella* grouped as sister to each other, followed by the *P. calida* and *P. gaviniae* group (potentially a novel genus), with *Erwinia* grouping basal to the other two genera. *Pantoea* was separated into four distinct lineages, where one (represented by a clade containing *P. agglomerans*, *P. allii*, *P. ananatis*, *P. anthophila*, *P. brenneri*, *P. conspicua*, *P. deleyi*, *P. eucalypti*, *P. stewartii* ssp. *indologenes*, *P. stewartii* ssp. *stewartii* and *P. vagans*) was sister to *P. septica*, which together formed the sister group of the third lineage (represented by the clade containing *P. cypripedii*, *P. dispersa*, *P. eucrina*, *P. rodasii*, *P. rwandensis*, *P. wallisii*, *Pantoea* sp. At-9b and GM01). The


(a)
Nucleotide saturation at the first codon position of the MLSA dataset
Uncorrected P Distance
GTR Distance
y = 0.9562x
R² = 0.9991
(b)
Nucleotide saturation at the second codon position of the MLSA dataset
Uncorrected P Distance
GTR Distance
y = 0.9755x
R² = 0.9998
(c)
Nucleotide saturation at the third codon position of the MLSA dataset
Uncorrected P Distance
GTR Distance
y = 0.7014x
R² = 0.8681
(d)
Saturation at the amino acid level for the Erwinia, Pantoea, Tatumella + outgroups dataset
Uncorrected P distance
JTT distance
y = 0.8759x
R² = 0.9893
(e)
Nucleotide saturation at the first codon position of the reduced Erwinia, Pantoea + Tatumella dataset
Uncorrected P Distance
GTR Distance
y = 0.8756x
R² = 0.9897
(f)
Nucleotide saturation at the second codon position of the reduced Erwinia, Pantoea + Tatumella dataset
Uncorrected P Distance
GTR Distance
y = 0.919x
R² = 0.9963
(g)
Nucleotide saturation at the third codon position of the reduced Erwinia, Pantoea + Tatumella dataset
Uncorrected P Distance
GTR Distance
y = 0.5184x
R² = 0.2946
(h)
Nucleotide saturation for the first and second codon position of the Erwinia, Pantoea, Tatumella + outgroups dataset
Uncorrected P distance
GTR distance
y = 0.8763x
R² = 0.9899

◀**Fig. 4** Substitution saturation plots of the first (*blue*; **a**), second (*green*; **b**) and third (*orange*; **c**) codon positions of the MLSA dataset. Uncorrected p-distances are indicated on the y-axis and GTR distances are indicated on the x-axis. The slopes of the linear regression lines for this dataset are 0.9562, 0.9755 and 0.7014 with R^2-values of 0.9991, 0.9998 and 0.8681, respectively. **d** Substitution saturation plot of the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid dataset. Uncorrected p-distances are indicated on the y-axis and JTT distances are indicated on the x-axis. The slope of the linear regression line is 0.8759 with an R^2-value of 0.9893. Substitution saturation plots of the first (*blue*; **e**), second (*green*; **f**) and third (*orange*; **g**) codon positions of the *Erwinia* + *Pantoea* + *Tatumella*_reduced dataset. Uncorrected p-distances are indicated on the y-axis and GTR distances are indicated on the x-axis. The slopes of the linear regression lines for this dataset are 0.8756, 0.919 and 0.5184 with R^2-values of 0.9897, 0.9963 and 0.2946, respectively. **h** Substitution saturation plot of the combined first and second codon positions for the *Erwinia* + *Pantoea* + *Tatumella*_reduced dataset. Uncorrected p-distances are indicated on the y-axis and GTR distances are indicated on the x-axis. The slope of the linear regression line is 0.8763 with an R^2-value of 0.9899

fourth lineage, represented by *Pantoea* sp. A4, was sister to these three lineages.

Robustness of the *Pantoea* phylogeny

The robustness of the phylogenies obtained from the nucleotide and amino acid *Erwinia* + *Pantoea* + *Tatumella* + Outgroups datasets were evaluated in terms of factors known to cause so-called "non-phylogenetic signal" (Jeffroy et al. 2006, Philippe et al. 2011), as well as potential biases introduced due to the choice of genes analysed (Rivera et al. 1998; Jain et al. 1999; Cohen et al. 2011). The causes of non-phylogenetic signal investigated were homoplasy, substitution saturation, codon usage bias, LBA and lineage specific rate heterogeneity. For identifying inherent biases due to the selected genes, different subsets of the shared genes were constructed. The subsets were either based on the type of selection experienced by the genes, or the functional classes to which the genes belong.

Non-phylogenetic signal—homoplasy

The contribution of homoplasious characters to the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups datasets was estimated using PAUP*. These analyses yielded HI values of 0.76, 0.74 and 0.53, respectively, for the dataset with all nucleotides included, the nucleotide dataset with the third codon positions excluded, and the amino acid dataset. Compared to the two nucleotide datasets, the amino acid dataset thus contained substantially fewer homoplasious characters over the tree topology (Bremer 1994) that could contribute to the non-phylogenetic signal (Philippe et al. 2011). The amino acid dataset is thus superior in that it contains fewer characters competing with the true phylogenetic signal during tree inference (Philippe et al. 2011).

Non-phylogenetic signal—substitution saturation

Substitution saturation (like homoplasy) contributes to the non-phylogenetic signal that competes with the true signal, which detracts from the robustness and accuracy of the inferred tree (Philippe and Forterre 1999; Xia et al. 2003; Jeffroy et al. 2006; Philippe et al. 2011). To estimate the level of saturation in our WGS-based datasets, the correlation between actual substitutions in the data (represented by p-distances) and substitutions inferred using an appropriate evolutionary model (represented by modelled distances) was inferred (Jeffroy et al. 2006; Philippe et al. 2011). For the three *Erwinia* + *Pantoea* + *Tatumella*_reduced datasets consisting, respectively, of the first codon positions, second codon positions, and first plus second codon positions, there is an almost one to one correlation between p-distances and the distances that compensate for potential saturation (Fig. 4e, f, h). This was also true for the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid dataset (Fig. 4d). These results thus suggest a limited effect of substitution saturation on the amino acid data and the nucleotide datasets including the first and second codon positions (Jeffroy et al. 2006; Philippe et al. 2011).

The uncorrected and modelled distances were, however, poorly correlated in the nucleotide dataset containing the third codon positions only (Fig. 4g). The latter dataset thus contains many more characters that have undergone multiple mutations over evolutionary time, which could explain why one of the nodes in the backbone of the tree inferred from this dataset lacked statistical support (Supplementary Fig. S1).

Table 2 Differences in relative synonymous codon usage between *Erwinia*, *Pantoea* and *Tatumella*

Differences between taxa	Total significant differences*	Amino Acids (Codons)
Pantoea and *Erwinia*	14	**A** (GCG, GCC), **D** (GAU, GAC), **G** (GGU), **I** (AUU), **K** (AAA, AAG), **L** (CUC, UUA, UUG), **S** (UCC, UCG), **V** (GUU)
Pantoea and *Tatumella*	37	**A** (GCU, GCG, GCA), **C** (UGU, UGC), **D** (GAU, GAC), **E** (GAG, GAA), **F** (UUU, UUC), **G** (GGU, GGG, GGC), **H** (CAC, CAU), **K** (AAA, AAG), **L** (CUU), **N** (AAC, AAU), **P** (CCU), **Q** (CAA, CAG), **R** (CGA, CGC, CGU), **S** (AGC, AGU, UCG, UCU), **T** (ACG, ACU), **V** (GUG, GUA), **Y** (UAC, UAU)
Erwinia and *Tatumella*	28	**A** (GCG, GCA), **C** (UGU, UGC), **D** (GAU, GAC), **G** (GGU, GGC), **H** (CAC, CAU), **K** (AAA, AAG), **N** (AAC, AAU), **P** (CCU), **R** (CGC, CGU), **S** (AGC, AGU, UCG, UCU), **T** (ACA, ACG, ACU), **V** (GUG, GUC), **Y** (UAC, UAU)
(*Pantoea*, *Erwinia*) and *Tatumella*	19	**A** (GCA), **C** (UGU, UGC), **G** (GGC), **H** (CAC, CAU), **N** (AAC, AAU), **P** (CCU), **R** (CGC, CGU), **S** (AGC, AGU, UCU), **T** (ACG, ACU), **V** (GUG), **Y** (UAC, UAU)

* Statistical significance as determined with pairwise two-tailed unpaired t-tests ($p < 0.05$)

Non-phylogenetic signal—codon composition bias

As an indication of codon usage biases, the RSCU of species in the three genera, *Erwinia*, *Pantoea* and *Tatumella*, were analysed. Upon comparison of either *Erwinia* (28 codons) or *Pantoea* (37 codons) to *Tatumella* (Table 2, Supplementary Fig. S2), it was clear that *Tatumella* utilizes a large number of codons for certain amino acids that are different from those used by *Erwinia* and *Pantoea*. This might reflect a bias toward certain nucleotides in *Tatumella* (Nei and Kumar 2000), particularly at third codon positions (Jeffroy et al. 2006). Like homoplasy and substitution saturation, such biases also contribute the non-phylogenetic signal that might overshadow the true signal during tree reconstruction (Galtier and Gouy 1995; Jeffroy et al. 2006). The apparent codon composition bias in *Tatumella* is therefore the likely cause of the somewhat longer branch for this genus in our various WGS-based phylogenies.

Non-phylogenetic signal—LBA

LBA is a tree reconstruction artefact which indicates a closer relationship between certain taxa, due to the divergent nature of these taxa compared to the rest of the taxa in the analysis (Bergsten 2005). The effect of LBA on the tree inferred from the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid dataset (Fig. 3a) was evaluated by removing and adding different combinations of taxa with long branches (Bergsten 2005). These taxa were the *Tatumella* group and the group containing *P. calida* and *P. gaviniae*. Exclusion of these groups (singly or combined) in the amino acid dataset did not alter the position of any of the remaining taxa (ingroup or outgroup), including the basal position of *Pantoea* sp. A4 within *Pantoea* (Supplementary Fig. S3). However, differential exclusion of these taxa appeared to alter the topology of the tree inferred using nucleotide data (Supplementary Fig. S4), where the presence of these two groups, but particularly the *P. calida* and *P. gaviniae* group, appears to influence the position of *Pantoea* sp. A4. Upon the inclusion of the *P. calida* and *P. gaviniae* group, the basal position of *Pantoea* sp. A4 changes to what is observed in the nucleotide topologies, whereas inclusion of *Tatumella* does not alter this basal position of *Pantoea* sp. A4. These data thus suggest that, despite attempts to counter LBA (i.e., the use of appropriate taxon selection and evolutionary models) (Zwickl and Hillis 2002; Heath et al. 2008; Nabhan and Sarkar 2012), the inclusion of certain taxa (specifically *P. calida* and *P. gaviniae*) in the nucleotide dataset has a significant effect on the accuracy of the phylogeny reconstructed from it.

We also tested the possible LBA-effect of outgroup selection on the ML tree inferred from the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid dataset. The results of the nine separate analyses

(which each included the 35 ingroup taxa and one of the nine outgroup taxa) showed that outgroup selection had a limited effect on the robustness of the tree inferred from the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid dataset (Supplementary Fig. S10). For all intrageneric relationships, the only variation observed involved the relationships (generally lacking statistical support) among the closely related *P. agglomerans*, *P. eucalypti* and *P. vagans*. The only outgroup that affected the intergeneric relationships was *Cronobacter*, which caused *P. calida* and *P. gaviniae* to group sister to the *Pantoea* + *Tatumella* + *Erwinia* clade. In the remaining eight analyses, these two species formed the sister taxon of the *Pantoea* + *Tatumella* clade similar to what is observed in the trees inferred from the 44-taxon *Erwinia* + *Pantoea* + *Tatumella* + Outgroups amino acid and nucleotide datasets. This suggests that the use of phylogenetic signal associated with the other outgroup taxa sufficiently compensated for the non-phylogenetic signal associated with the *Cronobacter* sequence (Philippe et al. 2011).

Non-phylogenetic signal—lineage specific rate heterogeneity

The first relative rate test utilized the amino acid sequences of a *Pantoea* isolate (*P. agglomerans*), an *Erwinia* isolate (*E. amylovora*) and an outgroup taxon (*S. marcescens*) (Supplementary Table S3). A *p* value of 0.62013 was obtained indicating that the null hypothesis of equal rates across the taxa could not be rejected. The second relative rate test utilized the amino acid sequence of a *Pantoea* isolate (*P. agglomerans*), a *Tatumella* isolate (*T. morbirosei*) and an *Erwinia* isolate (*E. amylovora*) as the more distantly related taxon (as is observed from the phylograms; Supplementary Table S4). A p-value of 0 was obtained, thus leading to the rejection of the null hypothesis of equal rates across taxa, indicating lineage specific rate heterogeneity. Various combinations of different representatives of the different genera generally resulted in similar results. These data are thus congruent with the results of the RSCU analysis and suggest that *Tatumella* evolves at a faster evolutionary rate compared to either *Erwinia* or *Pantoea*.

Comparison of data subsets—'Purifying', 'Diversifying' and 'Neutral' selection

Analysis with HyPhy showed that most of the 1039 genes included in the *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset likely experience purifying selection, which is consistent with what has been proposed for housekeeping or core genes involved in essential functions (Koonin 2005; Koonin and Wolf 2006; Alvarez-Ponce et al. 2016). Among the 1039 shared genes, 218 genes had dN/dS values higher than 1 (diversifying selection) and 820 genes had values lower than 1 (purifying selection) (Supplementary Fig. S5), while one gene were too truncated in some taxa to include in the analysis. Of the set of 1038 gene included in the analyses, only 13 formed part of the neutral or nearly neutral category. These three categories of genes could be expected to evolve at different rates (Alvarez-Ponce et al. 2016), as was clear from the trees inferred using the amino acid datasets (Supplementary Fig. S6). However, the tree obtained from the 'purifying' amino acid dataset was fully congruent with the one obtained from the dataset including the amino acids for all 1039 genes (compare Fig. 3a and Supplementary Fig. S6). This suggests that the majority of the core genome is under purifying selection and contributes to the overall phylogenetic signal in the combined dataset. The incongruence between the 'neutral' and 'purifying' amino acid trees is likely due, in part, to a lack of phylogenetic signal in the 'neutral' dataset that includes only thirteen genes. The unusual relationships inferred from the 'diversifying' amino acid dataset is probably due to the limited constraints in terms of how these genes evolve, which allows increased fixation of non-synonymous substitutions in these genes. Although differing topologies were observed for these datasets, the likelihood of the tree topology obtained for the amino acid *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset measured against the neutral and purifying amino acid datasets could not be rejected based on the SH tests (Fig. 2).

Comparison of data subsets—'Cellular functioning', 'Metabolism', 'Informational', 'External factors' and 'Unclassified' functional categories

To maintain functionality, the genes involved in a specific cellular process (particularly those

characterized by high levels of complexity) often evolve in concert and may follow similar evolutionary trajectories (Rivera et al. 1998; Jain et al. 1999; Daubin et al. 2002). The genes involved in certain processes are also more prone to HGT than others, despite representing part of the core genome component (Rivera et al. 1998; Jain et al. 1999). Therefore, to assess the possible influence that the functional categories might have had on our species tree, the 1039 genes were separated into their functional categories and subjected to phylogenetic analyses (Supplementary Fig. S7). The data subsets comprised of between 80 ('External factors') and 336 genes ('Unclassified'), with the 'Cellular functioning', 'Metabolism' and 'Informational' functional categories incorporating 240, 236 and 281 genes, respectively, with some genes being involved in multiple functional categories. The overall relationships among the ingroup taxa of all subset trees supported the full core genome protein sequence topology, with minor differences within *Erwinia* ('Cellular functioning' and 'External factors' tree topologies) and *Pantoea* ('Informational' tree topology). This suggests that the topology obtained from the concatenation of all shared protein sequences is not influenced by the functional constraints of the chosen genes or large-scale HGT, potentially leading to false phylogenies. Despite these minor topological differences, the species tree obtained from the amino acid *Erwinia* + *Pantoea* + *Tatumella* + Outgroups dataset did also not score significantly worse based on the SH test in terms of likelihood compared to the trees obtained from the data subsets (Fig. 2).

Problems with the MLSA and rMLSA trees

ML trees generated from the MLSA dataset (consisting of four protein-coding gene sequences, Supplementary Fig. S8) and the rMLSA dataset (consisting of gene sequences for 52 ribosomal proteins, Supplementary Fig. S9) all differed markedly from the tree inferred using the amino acid dataset for the 1039 shared genes (Fig. 3a with 3b, c). Upon comparison to the WGS-based phylogenies, it could be seen that the alternate topologies tested had significantly lower likelihood values based on the SH test compared to the trees obtained for each dataset during the respective ML analyses (Fig. 2). This indicates drastically different topologies for these datasets that are not reconcilable between these datasets.

In contrast to the 1039-shared gene tree, both the MLSA and rMLSA trees further included numerous branches lacking statistical support. To some extent this is due to the limited sizes of these datasets, which would accordingly also lack sufficient phylogenetic signal especially at the nucleotide level. This was particularly evident in the MLSA dataset, as has been suggested previously (Gevers et al. 2005). As a measure of phylogenetic noise, HI was 0.745 (all nucleotides), 0.603 (third nucleotide excluded) and 0.526 (amino acid), respectively. HI values for the ribosomal dataset were 0.700 (all nucleotides), 0.672 (third nucleotide excluded) and 0.588 (amino acid). Similar to the 1039-shared gene dataset, more homoplasious characters were thus present in the MLSA and rMLSA nucleotide datasets than their corresponding amino acid datasets.

As with the 1039-shared gene dataset, limited substitution saturation was detected in the first and second codon positions of genes included in the smaller MLSA (Fig. 4 a, b and c) and rMLSA datasets (results not shown). The phylogenies inferred from the nucleotide MLSA and rMLSA datasets containing only first and second codon positions were overall congruent with those inferred from the respective amino acid datasets (Supplementary Figs. S8 and S9). However, inclusion of the third codon positions in the analyses produced trees that were clearly different from the amino acid-based trees of the corresponding dataset (Supplementary Figs. S8 and S9).

The MLSA and rMLSA datasets further appeared to be particularly sensitive to LBA. The use of different outgroup-ingroup combinations generated distinct topologies, both in terms of inter- and intrageneric relationships amongst the ingroup taxa, especially in the MLSA dataset (Supplementary Fig. S10). Among the nine combinations tested, none were congruent with the tree topology inferred from the amino acid sequence of the 1039 shared genes.

Discussion

Among the genomes for twenty-three *Pantoea* strains (twelve of which were determined in this study), three *Tatumella* species, nine *Erwinia* strains and their nine outgroup taxa, a set of 1039 single-copy shared genes

were identified. These genes formed part of the core genomes of the species harbouring them and were most likely inherited in a vertical fashion (Hacker and Carniel 2001; Daubin et al. 2002). This core genomic component is also thought to be essential for survival as most of these genes are involved in complex processes requiring the interaction of these genes with one another, leading to concerted evolutionary paths (Rivera et al. 1998; Jain et al. 1999; Daubin et al. 2002; Cohen et al. 2011). Shared evolutionary trajectories are thus expected for groups of genes that are functionally constrained due to their intergenic interactions. Thus, the overall similarities of the phylogenies obtained for the different functional subsets were expected, as this overall core component should be evolutionarily relatively cohesive providing congruent phylogenetic hypotheses (Daubin et al. 2002).

The amino acid dataset for the 1039 genes used in this study contained much less non-phylogenetic signal than the corresponding nucleotide dataset. The term non-phylogenetic signal refers to the combined effects of different kinds of structured phylogenetic noise (Jeffroy et al. 2006; Philippe et al. 2011). Similar to what has been shown previously, the nucleotide dataset contained higher levels of substitution saturation, particularly at third codon positions (Xia et al. 2003; Jeffroy et al. 2006). The nucleotide dataset was also more homoplasious, potentially because the accumulation of convergent mutations in data with four character states is more pronounced than in amino acid data with 20 character states (Xia et al. 2003; Jeffroy et al. 2006). However, despite being less "noisy", the amino acid dataset remained affected by non-phylogenetic signal. In addition to containing low levels of homoplasy and substitution saturation, the codon usage bias detected in the nucleotide dataset likely gave rise to the lineage-specific rate heterogeneity observed in the amino acid dataset. The non-phylogenetic signals inherent to the amino acid dataset could, therefore, be problematic during tree reconstruction.

In this study, we attempted to limit the negative effects of non-phylogenetic signal during tree inference in three ways (Philippe et al. 2011). Firstly, we utilized strict criteria for identifying the genes included in the analyses, i.e., BLAST bit score ratios adjusted automatically depending on the data analysed (Blom et al. 2016). Although this might have led to the exclusion of less conserved genes, it allowed for the construction of a concatenated dataset consisting mainly of orthologous sequences (related via speciation or vertical descent) (Koonin 2005). Secondly, to avoid the artificial introduction of "noise", an iteration-based method, which takes into account relatedness during iterative pair-wise alignment, was used to generate optimal sequence alignments (Edgar 2004; Philippe et al. 2011). Thirdly, phylogenies were inferred using a probabilistic method (i.e., Maximum Likelihood) with appropriate models to approximate the evolution of individual genes making up the dataset (Philippe et al. 2011). Our findings clearly showed that this approach was highly effective for analysing the amino acid dataset, as the non-phylogenetic signal it included did not seem to influence the topology of the final tree. For example, LBA is one of the best-understood outcomes of non-phylogenetic signal (Philippe et al. 2011), yet the tree inferred from the amino acids of 1039 genes appeared to be relatively unaffected by this phenomenon.

To further interrogate the robustness of the tree inferred from the aligned amino acid sequences of 1039 genes, different subsets of these data were evaluated phylogenetically. The first set of analyses involved subsets based on selection, where almost 80% of the genes seemed to experience purifying selection due to high levels of functional conservation (Jain et al. 1999; Lan and Reeves 2000; Coenye et al. 2005). Not surprisingly, the phylogeny inferred from the amino acids for these genes matched the phylogeny inferred from the 1039 gene dataset (Sarkar and Guttman 2004; He et al. 2010). The tree inferred from the 13 neutrally evolving genes lacked resolution, probably due to inadequate phylogenetic signal, similar to what has been observed for other small datasets (Daubin et al. 2002; Coenye et al. 2005; Galtier and Daubin 2008; Bennett et al. 2012; Chan et al. 2012). The tree inferred from the 218 genes under diversifying selection also lacked resolution, but in this case it is likely due to the accumulation of non-phylogenetic signal introduced during diversifying evolution (Xia et al. 2003; Jeffroy et al. 2006). Overall, however, these results suggest that the majority of the core genome evolved in a cohesive manner due to the purifying selection acting on this genomic compartment.

The second set of analyses concentrated on five subsets of the 1039 shared genes involved in the different functional categories of the products encoded

by individual genes as well as unclassified genes. The trees inferred from all of these five amino acid datasets tested, generally matched the one inferred from the 1039 amino acid dataset. There were, however, small differences within the topologies obtained for the genes involved in 'Cellular functioning' and the 'Informational' genes, although the sister-groupings observed were without statistical support. Such subtle incongruences in topologies inferred from different functional subsets are not uncommon (Wolf et al. 2001; Lerat et al. 2003; Dutilh et al. 2004; Ma and Zeng 2004). In fact, much greater discordance is often seen for the phylogenies inferred from different functional subsets when distantly related bacteria are considered (Dutilh et al. 2004; Ma and Zeng 2004). Thus, despite minor differences observed from the different datasets, possibly due to "noise", the robust amino acid based phylogeny obtained for the full set of shared genes were reflected in all functional subset tree topologies.

Taken together, our findings suggest that the tree inferred from the amino acid data for the 1039 shared genes represents the best hypothesis of explaining the inter- and intrageneric relationships examined in this study. None of the various factors typically responsible for destabilizing phylogenetic trees (Philippe et al. 2011) appeared to significantly affect it. In other words, despite containing detectable levels of non-phylogenetic signal, the use of amino acid data (Glaeser and Kämpfer 2015), together with a suitable set of outgroup taxa and the application of appropriate evolutionary models fitted against each gene partition (Jeffroy et al. 2006; Philippe et al. 2011), provided the most robust phylogenetic hypothesis for describing the relationships within *Pantoea* and its relationships with *Tatumella* and *Erwinia*.

Pantoea, *Tatumella* and *Erwinia* were generally recovered as monophyletic groups. In accordance with what has been found previously (Brady et al. 2010b; Brady et al. 2012; Glaeser and Kämpfer 2015), it was also consistently observed that *Pantoea* and *Tatumella* group as sister to each other. In all phylogenetic analyses (amino acid and nucleotide), the two species *P. calida* and *P. gaviniae* appeared to form a unique and separate cluster. These two species thus represent a novel genus due to the distinctness of these taxa when compared to the closest related taxa (Gavini et al. 1989a; Konstantinidis and Tiedje 2005). Future description of a novel genus will be required to accommodate these and potentially other atypical *Pantoea* and *Erwinia* species not included in the study.

The robust species tree obtained in this study also allowed elucidating intrageneric relationships as comparison of the different phylogenies produced consistent species groupings within the respective genera. *Pantoea* sp. A4 was consistently recovered as part of *Pantoea* as suggested before (Hong et al. 2012), where it forms a basal lineage within the genus. Contrary to what was expected (Wu et al. 2013; Tian and Jing 2014), *Pantoea* sp. IMH consistently grouped within *Erwinia*, as the 16S rRNA gene initially used for identification purposes is known to lack resolution within the *Enterobacteriaceae* (Rezzonico et al. 2009; Glaeser and Kämpfer 2015) which could have led to the misidentification of this taxon. The description of *Pantoea* sp. IMH as an *Erwinia* species is thus required as it may also represent a novel species. Further study and comparison to other *Erwinia* species is however required to determine whether this isolate forms part of a novel exclusive and cohesive cluster within *Erwinia*.

The results of our study also showed that conventional MLSA and rMLSA are inadequate for inferring inter- and intrageneric relationships due to the limited number of loci used in the analyses. MLSA and rMLSA phylogenies yield inconsistent groupings that lack statistical support. Our analyses showed that this could mostly be attributed to a general lack of true phylogenetic signal from which to reconstruct trees. The little true signal present in the data was likely outcompeted by non-phylogenetic signal during tree building. Accordingly, the MLSA and rMLSA phylogenies were both exceedingly sensitive to LBA where outgroup selection severely affected the topology of the ingroup. Although improved taxon sampling could counter the effects of LBA (Hillis 1998; Zwickl and Hillis 2002; Heath et al. 2008), our results showed that this phenomenon remains a problem in smaller datasets. Therefore, apart from the intended use for species delineation, where these approaches have been applied successfully (Brady et al. 2008, 2010a, b, 2012; Glaeser and Kämpfer 2015), these trees lack robustness for investigating relationships at higher taxonomic levels.

Conclusions

The use of shared gene sets for phylogenomic analyses has proven to be a useful tool for obtaining species trees of bacteria (Daubin et al. 2002; Dutilh et al. 2008; Galtier and Daubin 2008; Segata and Huttenhower 2011) and provides better supported and robust phylogenies compared to the commonly employed molecular markers. It has, however, been suggested that the use of only a few genes with strong phylogenetic signal may be more feasible (Konstantinidis and Tiedje 2006; Salichos and Rokas 2013), but their identification will be difficult without the use of a robust phylogeny for comparison. The results presented here indicate that the choice of shared genes for analysis, as well as whether datasets are nucleotide or protein sequence based, remain important as different approaches may provide different evolutionary hypotheses, as has been suggested before (Rivera et al. 1998; Jain et al. 1999; Glaeser and Kämpfer 2015). The robust phylogeny obtained from this data will thus be invaluable for addressing questions pertaining to the evolutionary history of *Pantoea* and its related genera as this provides a framework for investigating how different biological traits, like pathogenicity and other potentially beneficial characteristics, have evolved in these different genera (Heath et al. 2008).

Acknowledgements We would like to acknowledge the Centre for Bioinformatics and Computational Biology, University of Pretoria, for the use of the facility and server access. For genome sequencing, we want to acknowledge the Ion Torrent Sequencing Facility at the University of Pretoria and Markus Oggenfuss and Jürg E. Frey for sequencing at Agroscope (Wädenswil, Switzerland). We would also like to acknowledge the Genome Research Institute (GRI) as well as the Centre of Excellence in Tree Health Biotechnology (CTHB) at the University of Pretoria for additional funding. THMS and BD acknowledge the funding by the Swiss Federal Office of Agriculture ACHILLES project (BLW/FOAG Project ACHILLES) as part of the Agroscope Research Programme ProfiCrops and the Department of Life Sciences and Facility Management of ZHAW.

References

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105

Alvarez-Ponce D, Sabater-Muñoz B, Toft C, Ruiz-González MX, Fares MA (2016) essentiality is a strong determinant of protein rates of evolution during mutation accumulation experiments in *Escherichia coli*. Genome Biol Evol 8:2914–2927

Anda M, Ohtsubo Y, Okubo T, Sugawara M, Nagata Y, Tsuda M, Minamisawa K, Mitsui H (2015) Bacterial clade with the ribosomal RNA operon on a small plasmid rather than the chromosome. Proc Natl Acad Sci USA 112: 14343–14347

Andam CP, Gogarten JP (2011) Biased gene transfer in microbial evolution. Nat Rev Microbiol 9:543–555

Angus AA, Agapakis CM, Fong S, Yerrapragada S, Estrada-de Los Santos P, Yang P, Song N, Kano S, Caballero-Mellado J, de Faria SM, Dakora FD, Weinstock G, Hirsch AM (2014) Plant-associated symbiotic *Burkholderia* species lack hallmark strategies required in mammalian pathogenesis. PLoS ONE 9:e83779

Aziz R, Bartels D, Best A, Dejongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M, Meyer F, Olsen G, Olson R, Osterman A, Overbeek R, Mcneil L, Paarmann D, Paczian T, Parrello B, Pusch G, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST server: rapid annotations using subsystems technology. BMC Genom 9:75

Bennett JS, Jolley KA, Earle SG, Corton C, Bentley SD, Parkhill J, Maiden MCJ (2012) A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. Microbiology 158:1570–1580

Bergsten J (2005) A review of long-branch attraction. Cladistics 21:163–193

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 Res 44:W22–W28

Boto L (2010) Horizontal gene transfer in evolution: facts and challenges. Proc R Soc B Biol Sci 277:819–827

Boucher Y, Douady CJ, Sharma AK, Kamekura M, Doolittle FW (2004) Intragenomic heterogeneity and intergenomic recombination among haloarchaeal rRNA genes. J Bacteriol 186:3980–3990

Brady C, Venter S, Cleenwerck I, Vancanneyt M, Swings J, Coutinho T (2007) A FALFP system for the improved identification of plant-pathogenic and plant-associated species of the genus *Pantoea*. Syst Appl Microbiol. doi:10.1111/j.1472-765X.2009.02692.x

Brady C, Cleenwerck I, Venter S, Vancanneyt M, Swings J, Coutinho T (2008) Phylogeny and identification of *Pantoea* species associated with plants, humans and the natural environment based on multilocus sequence analysis (MLSA). Syst Appl Microbiol 31:447–460

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. Int J Syst Evol Microbiol 59:2339–2345

Brady CL, Cleenwerck I, Venter SN, Engelbeen K, De Vos P, Coutinho TA et al (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. Int J Syst Evol Microbiol 60:2430–2440

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. Int J Syst Evol Microbiol 60:484–494

Brady CL, Goszczynska T, Venter SN, Cleenwerck I, De Vos P, Gitaitis RD, Coutinho TA (2011) *Pantoea allii* sp. nov., isolated from onion plants and seed. Int J Syst Evol Microbiol 61:932–937

Brady CL, Cleenwerck I, Van Der Westhuizen L, Venter SN, Coutinho TA, De Vos P (2012) *Pantoea rodasii* sp. nov., *Pantoea rwandensis* sp. nov. and *Pantoea wallisii* sp. nov., isolated from *Eucalyptus*. Int J Syst Evol Microbiol 62:1457–1464

Bremer KR (1994) Branch support and tree stability. Cladistics 10:295–304

Brown SD, Utturkar SM, Klingeman DM, Johnson CM, Martin SL, Land ML, 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*. J Bacteriol 194:5991–5993

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Chan JZM, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ (2012) Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. BMC Microbiol 12:302

Cleenwerck I, Vandemeulebroecke K, Janssens D, Swings J (2002) Re-examination of the genus *Acetobacter*, with descriptions of *Acetobacter cerevisiae* sp. nov. and *Acetobacter malorum* sp. nov. Int J Syst Evol Microbiol 52:1551–1558

Coenye T, Gevers D, van de Peer Y, Vandamme P, Swings J (2005) Towards a prokaryotic genomic taxonomy. Fed Eur Microbiol Soc Microbiol Rev 29:147–167

Cohen O, Gophna U, Pupko T (2011) The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. Mol Biol Evol 28:1481–1489

Conlan S, Thomas PJ, Deming C, Park M, Lau AF, Dekker JP, Snitkin ES, Clark TA, Luong K, Song Y, Tsai Y-C, Boitano M, Dayal J, Brooks SY, Schmidt B, Young AC, Thomas JW, Bouffard GG, Blakesley RW, Mullikin JC, Korlach J, Henderson DK, Frank KM, Palmore TN, Segre JA (2014) Single-molecule sequencing to track plasmid diversity of hospital-associated carbapenemase-producing *Enterobacteriaceae*. Sci Transl Med 6:254ra126

Conville PS, Witebsky FG (2007) Analysis of multiple differing copies of the 16S rRNA gene in five clinical isolates and three type strains of *Nocardia* species and implications for species assignment. J Clin Microbiol 45:1146–1151

Cruz AT, Cazacu AC, Allen CH (2007) *Pantoea agglomerans*—a plant pathogen causing human disease. J Clin Microbiol 45:1989–1992

Daubin V, Gouy M, Perrière G (2002) A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. Genome Res 12:1080–1090

de Baere T, Verhelst R, Labit C, Verschraegen G, Wauters G, Claeys G, Vaneechoutte M (2004) Bacteremic infection with *Pantoea ananatis*. J Clin Microbiol 42:4393–4395

de Maayer P, Chan W-Y, Blom J, Venter SN, Duffy B, Smits THM, Coutinho TA (2012) The large universal *Pantoea* plasmid LPP-1 plays a major role in biological and ecological diversification. BMC Genom 13:625

de Maayer P, Chan W-Y, Rubagotti E, Venter SN, Toth IK, Birch PRJ, 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 Genom 15:1–28

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9:687–705

Dutilh BE, Huynen MA, Bruno WJ, Snel B (2004) the consistent phylogenetic signal in genome trees revealed by reducing the impact of noise. J Mol Evol 58:527–539

Dutilh BE, Snel B, Ettema TJG, Huynen MA (2008) Signature genes as a phylogenomic tool. Mol Biol Evol 25:1659–1667

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797

Felsenstein, J. 2005. SEQBOOT—bootstrap, jackknife or permutation resampling of molecular sequence, restriction site, gene frequency or character data

Fouts DE, Matthias MA, Adhikarla H, Adler B, Amorim-Santos L, Berg DE, Bulach D, Buschiazzo A, Chang Y-F, Galloway RL, Haake DA, Haft DH, Hartskeerl R, Ko AI, Levett PN, Matsunaga J, Mechaly AE, Monk JM, Nascimento ALT, Nelson KE, Palsson B, Peacock SJ, Picardeau M, Ricaldi JN, Thaipandungpanit J, Wunder EA, Yang JR, Yang XF, Zhang J-J, Vinetz JM (2016) What makes a bacterial species pathogenic? Comparative genomic analysis of the genus *Leptospira*. PLoS Negl Trop Dis 10:e0004403

Fox GE, Wisotzkey JD, Jurtshuk PJ (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. Int J Syst Bacteriol 42:166–170

Galtier N, Daubin V (2008) Dealing with incongruence in phylogenomic analyses. Philos Trans R Soc B Biol Sci 363:4023–4029

Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA 92:11317–11321

Gavini F, Holmes B, Izard D, Beji A, Bernigaud A, Jakubczak E (1989a) Numerical taxonomy of *Pseudomonas alcaligenes*, *P. pseudoalcaligenes, P. mendocina, P. stutzeri*, and related bacteria. Int J Syst Evol Microbiol 39:135–144

Gavini F, Mergaert J, Beji A, Mielcarek C, Izard D, Kersters K, De Ley J (1989b) Transfer of *Enterobacter agglomerans* (Beijerinck 1888) Ewing and Fife 1972 to *Pantoea* gen. nov. as *Pantoea agglomerans* comb. nov. and description of *Pantoea dispersa* sp. nov. Int J Syst Bacteriol 39:337–345

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, van de Peer Y, Vandamme P, Thompson FL, Swings J (2005) Re-evaluating prokaryotic species. Nat Rev 3:733–739

Glaeser SP, Kämpfer P (2015) Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol 38:237–245

Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. Mol Biol Evol 19:2226–2238

Gordon A, Hannon GJ (2010) Fastx-toolkit. FASTQ/A short-reads pre-processing tools. Unpublished http://hannonlab.cshl.edu/fastx_toolkit

Gueule D, Fourny G, Ageron E, Le Flèche-Matéos A, Vandenbogaert M, Grimont PAD, Cilas C (2015) *Pantoea coffeiphila* sp. nov., cause of the 'potato taste' of Arabica coffee from the African Great Lakes region. Int J Syst Evol Microbiol 65:23–29

Guindon SP, Dufayard J-FO, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321

Hacker J, Carniel E (2001) Ecological fitness, genomic islands and bacterial pathogenicity. EMBO Rep 2:376–381

Hacker JRH, Dobrindt U, Kurth R (2012) Genome plasticity and infectious diseases. ASM Press, Washington

Hall T (2011) BioEdit: an important software for molecular biology. GERF Bull Biosci 2:60–61

He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, Martin MJ, Holt KE, Seth-Smith HMB, Quail MA, Rance R, Brooks K, Churcher C, Harris D, Bentley SD, Burrows C, Clark L, Corton C, Murray V, Rose G, Thurston S, van Tonder A, Walker D, Wren BW, Dougan G, Parkhill J (2010) Evolutionary dynamics of *Clostridium difficile* over short and long time scales. Proc Natl Acad Sci USA 107:7527–7532

Heath TA, Hedtke SM, Hillis DM (2008) Taxon sampling and the accuracy of phylogenetic analyses. J Syst Evol 46:239–257

Hillis DM (1998) Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst Biol 47:3–8

Hong K-W, Gan HM, 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. J Bacteriol 194:6610

Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: the complexity hypothesis. Proc Natl Acad Sci USA 96:3801–3806

Jain R, Rivera MC, Moore JE, Lake JA (2002) Horizontal gene transfer in microbial genome evolution. Theor Popul Biol 61:489–495

Jeffroy O, Brinkmann H, Delsuc FDR, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22:225–231

Jolley KA, Bliss CM, Bennett JS, Bratcher HB, Brehony C, Colles FM, Wimalarathna H, Harrison OB, Sheppard SK, Cody AJ, Maiden MCJ (2012) Ribosomal multi-locus sequence typing: universal characterisation of bacteria from domain to strain. PhD, University of Oxford

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci: CABIOS 8:275–282

Kamber T, Smits THM, Rezzonico F, Duffy B (2012) Genomics and current genetic understanding of *Erwinia amylovora* and the fire blight antagonist *Pantoea vagans*. Trees 26:227–238

Kim J (1996) General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst Biol 45:363–374

Kim HJ, Lee JH, Kang BR, Rong X, McspaddenGardener BB, Ji HJ, Park C-S, Kim YC (2012) Draft genome sequence of *Pantoea ananatis* B1-9, a nonpathogenic plant growth-promoting bacterium. J Bacteriol 194:729

Klenk HP, Göker M (2010) En route to a genome-based classification of archaea and bacteria? Syst Appl Microbiol 33:175–182

Konstantinidis KT, Tiedje JM (2005) Towards a genome-based taxonomy for prokaryotes. J Bacteriol 187:6258–6264

Konstantinidis KT, Tiedje JM (2006) Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. Appl Environ Microbiol 72:7286–7293

Konstantinidis KT, Tiedje JM (2007) Prokaryotic taxonomy and phylogeny in the genomic era: advancements and challenges ahead. Curr Opin Microbiol 10:504–509

Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet 39:309–338

Koonin EV, Wolf YI (2006) Evolutionary systems biology: links between gene evolution and function. Curr Opin Biotechnol 17:481–487

Kuck P, Longo G (2014) FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool 11:81

Lan R, Reeves PR (2000) Intraspecies variation in bacterial genomes: the need for a species genome concept. Trends Microbiol 8:396–401

Lang JM, Darling AE, Eisen JA (2013) Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS ONE 8:e62510

Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: the case of the γ-proteobacteria. PLoS Biol 1:e19

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. J Biotechnol 188:7–8

Lukjancenko O, Wassenaar T, Ussery D (2012) Comparison of 61 sequenced *Escherichia coli* genomes. Microb Ecol 60:708–720

Ma H-W, Zeng A-P (2004) Phylogenetic comparison of metabolic capacities of organisms at genome level. Mol Phylogenet Evol 31:204–213

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*. Curr Microbiol 72:207–212

Makarova K, Slesarev A, Wolf Y, Sorokin A, Mirkin B, Koonin E, Pavlov A, Pavlova N, Karamychev V, Polouchine N, Shakhova V, Grigoriev I, Lou Y, Rohksar D, Lucas S, Huang K, Goodstein DM, Hawkins T, Plengvidhya V, Welker D, Hughes J, Goh Y, Benson A, Baldwin K, Lee JH, Díaz-Muñiz I, Dosti B, Smeianov V, Wechter W,

Barabote R, Lorca G, Altermann E, Barrangou R, Ganesan B, Xie Y, Rawsthorne H, Tamir D, Parker C, Breidt F, Broadbent J, Hutkins R, O'Sullivan D, Steele J, Unlu G, Saier M, Klaenhammer T, Richardson P, Kozyavkin S, Weimer B, Mills D (2006) Comparative genomics of the lactic acid bacteria. Proc Natl Acad Sci USA 103:15611–15616

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Mergaert J, Verdonck L, Kersters K (1993) Transfer of *Erwinia ananas* (synonym, *Erwinia uredovora*) and *Erwinia stewartii* to the genus *Pantoea* emend. as *Pantoea ananas* (Serrano 1928) comb. nov. and *Pantoea stewartii* (Smith 1898) comb. nov., respectively, and description of *Pantoea stewartii* subsp. *indologenes* subsp. nov. Int J Syst Bacteriol 43:162–173

Mitchell A, Mitter C, Regier JC (2000) More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of *Noctuoidea* (Insecta: *Lepidoptera*). Syst Biol 49:202–224

Nabhan AR, Sarkar IN (2012) The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Brief Bioinform 13:122–134

Naum M, Brown EW, Mason-Gamer RJ (2008) Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the enterobacteriaceae? J Mol Evol 66:630–642

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York

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. Stand Genom Sci 11:1–11

Philippe H, Forterre P (1999) The rooting of the universal tree of life is not reliable. J Mol Evol 49:509–523

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 Biol 9:e1000602

Pond SLK, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. In: Nielsen R (ed) Statistical methods in molecular evolution. Springer, New York

Popp A, Cleenwerck I, Iversen C, de Vos P, Stephan R (2010) *Pantoea gaviniae* sp. nov. and *Pantoea calida* sp. nov., isolated from infant formula and an infant formula production environment. Int J Syst Evol Microbiol 60:2786–2792

Prakash O, Nimonkar Y, Vaishampayan A, Mishra M, Kumbhare S, Josef N, Shouche YS (2015) *Pantoea intestinalis* sp. nov., isolated from the human gut. Int J Syst Evol Microbiol 65:3352–3358

Prasanna AN, Mehra S (2013) Comparative phylogenomics of pathogenic and non-pathogenic *Mycobacterium*. PLoS ONE 8:e71248

Price MN, Dehal PS, Arkin AP (2010) FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490

Rezzonico F, Smits TH, Montesinos E, Frey JE, Duffy B (2009) Genotypic comparison of *Pantoea agglomerans* plant and clinical strains. BMC Microbiol 9:204

Rezzonico F, Smits THM, Born Y, Blom J, Frey JE, Goesmann A, Cleenwerck I, de Vos P, Bonaterra A, Duffy B, Montesinos E (2016) *Erwinia gerundensis* sp. nov., a cosmopolitan epiphyte originally isolated from pome fruit trees. Int J Syst Evol Microbiol 66:1583–1592

Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci 106:19126–19131

Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. Proc Natl Acad Sci USA 95:6239–6244

Rong C, Ma Y, Wang S, Liu Y, Chen S, Huang B, Wang J, Xu F (2016) *Pantoea hericii* sp. nov., isolated from the fruiting bodies of *Hericium erinaceus*. Curr Microbiol 72:738–743

Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–333

Sarkar SF, Guttman DS (2004) Evolution of the core genome of *Pseudomonas syringae*, a highly clonal, endemic plant pathogen. Appl Environ Microbiol 70:1999–2012

Segata N, Huttenhower C (2011) Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. PLoS ONE 6:e24704

Smits THM, 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. J Bacteriol 192:6486–6487

Smits THM, Rezzonico F, Kamber T, Blom J, Goesmann A, Ishimaru CA, Frey JE, Stockwell VO, Duffy B (2011) Metabolic versatility and antibacterial metabolite biosynthesis are distinguishing genomic features of the fire blight antagonist *Pantoea vagans* C9-1. PLoS ONE 6:e22247

Smits THM, Rezzonico F, López MM, Blom J, Goesmann A, Frey JE, Duffy B (2013) Phylogenetic position and virulence apparatus of the pear flower necrosis pathogen *Erwinia piriflorinigrans* CFBP 5888T as assessed by comparative genomics. Syst Appl Microbiol 36:449–456

Staley JT (2006) The bacterial species dilemma and the genomic-phylogenetic species concept. Philos Trans R Soc B Biol Sci 361:1899–1909

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313

Swofford DL (2002) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland

Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599–607

Tambong JT, Xu R, Kaneza C-A, Nshogozabahizi J-C (2014) An in-depth analysis of a multilocus phylogeny identifies leuS as a reliable phylogenetic marker for the genus *Pantoea*. Evolut Bioinform Online 10:115–125

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis Version 6.0. Mol Biol Evol 30:2725–2729

Tanaka YK, Horie N, Mochida K, Yoshida Y, Okugawa E, Nanjo F (2015) *Pantoea theicola* sp. nov., isolated from black tea. Int J Syst Evol Microbiol 65:3313–3319

Tavaré S (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. Lect Math Life Sci 17:57–86

Tian H, Jing C (2014) Genome sequence of the aerobic arsenate-reducing bacterium *Pantoea* sp. Genome Announc, Strain IMH. doi:10.1128/genomeA.00267-14

Walterson AM, Stavrinides J (2015) *Pantoea*: insights into a highly versatile and diverse genus within the Enterobacteriaceae. FEMS Microbiol Rev 39:968–984

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

Wang X, Yang F, von Bodman SB (2011) The genetic and structural basis of two distinct terminal side branch residues in stewartan and amylovoran exopolysaccharides and their potential role in host adaptation. Mol Microbiol 83:195–207

Woese CR (2000) Interpreting the universal phylogenetic tree. Proc Natl Acad Sci 97:8392–8396

Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV (2001) Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol Biol 1:8

Wu Q, Du J, Zhuang G, Jing C (2013) Bacillus sp. SXB and *Pantoea* sp. IMH, aerobic As (V)-reducing bacteria isolated from arsenic-contaminated soil. J Appl Microbiol 114:713–721

Xia X, Xie Z (2001) DAMBE: software package for data analysis in molecular biology and evolution. J Hered 92:371–373

Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Mol Phylogenet Evol 26:1–7

West-Eberhard MJ (2003) Developmentalplasticity and evolution. Oxford University Press

Zhang Y, Qiu S (2015) Examining phylogenetic relationships of *Erwinia* and *Pantoea* species using whole genome sequence data. Antonie Van Leeuwenhoek 108:1037–1046

Zwickl DJ, Hillis DM (2002) Increased taxon sampling greatly reduces phylogenetic error. Syst Biol 51:588–598