Improved genome of *Agrobacterium radiobacter* type strain provides new taxonomic insight into *Agrobacterium* genomospecies 4

Han Ming Gan [1] , **Melvin VL Lee** [2] , **Michael A Savka** [Corresp. 3]

[1] Deakin Genomics Centre, Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia
[2] School of Science, Monash University, Petaling Jaya, Selangor, Malaysia
[3] College of Science, The Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, New York, United States

Corresponding Author: Michael A Savka
Email address: massbi@rit.edu

The reported *Agrobacterium radiobacter* DSM30174T genome is highly fragmented, hindering robust comparative genomics and genome-based taxonomic analysis. We re-sequenced the *Agrobacterium radiobacter* type strain obtained from NCPPB, generating a dramatically improved genome with high contiguity. In addition, we sequenced the genome of *Agrobacterium tumefaciens* B6T, enabling for the first time, a proper comparative genomics of these contentious *Agrobacterium* species. We provide concrete evidence that the previously reported *A. radiobacter* type strain genome (Accession Number: ASXY01) is contaminated which explains its abnormally large genome size and fragmented assembly. We propose that *Agrobacterium tumefaciens* be reclassified as *A. radiobacter* subsp. *tumefaciens* and that *A. radiobacter* retains it species status with the proposed name of *A. radiobacter* subsp. *radiobacter* . This proposal is based, first on the high pairwise genome-scale average nucleotide identity supporting the amalgamation of both *A. radiobacter* and *A tumefaciens* into a single genomospecies. Second, core genome alignment and phylogenomic analysis indicates *A. radiobacter* NCPPB3001 is sufficiently divergent from *A. tumefaciens* to propose two independent sub-clades. Third, *A. tumefaciens* demonstrates the genomic potential to synthesize the L configuration of fucose in its lipid polysaccharide, fostering its ability to colonize plant cells more effectively than *A. radiobacter* .

Improved genome of *Agrobacterium radiobacter* type strain provides new taxonomic insight into *Agrobacterium* genomospecies 4

Han Ming Gan[1,2,3], Melvin VJ Lee[3], Michael A. Savka[4*]

[1] Deakin Genomics Centre, Deakin University, Geelong, Victoria, Australia
[2] Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, Victoria, Australia
[3] School of Science, Monash University Malaysia, Petaling Jaya, Selangor, Malaysia
[4] The Thomas H. Gosnell School of Life Sciences, Rochester, NY 14623, USA

*Corresponding author
Michael A. Savka
The Thomas H. Gosnell School of Life Sciences,
85 Lomb Memorial Dr., Rochester, NY 14623 United States
Email address: massbi@rit.edu

ABSTRACT

The reported *Agrobacterium radiobacter* DSM30174^T genome is highly fragmented, hindering robust comparative genomics and genome-based taxonomic analysis. We re-sequenced the *Agrobacterium radiobacter* type strain obtained from NCPPB, generating a dramatically improved genome with high contiguity. In addition, we sequenced the genome of *Agrobacterium tumefaciens* B6^T, enabling for the first time, a proper comparative genomics of these contentious *Agrobacterium* species. We provide concrete evidence that the previously reported *A. radiobacter* type strain genome (Accession Number: ASXY01) is contaminated which explains its abnormally large genome size and fragmented assembly. We propose that *Agrobacterium tumefaciens* be reclassified as *A. radiobacter* subsp. *tumefaciens* and that *A. radiobacter* retains it species status with the proposed name of *A. radiobacter* subsp. *radiobacter*. This proposal is based, first on the high pairwise genome-scale average nucleotide identity supporting the amalgamation of both *A. radiobacter* and *A tumefaciens* into a single genomospecies. Second, core genome alignment and phylogenomic analysis indicates *A. radiobacter* NCPPB3001 is sufficiently divergent from *A. tumefaciens* to propose two independent sub-clades. Third, *A. tumefaciens* demonstrates the genomic potential to synthesize the L configuration of fucose in its lipid polysaccharide, fostering its ability to colonize plant cells more effectively than *A. radiobacter.*

Keywords: *Agrobacterium*, *Agrobacterium radiobacter, Agrobacterium tumefaciens,* type strain, lipopolysaccharide, Ti plasmid, average nucleotide identity, phylogenomics

INTRODUCTION

The taxonomy and phylogeny of the genus *Agrobacterium* has proven to be complex and controversial. Bacteria of the genus *Agrobacterium* have been grouped into six species based on the disease phenotype associated, in part, with the resident disease-inducing plasmid: 1) *A. tumefaciens* causes crown gall on dicotyledonous plants, stone fruit and nut trees; 2) *A. rubi* causes crown gall on raspberries; 3) *A. vitis* causes gall formation only on grapevines; 4) *A. rhizogenes* causes hairy root proliferative disease on diverse plant hosts; 5) *A. larrymmoori* causes aerial tumors on weeping fig; and 6) *A. radiobacter* that fails to cause crown gall or disease on plants (Bouzar & Jones 2001; Conn 1942; Kerr & Panagopoulos 1977; Panagopoulos et al. 1978; Riker et al. 1930; Starr & Weiss 1943; Süle 1978). An alternative classification approach grouped *Agrobacterium* organisms into three biovars based on physiological and biochemical properties without consideration of disease phenotype (Keane et al. 1970; Kerr & Panagopoulos 1977; Panagopoulos et al. 1978). The species and biovar classification schemes do not coincide well, in a large part, because of the disease-inducing plasmids, tumor-inducing (pTi) and hairy root-inducing (pRi), are readily transmissible plasmids (Young et al. 2001).

Much controversy has existed in the three main components of bacterial taxonomy which include classification, nomenclature and identification. The genus *Agrobacterium* is a prime example with many proposals and oppositions regarding the amalgamation of *Agrobacterium* and *Rhizobium* over the last three or four decades (Farrand et al. 2003; Gaunt et al. 2001; Young et al. 2001; Young et al. 2003). However, more recent studies appear to favor the preservation of the genus *Agrobacterium* backed by strong genetic and genomic evidence (Gan & Savka 2018; Ramírez-Bahena et al. 2014). Within the genus *Agrobacterium*, the taxonomic status of *A. radiobacter* and *A. tumefaciens* remains contentious (Sawada et al. 1993; Young 2008; Young et al. 2006). *Agrobacterium radiobacter* (originally proposed as *Bacillus radiobacter*) is a non-pathogenic soil bacterium associated with nitrogen utilization isolated more than a century ago in 1902 (Beijerinck & van Delden 1902; Conn 1942). On the other hand, *A. tumefaciens* (previously *Bacterium tumefaciens*) is a plant pathogen capable of inducing tumorigenesis (Smith & Townsend 1907). However, the descriptive assignment for *A. tumefaciens* was later found to be contributed by a set of genes located on the large Ti plasmid that can be lost (Gordon & Christie 2014). In other words, the curing of Ti plasmid in *A. tumefaciens* will change its identity to the non-pathogenic species, *A. radiobacter*. Furthermore, comparative molecular analysis based on single-copy housekeeping genes also supports the close relatedness of *A. radiobacter* and *A. tumefaciens*, blurring the taxonomic boundaries between these species (Mousavi et al. 2015; Shams et al. 2013). As taxa are reclassified into different populations that do not conform to the characteristics of the original description, the given names lose their significant and descriptive importance. Consistent with the Judicial Commission according to the Rules of the International Code of Nomenclature of Bacteria, Tindall (2014) concluded that "the combination of *A. radiobacter* has priority over the combination *A. tumefaciens* when the two are treated as members of the same species based on *the principle of priority* as applied to the corresponding specific epithets". However, given that *A. tumefaciens* has been more widely studied than *A. radiobacter* due to its strong relevance to agriculture (Bourras et al. 2015), it remains unclear but interesting to see if the broader scientific community will obey this rule by adopting the recommended species name change in future studies.

To our knowledge, a detailed comparative genomics analysis of *A. radiobacter* and *A. tumefaciens* type strains has yet been reported to date despite their genome availability (Zhang et al. 2014). The high genomic relatedness of both type strains was briefly mentioned by Kim and

Gan (2017) through whole genome alignment and pairwise nucleotide identity calculation from homologous regions. However, evidence is now mounting that the *A. radiobacter* DSM 30147^T reported by Zhang et al. (2014) is contaminated, warranting immediate investigation (Jeong et al. 2016). The assembled genome is nearly 7 MB, the largest among *Agrobacterium* currently sequenced at that time with up to 6,853 predicted protein-coding genes contained in over 600 contigs. At sequencing depth of nearly 200×, its genome assembly is unusually fragmented even for a challenging microbial genome (Utturkar et al. 2017). Furthermore, the phylogenomic placement of *A. radiobacter* DSM 30147^T based on this genome assembly has been questionable as evidenced by its basal position and substantially longer branch length relative to other members of the genomospecies 4 (Gan & Savka 2018). The overly fragmented nature of this assembly also precludes fruitful comparative genomics focusing on gene synteny analysis. More importantly, analysis done on a contaminated assembly but with the assumption that it is not, will likely lead to incorrect biological interpretations (Allnutt et al. 2018).

In this study, we sequenced the whole genome of *A. radiobacter* using a type strain that was sourced from the National Collection of Plant Pathogenic Bacteria (NCPPB). We produced a contiguous genome assembly exhibiting genomic statistics that are more similar to other *Agrobacterium* assembled genomes. We show here, through comparative genomics and phylogenetics, that the previously assembled *A. radiobacter* DSM 30147^T genome contains substantial genomic representation from another *Agrobacterium* sp. isolated and sequenced by the same lab, consistent with our initial suspicion of strain contamination. Using the newly assembled genome for subsequent comparative analysis, we provide conclusive genomic evidence that *A. radiobacter* DSM 30147^T and *A. tumefaciens* B6^T are the same genomic species. However, strain DSM 30147^T should not be considered as a merely non-tumorigenic strain of *A. tumefaciens* as substantial genomic variation exists between these two type strains notably in the nucleotide sugar metabolism pathway that may contribute to their ecological niche differentiation.

MATERIALS & METHODS

DNA extraction and whole genome sequencing

Approximately 10 bacterial colonies were scrapped using a sterile P200 pipette tip from a 3-day-old nutrient agar culture and resuspended in lysis buffer with proteinase K (Sokolov 2000) followed by incubation at 56 ^{o}C for 3 hours. DNA purification was performed as previously described. The extracted DNA was normalized to 0.2 ng/uL and prepared using the Nextera XT library preparation according to the manufacturer's instructions. Sequencing of the constructed library was performed on a MiSeq desktop sequencer located at the Monash University Malaysia Genomics Facility (2 × 250 bp run configuration).

De novo assembly and genome completeness assessment

Raw paired-end reads were adapter-trimmed using Trimmomatic v0.36 (Bolger et al. 2014) followed by error-correction and de novo assembly using Spades Assembler v3.9 (Bankevich et al. 2012). Genome completeness was assessed with BUSCOv3 (Rhizobiales database) (Waterhouse et al. 2017).

Protein clustering

Gene prediction used Prodigal v2.6 (Hyatt et al. 2010). Clustering of the predicted proteins performed with CD-HIT using the settings "-C 0.95, -T 0.8" (Li & Godzik 2006). Identification of unique and shared clusters were done using basic unix commands e.g. csplit, grep, sort and uniq. Protein sequences were aligned with mafft v7.3 using the the most accurate setting (--localpair --maxiterate 1000) followed by phylogenetic tree construction via IqTree v1.65 with optimized model (Kalyaanamoorthy et al. 2017; Nguyen et al. 2014). Visualization and annotation of phylogenetic trees used Figtree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Pan-genome construction and phylogenomics

Whole genome sequences were annotated with Prokka v1.1 using the default setting (Seemann 2014). The Prokka-generated gff files were used as the input for Roary to calculate the pan-genome (Page et al. 2015). Maximum likelihood tree construction of the core-genome alignment and tree visualization used FastTree2 v2.1.10 (-nt -gtr) and FigTree v 1.4.3, respectively (Price et al. 2010).

Detection and visualization of Ti plasmid

Genome sequences of each member of the genomospecies 4 except for the problematic DSM37014 strain were used as the query for blastN search (-evalue 1e-100) against the octopine-type Ti plasmid. The result of the similarity search was subsequently visualized in Blast Ring Image Generator (BRIG) v0.95.

Genome annotation and KEGG pathway reconstruction

Whole genome sequences of *A. tumefaciens* B6^T and *A. radiobacter* NCPPB 3001^T were submitted to the online server GhostKoala (Kanehisa et al. 2016) for annotation and the annotated genomes were subsequently used to reconstruct KEGG pathways in the same webserver. Identification of proteins with TIGRFAM signatures of interest used HMMsearch v3.1b2 with the option "--cut_tc" activated to filter for only protein hits passing the TIGRFAM trusted cutoff values (Johnson et al. 2010).

RESULTS

An improved *Agrobacterium radiobacter* type strain genome

The newly assembled genome of *A. radiobacter* type strain that was sourced from the National Collection of Plant Pathogenic Bacteria (NCPPB) is approximately 30% smaller than the first reported *A. radiobacter* DSM 30147 genome with 96% less contigs (22 vs 612), 20-fold longer N50 (480 kb vs 23 kb) and assembled length that is much more similar to other *Agrobacterium* spp. (Table 1). In addition, it is near-complete with 685 out of 686 BUSCO Rhizobiale single-copy genes detected as either partial or complete with minimal evidence of contamination as indicated by the near absence of duplicated single-copy gene(<0.1%). On the contrary, the current DSM 30147 genome is missing 25.1% of the single copy gene with up to 34.8% duplication rate. At the time of this manuscript writing, another genome of *A. radiobacter* type strain that was sourced from another culture collection centre e.g. the Belgian Coordinated

Collections of Microorganisms has been deposited in the NCBI wgs database (*A. radiobacter* LM140, Table 1) with assembly statistics that are highly similar to the type strain genome reported in this study.

The inflated genome size of *Agrobacterium radiobacter* DSM 30147(T) is due to technical errors

Instead of sharing a recent common ancestor as would be expected for a recently duplicated gene, the duplicated single copy genes coding for seryl-tRNA synthetase in A. radiobacter DSM 30147^T were placed in two distinct clusters with one affiliated to genomospecies 4 and the other affiliated to genomospecies 7 (Figure 1A). Such an unexpected clustering pattern raises the suspicion of genome assembly from two or more non-clonal bacterial strains. In addition, by performing comparison at the genome-scale based on whole proteome clustering of *A. radiobacter* DSM 30147^T /NCPPB 3001^T (Previous study, GCF_000421945 ; This study, GCF_001541305), *A.* sp. TS43 (unpublished, GCF_001526605) and *A. tumefaciens* B6 (GCF_001541315), we observed a high number of proteins that were exclusively shared between Zhang et al *A. radiobacter* DSM 30147 and *A.* sp. TS43 belonging to genomospecies 7. Coincidentally, despite not sharing the same Bioproject ID, the whole genomes of strains DSM 30147^T and TS43 were sequenced by the Zhang et al., and submitted to NCBI on the same date, 30-May-2013, hinting strain contamination during sample processing in the lab.

Genome-scale average nucleotide identity calculation supports the amalgamation of *A.radiobacter* and *A.tumefaciens* into a single genomospecies

Single gene tree shows that *A. radiobacter* NCPPB 3001^T and *A. tumefaciens* $B6^T$ belong to the genomospecies 4 clade (Figure 1A), corroborating with the PhyloPhlAN phylogenomic tree that was constructed based on the alignment of 400 universal single-copy proteins (blue colored branches in Supplemental Figure 1). The pairwise average nucleotide (ANI) among strains within this clade is consistently more than 95% further supporting their affiliation to the same genomospecies (Figure 2). As expected, pairwise ANI of less than 92% was observed when they were compared with strains from genomospecies 7 (strains RV3 and Zutra 3/1). A 100% pairwise ANI was observed between *A. radiobacter* type strains that were sourced from NCPPB and LMG. In addition, non-type strains B140/95 and CFBP5621 also exhibit a strikingly high pairwise ANI (>99%) to the type strains of *A. tumefaciens* and *A. radiobacter*, respectively, leading to the formation of sub-clusters within genomospecies 4 (Figure 2).

Is *A. radiobacter* a non-tumorigenic strain of *A. tumefaciens*?

A majority of the currently sequenced strains from genomospecies 4 are non-tumorigenic as evidenced by the near complete lack of genomic region with significant nucleotide similarity to the octopine-type Ti reference plasmid (Figure 3). Of the 14 genomes analyzed, only strains $B6^T$ and B140/95 exhibit a complete coverage of the Ti plasmid with near 100% sequence identity while strain 186 shows hits mainly to the essential gene clusters of a Ti plasmid such as the *vir* gene cluster (black rings and gene labels in Figure 3) at a substantially lower sequence identity (50%<x<90%) (Figure 3), suggesting that it may be harboring a dissimilar variant of Ti plasmid e.g. different opine type. In addition, although lacking hits to the virulence gene of the Ti plasmid, the *tra* and *trb* clusters involved in plasmid conjugal transfer are present in strains Kerr 14, CCNWGS0286 and UNC420CL41Cvi. Despite belonging to the same genomospecies, core genome alignment and phylogenomic analysis indicates that *A. radiobacter* NCPPB3001^T is sufficiently divergent from A. tumefaciens B6^T leading to their separation into two distinct sub-clades (Figure 4A). This is also resonated by their different sub-cluster placement in the pairwise ANI heatplot (Figure 3). Furthermore, strains from both subclades could be broadly differentiated by the set of core accessory genes that they harbor (Figure 4B). Therefore, even though *A. radiobacter* does not harbor a Ti plasmid, it cannot be considered as a non-tumorigenic strain of *A. tumefaciens* given multiple lines of evidence indicating its substantial genomic divergence from *A. tumefaciens*.

Agrobacterium genomospecies 4 strains differ in their genomic potential for nucleotide sugar metabolism

A manual inspection of the core accessory genomes uniquely shared by *A. tumefaciens* strains B6 and B140/95 identified a homolog cluster containing GDP-L-fucose synthase (EC 1.1.1.271) that is involved in the enzymatic production of GDP-L-fucose from GDP-4-dehydro-6-deoxy-D-mannose and NADH (Figure 4A and Table 2). As expected, the genes coding for this enzyme and GDP-mannose 4,6-dehydratase involved in the conversion of GDP-alpha-D-mannose to GDP-4-dehydro-6-deoxy-D-mannose, are absent in the *A. radiobacter* NCPPB3001 genome (Figure 4B). Intriguingly, HMMsearch scan revealed the presence of two protein hits to the TIGR01479 HMM profile in *A. tumefaciens* B6 that corresponds to D-mannose 1,6-phosphomutase (EC 5.4.2.8) required for the synthesis of D-mannose 6-phosphate. In addition to strain B6, its close relative, strain B140/95, and a more distantly related strain Kerr14 also harbor two copies of this gene. However, one of the D-mannose 1,6-phosphomutases in strain Kerr14 is more divergent with a lower TIGRFAM HMM sequence score (Table 2). Furthermore, it exhibits less than 70% protein identity to the *A. tumefaciens* B6 and B140/95 homologs, forming a private protein cluster in the pan-genome (data not shown). Individual comparison of the reconstructed KEGG pathways in *A. radiobacter* and *A. tumefaciens* B6 revealed another stark contrast in the anabolism of dTDP-L-rhamnose which is commonly found in the O-antigen of LPS in gram-negative bacteria (Figure 4C and D). Surprisingly, the entire enzyme set required for the generation of dTDP-L-rhamnose from D-glucose-phosphate (Table 2) is absent in *A. tumefaciens* B6, suggesting that this common nucleotide sugar may be absent from the LPS O-antigen of strain B6.

DISCUSSION

We re-sequenced the genome of *Agrobacterium radiobacter* type strain using strain directly obtained from NCPPB. The gDNA was prepped and sequenced in a genomics facility that routinely sequences mostly decapod crustacean mitogenomes (Gan et al. 2016a; Gan et al. 2016b; Tan et al. 2015) and occasionally microbial genomes (Gan et al. 2015; Gan et al. 2014; Wong et al. 2014) without prior history of processing any member from the *Agrobacterium* genomospecies 4 clade. The assembled *A. radiobacter* genome reported in this study exhibits assembly statistics that are consistent with a high-quality draft genome such as high genome completeness and contiguity, near-zero contamination/duplication and comparable genome size to other closely related strains (Gan et al. 2018; Parks et al. 2015). Furthermore, given the improved contiguity and dramatic reduction in the number of contigs of this newly assembled draft genome, we recommend using this genome in place of the previously published draft genome for future *Agrobacterium* comparative studies.

The distinct separation of *Agrobacterium* genomospecies 4 and 7 at 95% ANI cutoff corroborates with the previously established "genomic yardstick" for species differentiation (Konstantinidis & Tiedje 2005; Richter & Rosselló-Móra 2009). Using this percentage cutoff, the ANI approach has been successfully used to provide a near "black-and-white" pattern of species separation in even some of the most diverse bacterial genera such as *Pseudomonas, Arcobacter* and *Stenotrophomonas* (Pérez-Cataluña et al. 2018; Tran et al. 2017; Vinuesa et al. 2018). Given the increasing evidence highlighting the robustness and reliability of the ANI approach in species delineation, the pairwise ANI between *A. tumefaciens* and *A. radiobacter* type strains that is at least 2.5% higher than the 95% cutoff value is rigorous evidence that they belong to the same genomospecies, effectively serving as the final nail in the coffin for the decade-long debate on their taxonomic status. The amalgamation of *A. radiobacter* and *A. tumefaciens* into a single species have been repeatedly suggested in the past few years but was complicated by the special status of *A. tumefaciens* as the type species of the genus *Agrobacterium* despite the priority that *A. radiobacter* has over *A. tumefaciens* as it was isolated and described 3 years before *A. tumefaciens* (Young et al. 2001; Young et al. 2003). Despite sharing numerous morphological and biochemical features, differences in genomic features such as pairwise ANI, phylogenomic clustering and core accessory gene contents do exist among members in *Agrobacterium* genomospecies 4 that can facilitate the identification of genotypic and phenotypic variants for delimiting sub-species relationships (Brenner et al. 2015; Jezbera et al. 2011; Meier-Kolthoff et al. 2014; Tan et al. 2013).

To date the LPS for both type strains have been determined (De Castro et al. 2004; De Castro et al. 2002). In stark contrast to *A. radiobacter*, the *A. tumefaciens* LPS consists of D-arabinose and L-fucose that have yet been reported to date in another members of the genus *Agrobacterium (De Castro et al. 2002)*. The presence of the L configuration of fucose is considered to be rare even among plant pathogenic bacteria but may be associated with the ability of *A. tumefaciens* to colonize or bind to wounded plant cell (Lippincott et al. 1977; Whatley et al. 1976; Whatley & Spiess 1977). It has been previously shown that the LPS of *A. tumefaciens* but not *A. radiobacter* can bind to the plant cells thus providing protection against subsequent infection by pathogenic strains (Whatley et al. 1976). The presence and absence of nucleotide sugars in the O-chain constituent of LPS in both type strains corroborates with their observed genomic potential in the nucleotide sugar metabolism pathway thus underscoring the utility of comparative genomics in facilitating the prediction of microbial host range and ecological niche (Klosterman et al. 2011). For example, the absence of L-rhamnose and L-fucose in the LPS of *A. tumefaciens* B6 and *A. radiobacter* DSM30147, respectively, is consistent with the lack of genes coding for enzymes involved with the particular nucleotide sugar metabolism. Generation of *Agrobacterium tumefaciens* B6 LPS mutant via targeted gene deletion (Kaczmarczyk et al. 2012) or the classical but more laborious transposon mutagenesis approach followed by characterization of the LPS mutant host-range and phytopathogenicity will be instructive (Gan et al. 2011; Reuhs et al. 2005).

Our current genomic sampling indicates that the Ti plasmid appears to be restricted to the *A. tumefaciens* subclade. The maintenance of the Ti plasmid is metabolically taxing given its large size (Barker et al. 1983; Glick 1995). Even if the Ti plasmid was conjugally transfer for example, to *A. radiobacter*, the inability of *A. radiobacter* to colonize plant host as evidenced by its LPS incompatibility will not confer an advantage to the new plasmid host in a natural environment (Thomashow et al. 1980). Furthermore, in the absence of high density AHL signals which is required to trigger Ti plasmid conjugation (Fuqua & Winans 1994; Pappas 2008; Zhang et al. 2002), the newly acquired Ti plasmid in *A. radiobacter* may be cured in its natural soil habitat after a few generations. Although the spontaneous transfer of the Ti plasmid from tumorigenic *A. tumefaciens* to *A. radiobacter* K84 has been reported previously, strain K84 was re-classified based on a recent core gene analysis to *Rhizobium rhizogenes* K84 (Velázquez et al. 2010; Vicedo et al. 1996), reiterating the pervasive taxonomic inconsistency within the genus *Agrobacterium* that may have confound previous biological interpretations (De Ley et al. 1966; Lindström et al. 1995; Young 2008). Given that a large majority of *Agrobacterium* genetics was performed during the pre-NGS era (Gan & Savka 2018), it remains unknown as to how many *A. tumefaciens* and *A. radiobacter* strains have been molecularly misclassified due to their high genomic relatedness.

The inability to accurately identify plasmid and chromosomal-derived contigs among the draft genomes means that some of the core accessory genes among tumorigenic strains may be plasmid-derived and should be treated with caution as the low-copy-number Ti-plasmid is prone to curing in the absence of AHL signals. Despite the value of complete genome assembly in enabling the accurate partitioning of plasmid and chromosomal genomic region (Arredondo-Alonso et al. 2017), the representation of complete *Agrobacterium* genomes in current database is still very low as a majority of the genomes were assembled from short Illumina reads that cannot effectively span repetitive region (Wibberg et al. 2011; Wood et al. 2001). Furthermore, most *Agrobacterium* strains harbor multiple large plasmids that further complicate short-read-only assembly graph (Kado et al. 1981; Lowe et al. 2009; Shao et al. 2018). However, the advent of high throughput long-read sequencing that can span large repetitive region in recent years is likely going to overcome this limitation allowing a more accurate depiction of microbial pangenome (Gan et al. 2012; Gan et al. 2017; Schmid et al. 2018a; Schmid et al. 2018b). Future hybrid genome assemblies (Illumina and Nanopore/PacBio reads) of members from genomospecies 4, particularly *A. tumefaciens* and *A. radiobacter* strains with comprehensive metadata and reliable phenotypic information, will be instructive.

CONCLUSIONS

Despite belonging to the same genomospecies, *A. tumefaciens* and *A. radiobacter* are by no means clonal at the chromosomal level and instead demonstrate sufficient genomic characters that qualify their separation into two sub-species. In addition, the difference in the LPS profile among two type strains will have implications to host specificity leading to geographical separation. In the spirit of preserving the naming of both species but at the same time respecting the taxonomic jurisdiction for strain priority, we propose *A. tumefaciens* to be reclassified as *A.*

radiobacter subsp. tumefaciens and for *A. radiobacter* to retains its species status with the proposed name of *A. radiobacter subsp. radiobacter.*

ACKNOWLEDGEMENTS

Allnutt T, Yan CZY, Crowley TM, and Gan HM. 2018. Commentary: Genome Sequence of Vibrio parahaemolyticus VP152 Strain Isolated From Penaeus indicus in Malaysia. *Frontiers in microbiology* 9. 10.3389/fmicb.2018.00865

Arredondo-Alonso S, Willems RJ, van Schaik W, and Schürch AC. 2017. On the (im) possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microbial genomics* 3.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, and Prjibelski AD. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* 19:455-477.

Barker R, Idler K, Thompson D, and Kemp J. 1983. Nucleotide sequence of the T-DNA region from theA grobacterium tumefaciens octopine Ti plasmid pTi15955. *Plant Molecular Biology* 2:335-350.

Beijerinck M, and van Delden A. 1902. On a colourless bacterium, whose carbon food comes from the athmosphere. *Koninklijke Nederlandse Akademie van Wetenschappen Proceedings Series B Physical Sciences* 5:398-413.

Bolger AM, Lohse M, and Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114-2120.

Bourras S, Rouxel T, and Meyer M. 2015. Agrobacterium tumefaciens Gene Transfer: How a Plant Pathogen Hacks the Nuclei of Plant and Nonplant Organisms. *Phytopathology* 105:1288-1301. 10.1094/phyto-12-14-0380-rvw

Bouzar H, and Jones JB. 2001. Agrobacterium larrymoorei sp. nov., a pathogen isolated from aerial tumours of Ficus benjamina. *International Journal of Systematic and Evolutionary Microbiology* 51:1023-1026.

Brenner DJ, Staley JT, and Krieg NR. 2015. Classification of Procaryotic Organisms and the Concept of Bacterial Speciation. *Bergey's Manual of Systematics of Archaea and Bacteria*:1-9.

Conn HJ. 1942. Validity of the Genus Alcaligenes. *Journal of Bacteriology* 44:353-360.

De Castro C, Bedini E, Garozzo D, Sturiale L, and Parrilli M. 2004. Structural Determination of the O-Chain Moieties of the Lipopolysaccharide Fraction from Agrobacterium radiobacter DSM 30147. *European Journal of Organic Chemistry* 2004:3842-3849.

De Castro C, De Castro O, Molinaro A, and Parrilli M. 2002. Structural determination of the O-chain polysaccharide from Agrobacterium tumefaciens, strain DSM 30205. *European journal of biochemistry* 269:2885-2888.

De Ley J, Bernaerts M, Rassel A, and Guilmot J. 1966. Approach to an improved taxonomy of the genus Agrobacterium. *Microbiology* 43:7-17.

Farrand SK, Van Berkum PB, and Oger P. 2003. Agrobacterium is a definable genus of the family Rhizobiaceae. *Int J Syst Evol Microbiol* 53:1681-1687. 10.1099/ijs.0.02445-0

Fuqua WC, and Winans SC. 1994. A LuxR-LuxI type regulatory system activates Agrobacterium Ti plasmid conjugal transfer in the presence of a plant tumor metabolite. *Journal of Bacteriology* 176:2796-2806.

Gan HM, Chew TH, Tay Y-L, Lye SF, and Yahya A. 2012. Genome sequence of Hydrogenophaga sp. strain PBC, a 4-aminobenzenesulfonate-degrading bacterium. *Journal of Bacteriology* 194:4759-4760.

Gan HM, Gan HY, Ahmad NH, Aziz NA, Hudson AO, and Savka MA. 2015. Whole genome sequencing and analysis reveal insights into the genetic structure, diversity and evolutionary relatedness of luxI and luxR homologs in bacteria belonging to the

Sphingomonadaceae family. *Frontiers in Cellular and Infection Microbiology* 4. 10.3389/fcimb.2014.00188
Gan HM, Ibrahim Z, Shahir S, and Yahya A. 2011. Identification of genes involved in the 4-aminobenzenesulfonate degradation pathway of Hydrogenophaga sp. PBC via transposon mutagenesis. *FEMS microbiology letters* 318:108-114. 10.1111/j.1574-6968.2011.02245.x
Gan HM, Lee MVJ, and Savka MA. 2018. High-Quality Draft Genome Sequence of the Type Strain of Allorhizobium vitis, the Primary Causal Agent of Grapevine Crown Gall. *Microbiol Res Announc* 7:e01045-01018.
Gan HM, Lee YP, and Austin CM. 2017. Nanopore long-read guided complete genome assembly of Hydrogenophaga intermedia, and genomic insights into 4-aminobenzenesulfonate, p-aminobenzoic acid and hydrogen metabolism in the genus Hydrogenophaga. *Frontiers in microbiology* 8:1880.
Gan HM, and Savka MA. 2018. One More Decade of Agrobacterium Taxonomy.
Gan HM, Tan MH, and Austin CM. 2016a. The complete mitogenome of the red claw crayfish Cherax quadricarinatus (Von Martens, 1868)(Crustacea: Decapoda: Parastacidae). *Mitochondrial DNA Part A* 27:385-386.
Gan HM, Tan MH, Eprilurahman R, and Austin CM. 2016b. The complete mitogenome of Cherax monticola (Crustacea: Decapoda: Parastacidae), a large highland crayfish from New Guinea. *Mitochondrial DNA Part A* 27:337-338. 10.3109/19401736.2014.892105
Gan HY, Gan HM, Savka MA, Triassi AJ, Wheatley MS, Smart LB, Fabio ES, and Hudson AO. 2014. Whole-Genome Sequences of 13 Endophytic Bacteria Isolated from Shrub Willow (*Salix*) Grown in Geneva, New York. *Genome Announcements* 2. 10.1128/genomeA.00288-14
Gaunt M, Turner S, Rigottier-Gois L, Lloyd-Macgilp S, and Young J. 2001. Phylogenies of atpD and recA support the small subunit rRNA-based classification of rhizobia. *International Journal of Systematic and Evolutionary Microbiology* 51:2037-2048.
Glick BR. 1995. Metabolic load and heterologous gene expression. *Biotechnology advances* 13:247-261.
Gordon JE, and Christie PJ. 2014. The Agrobacterium Ti Plasmids. *Microbiol Spectr* 2. 10.1128/microbiolspec.PLAS-0010-2013
Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, and Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* 11:119.
Jeong H, Pan J-G, and Park S-H. 2016. Contamination as a major factor in poor Illumina assembly of microbial isolate genomes. *bioRxiv*. 10.1101/081885
Jezbera J, Jezberová J, Brandt U, and Hahn MW. 2011. Ubiquity of Polynucleobacter necessarius subspecies asymbioticus results from ecological diversification. *Environmental microbiology* 13:922-931.
Johnson LS, Eddy SR, and Portugaly E. 2010. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC bioinformatics* 11:431.
Kaczmarczyk A, Vorholt JA, and Francez-Charlot A. 2012. Markerless Gene Deletion System for Sphingomonads. *Applied and Environmental Microbiology* 78:3774-3777. 10.1128/AEM.07347-11
Kado C, amp, and Liu S. 1981. Rapid procedure for detection and isolation of large and small plasmids. *Journal of Bacteriology* 145:1365-1373.
Kalyaanamoorthy S, Minh BQ, Wong TK, von Haeseler A, and Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods* 14:587.
Kanehisa M, Sato Y, and Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *Journal of molecular biology* 428:726-731.

Keane P, Kerr A, and New P. 1970. Crown gall of stone fruit II. Identification and nomenclature of Agrobacterium isolates. *Australian Journal of Biological Sciences* 23:585-596.
Kerr A, and Panagopoulos C. 1977. Biotypes of Agrobacterium radiobacter var. tumefaciens and their biological control. *Journal of Phytopathology* 90:172-179.
Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, Chen Z, Henrissat B, Lee Y-H, Park J, Garcia-Pedrajas MD, Barbara DJ, Anchieta A, de Jonge R, Santhanam P, Maruthachalam K, Atallah Z, Amyotte SG, Paz Z, Inderbitzin P, Hayes RJ, Heiman DI, Young S, Zeng Q, Engels R, Galagan J, Cuomo CA, Dobinson KF, and Ma L-J. 2011. Comparative Genomics Yields Insights into Niche Adaptation of Plant Vascular Wilt Pathogens. *PLoS Pathogens* 7:e1002137. 10.1371/journal.ppat.1002137
Konstantinidis KT, and Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences* 102:2567-2572.
Li W, and Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658-1659.
Lindström K, Van Berkum P, Gillis M, Martinez E, Novikova N, and Jarvis B. 1995. Report from the roundtable on Rhizobium taxonomy. *Nitrogen Fixation: Fundamentals and Applications*: Springer, 807-810.
Lippincott BB, Whatley MH, and Lippincott JA. 1977. Tumor induction by Agrobacterium involves attachment of the bacterium to a site on the host plant cell wall. *Plant physiology* 59:388-390.
Lowe N, Gan HM, Chakravartty V, Scott R, Szegedi E, Burr TJ, and Savka MA. 2009. Quorum-sensing signal production by Agrobacterium vitis strains and their tumor-inducing and tartrate-catabolic plasmids. *FEMS microbiology letters* 296:102-109.
Meier-Kolthoff JP, Hahnke RL, Petersen J, Scheuner C, Michael V, Fiebig A, Rohde C, Rohde M, Fartmann B, and Goodwin LA. 2014. Complete genome sequence of DSM 30083 T, the type strain (U5/41 T) of Escherichia coli, and a proposal for delineating subspecies in microbial taxonomy. *Standards in genomic sciences* 9:2.
Mousavi SA, Willems A, Nesme X, de Lajudie P, and Lindstrom K. 2015. Revised phylogeny of Rhizobiaceae: proposal of the delineation of Pararhizobium gen. nov., and 13 new species combinations. *Syst Appl Microbiol* 38:84-90. 10.1016/j.syapm.2014.12.003
Nguyen L-T, Schmidt HA, von Haeseler A, and Minh BQ. 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution* 32:268-274.
Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, and Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 31:3691-3693.
Panagopoulos C, Psallidas P, and Alivizatos A. 1978. Studies on biotype 3 of Agrobacterium radiobacter var. tumefaciens. Proceedings of the IVth International Conference on Plant Pathogenic Bacteria Vol 1: Sta. Path. Veg. Phytobact. p 221-228.
Pappas KM. 2008. Cell–cell signaling and the Agrobacterium tumefaciens Ti plasmid copy number fluctuations. *Plasmid* 60:89-107.
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, and Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*:gr. 186072.186114.
Pérez-Cataluña A, Salas-Massó N, Dieguez ABL, Balboa S, Lema A, Romalde JL, and Figueras MJ. 2018. Revisiting the taxonomy of the genus Arcobacter: getting order from the chaos. *Frontiers in microbiology* 9:2077.
Price MN, Dehal PS, and Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* 5:e9490.
Ramírez-Bahena MH, Vial L, Lassalle F, Diel B, Chapulliot D, Daubin V, Nesme X, and Muller D. 2014. Single acquisition of protelomerase gave rise to speciation of a large and

diverse clade within the Agrobacterium/Rhizobium supercluster characterized by the presence of a linear chromid. *Molecular phylogenetics and evolution* 73:202-207. https://doi.org/10.1016/j.ympev.2014.01.005

Reuhs BL, Relić B, Forsberg LS, Marie C, Ojanen-Reuhs T, Stephens SB, Wong C-H, Jabbouri S, and Broughton WJ. 2005. Structural Characterization of a Flavonoid-Inducible Pseudomonas aeruginosa A-Band-Like O Antigen of Rhizobium sp. Strain NGR234, Required for the Formation of Nitrogen-Fixing Nodules. *Journal of Bacteriology* 187:6479-6487. 10.1128/jb.187.18.6479-6487.2005

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

Riker A, Banfield W, Wright W, Keitt G, and Sagen HE. 1930. Studies on infectious hairy root of nursery apple trees. *Journal of Agricultural Research* 41.

Sawada H, Ieki H, Oyaizu H, and Matsumoto S. 1993. Proposal for rejection of Agrobacterium tumefaciens and revised descriptions for the genus Agrobacterium and for Agrobacterium radiobacter and Agrobacterium rhizogenes. *Int J Syst Bacteriol* 43:694-702. 10.1099/00207713-43-4-694

Schmid M, Frei D, Patrignani A, Schlapbach R, Frey JE, Remus-Emsermann MN, and Ahrens CH. 2018a. Pushing the limits of de novo genome assembly for complex prokaryotic genomes harboring very long, near identical repeats. *bioRxiv*:300186.

Schmid M, Muri J, Melidis D, Varadarajan AR, Somerville V, Wicki A, Moser A, Bourqui M, Wenzel C, and Eugster-Meier E. 2018b. Comparative genomics of completely sequenced Lactobacillus helveticus genomes provides insights into strain-specific genes and resolves metagenomics data down to the strain level. *Frontiers in microbiology* 9:63.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30:2068-2069.

Shams M, Vial L, Chapulliot D, Nesme X, and Lavire C. 2013. Rapid and accurate species and genomic species identification and exhaustive population diversity assessment of Agrobacterium spp. using recA-based PCR. *Syst Appl Microbiol* 36:351-358. 10.1016/j.syapm.2013.03.002

Shao S, Zhang X, van Heusden GPH, and Hooykaas PJ. 2018. Complete sequence of the tumor-inducing plasmid pTiChry5 from the hypervirulent Agrobacterium tumefaciens strain Chry5. *Plasmid* 96:1-6.

Smith EF, and Townsend CO. 1907. A plant-tumor of bacterial origin. *Science* 25:671-673.

Sokolov EP. 2000. An improved method for DNA isolation from mucopolysaccharide-rich molluscan tissues. *Journal of Molluscan Studies* 66:573-575.

Starr M, and Weiss J. 1943. Growth of phytopathogenic bacteria in a synthetic asparagin medium. *Phytopathology* 33:314-318.

Süle S. 1978. Biotypes of Agrobacterium tumefaciens in Hungary. *Journal of Applied Bacteriology* 44:207-213.

Tan JL, Khang TF, Ngeow YF, and Choo SW. 2013. A phylogenomic approach to bacterial subspecies classification: proof of concept in Mycobacterium abscessus. *BMC genomics* 14:879.

Tan MH, Gan HM, Schultz MB, and Austin CM. 2015. MitoPhAST, a new automated mitogenomic phylogeny tool in the post-genomic era with a case study of 89 decapod mitogenomes including eight new freshwater crayfish mitogenomes. *Molecular phylogenetics and evolution* 85:180-188.

Thomashow M, Panagopoulos C, Gordon M, and Nester E. 1980. Host range of Agrobacterium tumefaciens is determined by the Ti plasmid. *Nature* 283:794.

Tran PN, Savka MA, and Gan HM. 2017. In-silico taxonomic classification of 373 genomes reveals species misidentification and new genospecies within the genus Pseudomonas. *Frontiers in microbiology* 8:1296.

Utturkar SM, Klingeman DM, Hurt RA, and Brown SD. 2017. A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies. *Frontiers in microbiology* 8:1272. 10.3389/fmicb.2017.01272

Velázquez E, Palomo JL, Rivas R, Guerra H, Peix A, Trujillo ME, García-Benavides P, Mateos PF, Wabiko H, and Martínez-Molina E. 2010. Analysis of core genes supports the reclassification of strains Agrobacterium radiobacter K84 and Agrobacterium tumefaciens AKE10 into the species Rhizobium rhizogenes. *Systematic and applied microbiology* 33:247-251.

Vicedo B, López MJ, Asíns MJ, and López MM. 1996. Spontaneous Transfer of the Ti Plasmid of Agrobacterium tumefaciens and the Nopaline Catabolism Plasmid of A. radiobacter Strain K84. *Phytopathology* 86:528-534.

Vinuesa P, Ochoa-Sánchez LE, and Contreras-Moreira B. 2018. GET_PHYLOMARKERS, a software package to select optimal orthologous clusters for phylogenomics and inferring pan-genome phylogenies, used for a critical geno-taxonomic revision of the genus Stenotrophomonas. *Frontiers in microbiology* 9.

Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, Kriventseva EV, and Zdobnov EM. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution* 35:543-548.

Whatley M, Bodwin J, Lippincott B, and Lippincott J. 1976. Role of Agrobacterium cell envelope lipopolysaccharide in infection site attachment. *Infection and immunity* 13:1080-1083.

Whatley MH, and Spiess LD. 1977. Role of bacterial lipopolysaccharide in attachment of Agrobacterium to moss. *Plant physiology* 60:765-766.

Wibberg D, Blom J, Jaenicke S, Kollin F, Rupp O, Scharf B, Schneiker-Bekel S, Sczcepanowski R, Goesmann A, and Setubal JC. 2011. Complete genome sequencing of Agrobacterium sp. H13-3, the former Rhizobium lupini H13-3, reveals a tripartite genome consisting of a circular and a linear chromosome and an accessory plasmid but lacking a tumor-inducing Ti-plasmid. *Journal of biotechnology* 155:50-62.

Wong YM, Juan JC, Gan HM, and Austin CM. 2014. Draft Genome Sequence of Clostridium perfringens Strain JJC, a Highly Efficient Hydrogen Producer Isolated from Landfill Leachate Sludge. *Genome Announcements* 2. 10.1128/genomeA.00064-14

Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, and Almeida NF. 2001. The genome of the natural genetic engineer Agrobacterium tumefaciens C58. *Science* 294:2317-2323.

Young J, Kuykendall L, Martinez-Romero E, Kerr A, and Sawada H. 2001. A revision of Rhizobium Frank 1889, with an emended description of the genus, and the inclusion of all species of Agrobacterium Conn 1942 and Allorhizobium undicola de Lajudie et al. 1998 as new combinations: Rhizobium radiobacter, R. rhizogenes, R. rubi, R. undicola and R. vitis. *International Journal of Systematic and Evolutionary Microbiology* 51:89-103.

Young J, Kuykendall L, Martinez-Romero E, Kerr A, and Sawada H. 2003. Classification and nomenclature of Agrobacterium and Rhizobium–a reply to Farrand et al.(2003). *International Journal of Systematic and Evolutionary Microbiology* 53:1689-1695.

Young JM. 2008. Agrobacterium—Taxonomy of plant-pathogenic Rhizobium species. *Agrobacterium: From biology to biotechnology*: Springer, 183-220.

Young JM, Pennycook SR, and Watson DR. 2006. Proposal that Agrobacterium radiobacter has priority over Agrobacterium tumefaciens. Request for an opinion. *Int J Syst Evol Microbiol* 56:491-493. 10.1099/ijs.0.64030-0

Zhang H-B, Wang L-H, and Zhang L-H. 2002. Genetic control of quorum-sensing signal turnover in Agrobacterium tumefaciens. *Proceedings of the National Academy of Sciences* 99:4638-4643.

Zhang L, Li X, Zhang F, and Wang G. 2014. Genomic analysis of Agrobacterium radiobacter DSM 30147(T) and emended description of A. radiobacter (Beijerinck and van Delden 1902) Conn 1942 (Approved Lists 1980) emend. Sawada et al. 1993. *Stand Genomic Sci* 9:574-584. 10.4056/sigs.4688352

Figure 1

Phylogenetic and genomic evidence indicating contamination in the published *A. radiobacter* DSM 30147T genome.

(A) Maximum likelihood phylogenetic tree of seryl-tRNA synthetases from *Agrobacterium* genomospecies 4 and 7. Codes after the tildes are contigs containing the corresponding homologs. Node labels indicate ultra-fast bootstrap support value and branch length indicates number of substitutions per site. Duplicated homologs in the problematic *A. radiobacter* DSM 30147 genome were colored red. (B) Venn diagram of the core proteome of selected *Agrobacterium* strains from genomospecies 4. Numbers in the overlapping regions indicate the number of proteins that were shared by two or more groups at 95% protein identity cutoff.

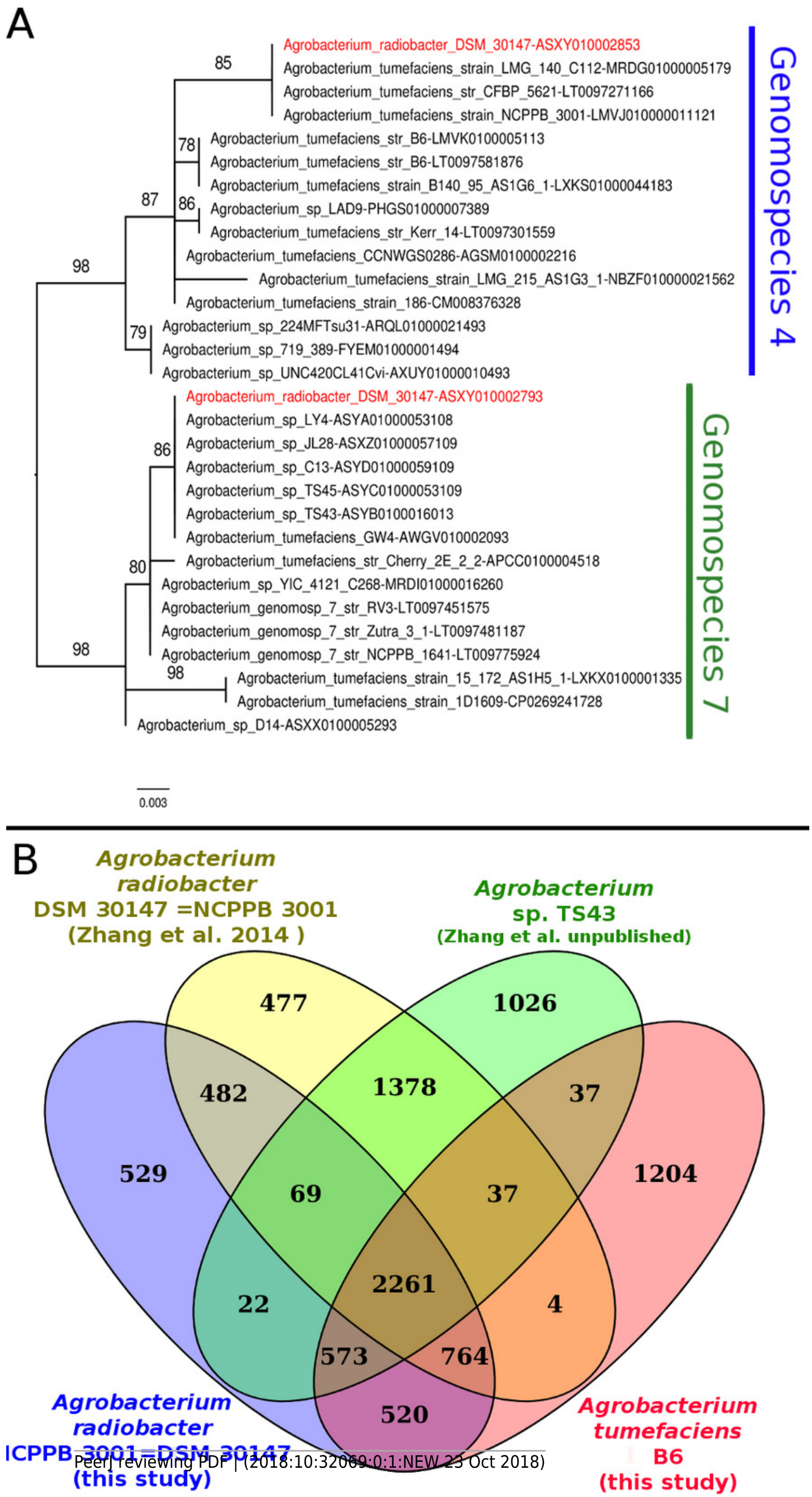
A
Agrobacterium_radiobacter_DSM_30147-ASXY010002853
Agrobacterium_tumefaciens_strain_LMG_140_C112-MRDG01000005179
Agrobacterium_tumefaciens_str_CFBP_5621-LT0097271166
Agrobacterium_tumefaciens_strain_NCPPB_3001-LMVJ010000011121
Agrobacterium_tumefaciens_str_B6-LMVK0100005113
Agrobacterium_tumefaciens_str_B6-LT0097581876
Agrobacterium_tumefaciens_strain_B140_95_AS1G6_1-LXKS01000044183
Agrobacterium_sp_LAD9-PHGS01000007389
Agrobacterium_tumefaciens_str_Kerr_14-LT0097301559
Agrobacterium_tumefaciens_CCNWGS0286-AGSM0100002216
Agrobacterium_tumefaciens_strain_LMG_215_AS1G3_1-NBZF010000021562
Agrobacterium_tumefaciens_strain_186-CM008376328
Agrobacterium_sp_224MFTsu31-ARQL01000021493
Agrobacterium_sp_719_389-FYEM01000001494
Agrobacterium_sp_UNC420CL41Cvi-AXUY01000010493
Agrobacterium_radiobacter_DSM_30147-ASXY010002793
Agrobacterium_sp_LY4-ASYA01000053108
Agrobacterium_sp_JL28-ASXZ01000057109
Agrobacterium_sp_C13-ASYD01000059109
Agrobacterium_sp_TS45-ASYC01000053109
Agrobacterium_sp_TS43-ASYB0100016013
Agrobacterium_tumefaciens_GW4-AWGV010002093
Agrobacterium_tumefaciens_str_Cherry_2E_2_2-APCC0100004518
Agrobacterium_sp_YIC_4121_C268-MRDI01000016260
Agrobacterium_genomosp_7_str_RV3-LT0097451575
Agrobacterium_genomosp_7_str_Zutra_3_1-LT0097481187
Agrobacterium_genomosp_7_str_NCPPB_1641-LT009775924
Agrobacterium_tumefaciens_strain_15_172_AS1H5_1-LXKX0100001335
Agrobacterium_tumefaciens_strain_1D1609-CP0269241728
Agrobacterium_sp_D14-ASXX0100005293
85
78
87
86
98
79
86
80
98
98
Genomospecies 4
Genomospecies 7
0.003
B
Agrobacterium radiobacter DSM 30147 =NCPPB 3001 (Zhang et al. 2014)
Agrobacterium sp. TS43 (Zhang et al. unpublished)
Agrobacterium radiobacter
ICPPB 3001=DSM 30147
(this study)
Agrobacterium tumefaciens B6 (this study)
477
1026
482
1378
37
529
69
37
1204
2261
22
4
573
764
520

Figure 2

A heatmap showing the hierarchical clustering of *Agrobacterium* strains based on genomic distance.

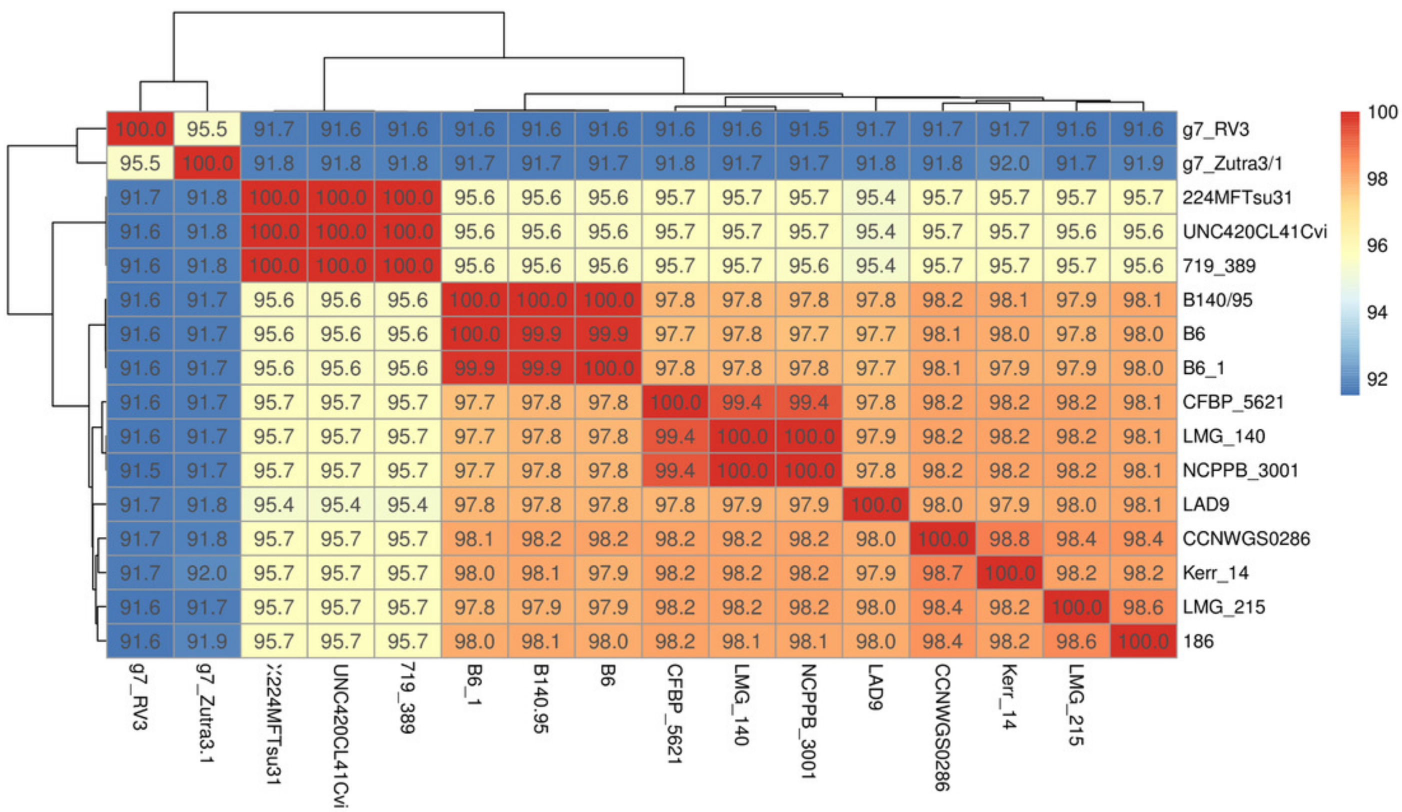

Figure 3

Prevalence and sequence conservation of the octopine-type Ti plasmid among *Agrobacterium* genomospecies 4.

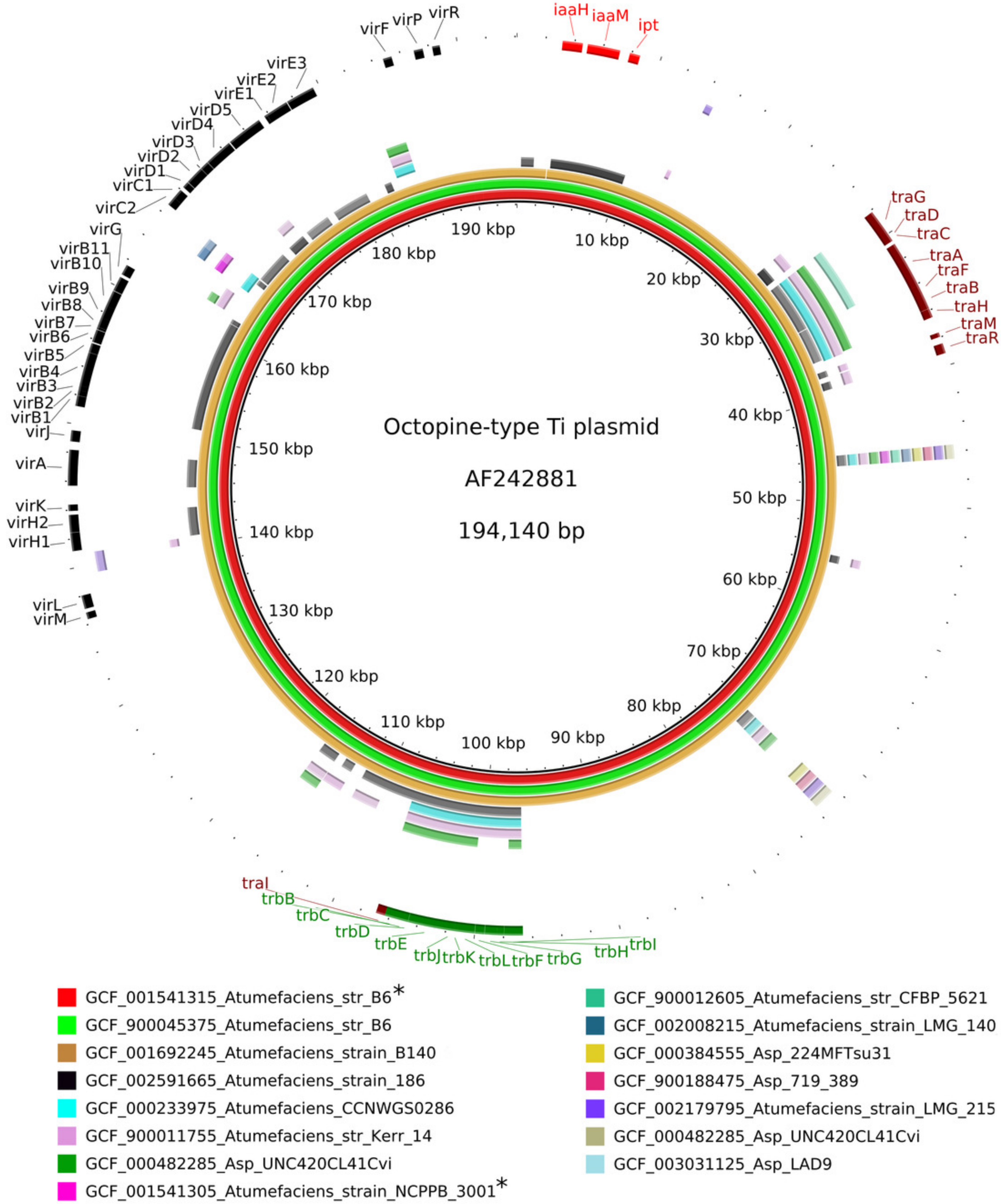
Octopine-type Ti plasmid
AF242881
194,140 bp
10 kbp
20 kbp
30 kbp
40 kbp
50 kbp
60 kbp
70 kbp
80 kbp
90 kbp
100 kbp
110 kbp
120 kbp
130 kbp
140 kbp
150 kbp
160 kbp
170 kbp
180 kbp
190 kbp
iaaH
iaaM
ipt
traG
traD
traC
traA
traF
traB
traH
traM
traR
traI
trbB
trbC
trbD
trbE
trbJ
trbK
trbL
trbF
trbG
trbH
trbI
virF
virP
virR
virE3
virE2
virE1
virD5
virD4
virD3
virD2
virD1
virC1
virC2
virG
virB11
virB10
virB9
virB8
virB7
virB6
virB5
virB4
virB3
virB2
virB1
virJ
virA
virK
virH2
virH1
virL
virM
GCF_001541315_Atumefaciens_str_B6*
GCF_900045375_Atumefaciens_str_B6
GCF_001692245_Atumefaciens_strain_B140
GCF_002591665_Atumefaciens_strain_186
GCF_000233975_Atumefaciens_CCNWGS0286
GCF_900011755_Atumefaciens_str_Kerr_14
GCF_000482285_Asp_UNC420CL41Cvi
GCF_001541305_Atumefaciens_strain_NCPPB_3001*
GCF_900012605_Atumefaciens_str_CFBP_5621
GCF_002008215_Atumefaciens_strain_LMG_140
GCF_000384555_Asp_224MFTsu31
GCF_900188475_Asp_719_389
GCF_002179795_Atumefaciens_strain_LMG_215
GCF_000482285_Asp_UNC420CL41Cvi
GCF_003031125_Asp_LAD9

Figure 4

Genomic divergence among genomospecies 4 strains.

(A) Unrooted maximum likelihood tree constructed based on the core genome alignment. Branch length and node labels indicate number of substitutions per site and FastTree2 SH-like support values, respectively. Putative subclades were colored blue, red and purple. (B) Distribution of gene clusters that are unique to a strain or exclusively shared by some strains. Taxa were arranged based on their clustering pattern in the phylogenomic tree.

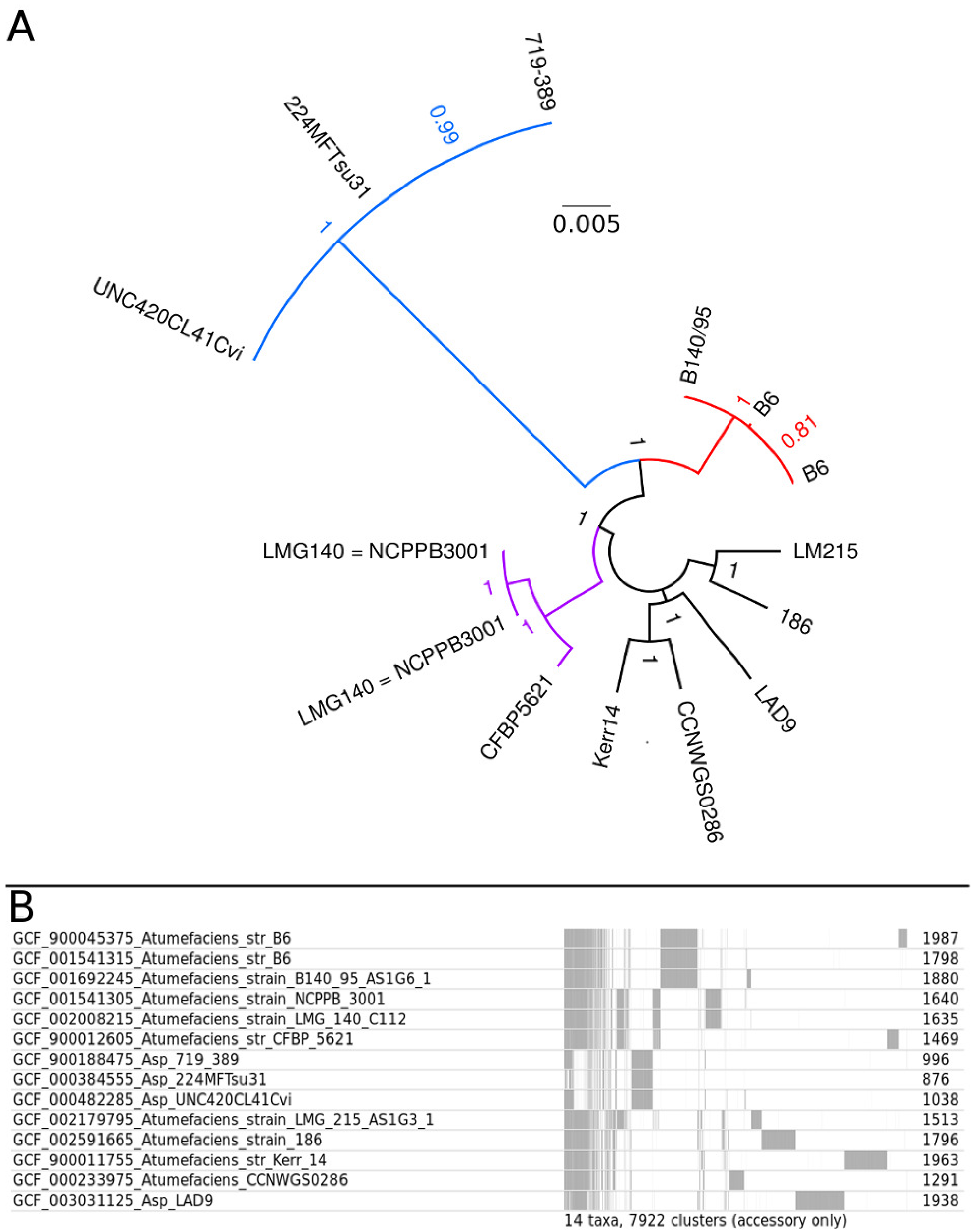
A
719-389
0.99
224MFTsu31
1
0.005
UNC420CL41Cvi
B140/95
1
B6
0.81
B6
1
1
LMG140 = NCPPB3001
LM215
1
1
186
1
LMG140 = NCPPB3001
1
1
LAD9
CFBP5621
Kerr14
CCNWGS0286
B
GCF_900045375_Atumefaciens_str_B6 1987
GCF_001541315_Atumefaciens_str_B6 1798
GCF_001692245_Atumefaciens_strain_B140_95_AS1G6_1 1880
GCF_001541305_Atumefaciens_strain_NCPPB_3001 1640
GCF_002008215_Atumefaciens_strain_LMG_140_C112 1635
GCF_900012605_Atumefaciens_str_CFBP_5621 1469
GCF_900188475_Asp_719_389 996
GCF_000384555_Asp_224MFTsu31 876
GCF_000482285_Asp_UNC420CL41Cvi 1038
GCF_002179795_Atumefaciens_strain_LMG_215_AS1G3_1 1513
GCF_002591665_Atumefaciens_strain_186 1796
GCF_900011755_Atumefaciens_str_Kerr_14 1963
GCF_000233975_Atumefaciens_CCNWGS0286 1291
GCF_003031125_Asp_LAD9 1938
14 taxa, 7922 clusters (accessory only)

Figure 5

KEGG pathway of nucleotide sugar metabolism associated with *Agrobacterium* lipopolysaccharide synthesis.

(A) and (B), genomic potential of *A. tumefaciens* B6 and *A. radiobacter* DSM 30147,respectively, in the biosynthesis of dTDP-L-rhamnose. (C) and (D), genomic potential of *A. tumefaciens* B6 and *A. radiobacter* DSM 30147,respectively, in the biosynthesis of GDP-L-Fucose. Numbers in boxes indicate Enzyme Commission numbers. White and green boxes indicate absence and presence of the corresponding enzymes, respectively, based on GhostKoala annotation.

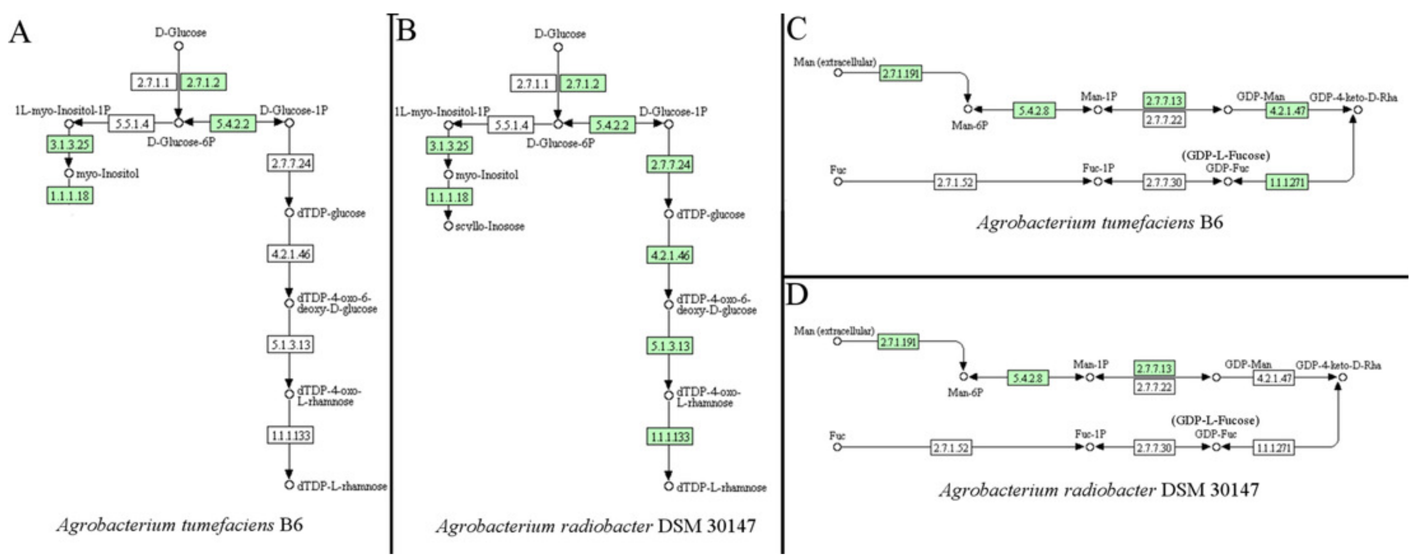

Table 1 (on next page)

Genome statistics of publicly available *Agrobacterium* genomospecies 4 whole genome sequences

Table 1: Genome statistics of publicly available Agrobacterium genomospecies 4 whole genome sequences

Assembly accession	Strain	Isolation Source	Country	Size	GC%	# Contig
GCF_900045375	B6	Apple Gall (Iowa)	USA	5.8	59.07	4
GCF_001541315	B6	Apple Gall (Iowa)	USA	5.6	59.32	52
GCF_001692245	B140/95	Peach/Almond Rootstock	USA	5.7	59.23	45
GCF_002179795	LMG 215	Humulus lupulus gall (USA)	USA	5.4	59.48	33
GCF_000233975	CCNWGS0286	R. pseudoacacia nodules	China	5.2	59.53	49
GCF_900011755	Kerr 14= LMG 15 = CFBP 5761	Soil around Prunus dulcis	Australia	5.9	59.04	5
GCF_002591665	186	English Walnut gall	California	5.7	59.42	22
GCF_002008215	LMG 140 = NCPPB 3001 =CFBP 5522= DSM 30147	saprobic soil	Germany	5.5	59.34	22
GCF_000421945	LMG 140 = NCPPB 3001 =CFBP 5522= DSM 30147	saprobic soil	Germany	7.17	59.86	612
GCF_001541305	LMG 140 = NCPPB 3001 =CFBP 5522= DSM 30147	saprobic soil	Germany	5.5	59.36	22
GCF_900012605	CFBP 5621	Lotus corniculata, root tissue commensal	France	5.4	59.32	3
GCF_003031125	LAD9 (CGMCC No. 2962)	landfill leachate treatment system	China	5.9	59.13	49
GCF_000384555	224MFTsu31	rhizosphere of L. luteus in Hungary, formerly R. lupini H13-3	USA	4.8	59.73	21
GCF_900188475	719_389	Rhizosphere and endosphere of Arabidopsis thaliana.	USA	4.9	59.73	18

GCF_000384555	UNC420CL41Cvi	Plant associated	USA	5	59.69	18

1

Table 2 (on next page)

Identification of *Agrobacterium* proteins with TIGRFAM domains involved in the biosynthesis of nucleotide sugar.

Numbers indicate bit scores calculated based on protein alignment to the model with higher scores indicating stronger and more significant hits.

Table 2. Identification of *Agrobacterium* proteins with TIGRFAM domains involved in the biosynthesis of nucleotide sugar. Numbers indicate bit scores calculated based on protein alignment to the model with higher scores indicating stronger and more significant hits.

Assembly ID	Strain	TIGR01479 (EC 5.4.2.8)		TIGR01472 (EC 4.2.1.47)	TIGR01207 (EC 2.7.7.24)	TIGR01181 (EC 4.2.1.46)	TIGR01221 (EC 5.1.3.13)	TIGR01214 (EC 1.1.1.133)
		1st hit	2nd hit					
GCF_900045375	B6	690.2	566.6	589.5				
GCF_001541315	B6	690.2	566.6	589.5				
GCF_001692245	B140/95	690.2	566.6	589.5				
GCF_900011755	Kerr14	691.3	690.2	428.6*				
GCF_001541305	NCPPB3001	690.2			494.6	488.5	215.4	331.5
GCF_002008215	LMG140	690.2			494.6	488.5	215.4	331.5
GCF_900012605	CFBP5621	689.3			494.6	489.5	215.4	331.5
GCF_002591665	186	689.3			494.6	488.5	215.4	331.8
GCF_003031125	LAD9	688.5			494.4	487.9	215.4	329.9
GCF_000233975	CCNWGS	644.8			494.6	487.5	215.4	331.8
GCF_002179795	LMG215	690.2						
GCF_000384555	224MFTsu31	644.8						
GCF_000482285	UNC420CL41Cvi	644.8						
GCF_900188475	719_389	687.5						

*Formed a separate protein cluster from the rest of genomospecies 4 GDP-mannose-4,6-dehydratase orthologs (<70% pairwise protein identity)