Reconstructing the Phylogeny of *Corynebacteriales* while Accounting for Horizontal Gene Transfer

Nilson Da Rocha Coimbra[1,2], Aristoteles Goes-Neto[2], Vasco Azevedo[2], and Aïda Ouangraoua[1,*]

[1]Department of Computer Science, University of Sherbrooke, Quebec, Canada

[2]Programa Interunidades de Pós-graduação em Bioinformática, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

*Corresponding author: E-mail: aida.ouangraoua@usherbrooke.ca.

Accepted: March 17, 2020

Abstract

Horizontal gene transfer is a common mechanism in Bacteria that has contributed to the genomic content of existing organisms. Traditional methods for estimating bacterial phylogeny, however, assume only vertical inheritance in the evolution of homologous genes, which may result in errors in the estimated phylogenies. We present a new method for estimating bacterial phylogeny that accounts for the presence of genes acquired by horizontal gene transfer between genomes. The method identifies and corrects putative transferred genes in gene families, before applying a gene tree-based summary method to estimate bacterial species trees. The method was applied to estimate the phylogeny of the order *Corynebacteriales*, which is the largest clade in the phylum *Actinobacteria*. We report a collection of 14 phylogenetic trees on 360 *Corynebacteriales* genomes. All estimated trees display each genus as a monophyletic clade. The trees also display several relationships proposed by past studies, as well as new relevant relationships between and within the main genera of *Corynebacteriales*: *Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*, and *Gordonia*. An implementation of the method in Python is available on GitHub at https://github.com/UdeS-CoBIUS/EXECT (last accessed April 2, 2020).

Key words: phylogeny estimation, bacteria, horizontal gene transfer, *Corynebacterium*, *Mycobacterium*.

Introduction

One of the major discoveries in the 20th century is the bacterial production of antibiotics, which are useful in treating bacterial infections (Bister et al. 2004; Fair and Tor 2014). The ongoing evolution of bacteria, however, contributes to the appearance of new bacterial species, including new antibiotic-resistant pathogenic species (Fischbach and Walsh 2009). In terms of genome structure, bacterial species differ from each other in the content and arrangement of genes in their genomes, which results from genome rearrangement, gene duplication, gene loss, and horizontal gene transfer (HGT) events (Gogarten et al. 2002). In particular, HGT has been shown to be a primary force underlying antibiotic-resistance and virulent genes spreading in Bacteria (Ruiz et al. 2011; Zhi et al. 2017).

HGT is the transfer of genetic material through a process different from vertical inheritance (Soucy et al. 2015). The modules of genetic transfer are usually genes, but it was also shown that HGT can occur at the level of protein domains (Chan et al. 2009). The prevalence of HGT events in bacterial evolution limits the use of phylogenetic methods that assume only vertical inheritance evolutionary events. Traditionally, alignments of sequences of 16S rRNA genes have been used to estimate bacterial phylogenies and study bacterial diversity. This approach relies on the assumption that 16S rRNA genes constitute essential genes that are only vertically inherited. Several studies have, however, reported evidence for HGT of 16S rRNA genes (Yap et al. 1999; Schouls et al. 2003; Kitahara and Miyazaki 2013; Miyazaki et al. 2017). Moreover, the identification and classification bacterial species based solely on 16S rRNA genes often lead to errors in phylogenetic estimations (Rajendhran and Gunasekaran 2011). The reason could be the intragenomic heterogeneity in bacterial rRNA as well as the presence of mosaicism and multiple copies of 16S rRNA genes in genomes, which may result from HGT events (Klappenbach et al. 2000; Schouls et al. 2003). In this context, the main contribution of this work is a method for estimating bacterial phylogenies with sets of gene families but without assuming only vertical inheritance in the evolution of gene families.

Phylogenetic reconstruction usually relies on two steps: first, the identification of groups of orthologous sequences in genomes, and second, the construction of a tree explaining the evolution within orthology groups by vertical inheritance (Felsenstein 1985). Therefore, computing accurate orthology groups in the first step is a prerequisite for reconstructing accurate phylogenies in the second step. For the second step, phylogeny methods can be classified into three main approaches: alignment-based, gene order-based, and gene tree-based methods (Wolf et al. 2002).

Alignment-based methods infer the species tree based on a concatenation of multiple sequence alignments on the orthology groups. This approach has been widely used because it scales well to a large number of orthology groups and species (Rokas et al. 2003; Ciccarelli et al. 2006). Nonetheless, alignment-based methods do not allow for accounting for the impact of genome content and structure evolution in estimating species diversity (Saitou and Nei 1987). Another class of alignment-based methods are whole genome single-nucleotide polymorphism-based methods that start by removing signals from recombination and then build a species tree using whole genome alignments, single-nucleotide polymorphisms, and maximum-likelihood (ML) approaches (Castillo-Ramírez et al. 2012; Comas et al. 2013).

Gene order-based methods infer a species tree based on the difference between genomes in terms of gene content and arrangement (Sankoff and Blanchette 1998; Bourque et al. 2004; Belda et al. 2005). They allow for accounting for the evolution of gene content and arrangement as well as gene conservation or splitting. Gene order-based methods are suitable for reconstructing phylogenies of closely related species (Moret et al. 2013). Nevertheless, this approach is limited by the complexity in scaling up to large data sets and the lack of a well-defined model of gene order evolution (Moret et al. 2001).

Gene tree-based methods consist in using a set of gene trees—one for each orthology group—in order to estimate a species tree that could explain the evolution of gene families within the species tree (Suyama and Bork 2001). Such methods are currently in limited use in estimating bacterial phylogeny because they are very sensitive to the presence of erroneous genes in orthology groups caused by HGT events, leading to a disruption in the phylogenetic signals (Ravenhall et al. 2015). Thus, the detection and discarding of transferred genes from orthology groups is a prerequisite for using gene tree-based methods for estimating bacterial phylogeny.

Computational methods for HGT detection can be classified into two main approaches: parametric methods and comparative methods. Parametric methods are intragenomic, and exploit sequence composition changes along a genome sequence to infer putative HGT regions. Comparative methods are intergenomic and include alignment-based and phylogeny-based methods. Alignment-based methods—such as MobilomeFINDER (Ou et al. 2007)—make use of the alignment between closely related genomes to infer HGT. Phylogeny-based methods—such as NOTUNG (Chen et al. 2000)—exploit the inconsistencies between gene trees and species trees to infer HGT (Lerat et al. 2005; Ravenhall et al. 2015; Jeong et al. 2019). On the one hand, parametric methods—such as IslandPath-DIMOB (Bertelli and Brinkman 2018)—take advantage of relying solely on genome sequences by an intrinsic analysis. They achieve average recall rates with high precision rates. On the other hand, comparative methods are limited by their requirement of an accurate species tree, which, in turn, is challenging to build in the presence of HGT (Lasek-Nesselquist et al. 2012). Nevertheless, when a preliminary, partially resolved species tree is available, the congruence of a gene tree with this species tree can be used to correct the misclassification of some genes as transferred genes inferred by parametric methods.

This article presents a gene tree-based method accounting for HGT events in estimating bacterial phylogenies (fig. 1). After collecting the input data set, which consists of genome sequences with the locations of their coding DNA sequences (CDSs) representing genes (Step 1), the method starts by detecting putative transferred genes (PTGs) with a parametric method (Step 2). PTGs are identified using an intragenomic genomic island (GI) detection method in order to avoid the circular argument of detecting HGT using a species tree, and then removing HGT form gene trees to compute a species tree. In parallel, genes are clustered into homology groups based on their CDS similarities (Step 3). Subsequently, putative orthology groups, containing a single gene per genome, are used to build a preliminary, partially resolved species tree with an alignment-based phylogenetic method (Step 4). The preliminary species tree is used to correct misclassified PTGs in homology groups with a phylogenetic approach. The latter consists in comparing the gene tree of each homology group with the species tree in order to compare the phylogenetic position of PTG in the two trees. A PTG whose location induces no HGT in the reconciliation between the gene tree and the species tree is reclassified as a vertically inherited gene (Step 5). Lastly, the remaining transferred genes are removed from homology groups. The latter are used to build gene trees and the final species trees using phylogenetic gene tree-based methods (Step 6).

The pipeline was applied to estimate the phylogeny of *Corynebacteriales*, the largest clade of the phylum *Actinobacteria* in terms of number of available genomes. The input data set comprised 360 complete genome sequences obtained in Step 1, and the output consisted of 10 distinct phylogenetic trees on the 360 genomes that were estimated with 2 gene tree-based phylogenetic methods, ASTRID (Vachaspati and Warnow 2015) and ASTRAL II (Mirarab and Warnow 2015) in Step 6. The similarity of the estimated phylogenies was compared by computing the percentage of conserved clades between each pair of trees. The final phylogeny of *Corynebacteriales* was obtained by computing consensus

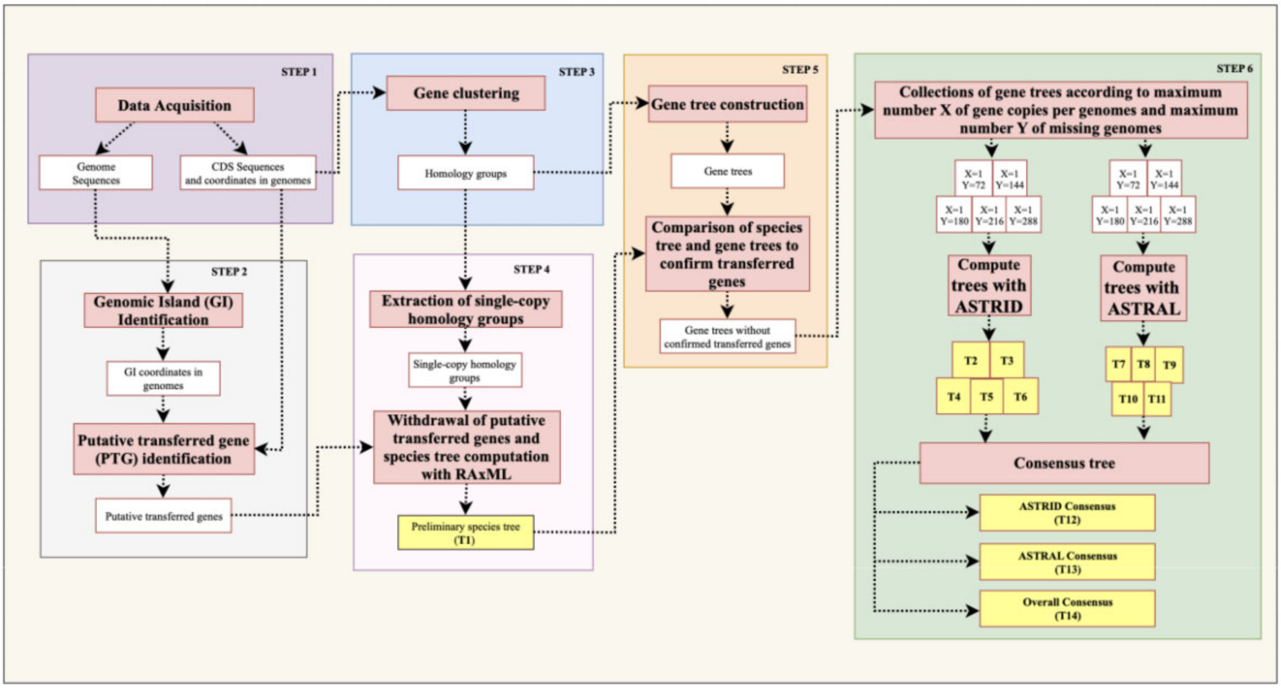
STEP 1
Data Acquisition
Genome Sequences
CDS Sequences and coordinates in genomes
STEP 2
Genomic Island (GI) Identification
GI coordinates in genomes
Putative transferred gene (PTG) identification
Putative transferred genes
STEP 3
Gene clustering
Homology groups
STEP 4
Extraction of single-copy homology groups
Single-copy homology groups
Withdrawal of putative transferred genes and species tree computation with RAxML
Preliminary species tree (T1)
STEP 5
Gene tree construction
Gene trees
Comparison of species tree and gene trees to confirm transferred genes
Gene trees without confirmed transferred genes
STEP 6
Collections of gene trees according to maximum number X of gene copies per genomes and maximum number Y of missing genomes
X=1 Y=72
X=1 Y=144
X=1 Y=180
X=1 Y=216
X=1 Y=288
Compute trees with ASTRID
Compute trees with ASTRAL
T2
T3
T4
T5
T6
T7
T8
T9
T10
T11
Consensus tree
ASTRID Consensus (T12)
ASTRAL Consensus (T13)
Overall Consensus (T14)

Fig. 1.—Overview of the method, which consists of six steps.

trees using the majority-rule consensus (Retief 2000) in two phases. First, the collections of trees obtained with ASTRAL and ASTRID were reduced to two trees. Then, those two trees and the preliminary species tree from Step 4 were reduced to a single tree.

Materials and Methods

Figure 1 depicts the entire method used to estimate the phylogeny of *Corynebacteriales*. The details and rationale underlying each step in the method are described below.

Step 1: Data Acquisition

Step 1 consists in acquiring the *Corynebacteriales* genome and gene data. All complete *Corynebacteriales* genome sequences were retrieved from the REFSEQ NCBI database, release 81 (Maglott et al. 2005). The CDSs and gene coordinates were subsequently extracted using the genome annotations.

Step 2: Detecting GIs and Identifying PTGs

Identifying horizontally transferred genes in bacterial genomes is a prerequisite to computing a bacterial phylogenetic tree using gene tree-based phylogeny methods. Horizontally transferred genes can be located in GIs, which are large segments of DNA (10–200 kb) acquired by horizontal transfer (Langille et al. 2010). There are several approaches for detecting the GI regions in genomes. Some methods—such as MobilomeFINDER (Ou et al. 2007)—make use of a comparative genomics approach and identify GIs as deleted or inserted regions inferred by aligning closely related genomes. Other comparative genomics methods—such as NOTUNG (Chen et al. 2000)—identify HGT by detecting discordance between a gene tree and a species tree. Other methods, referred to as parametric, make use of a sequence composition approach that defines GIs as regions with dinucleotide (G + C) bias or codon usage bias containing associated mobility genes. Compared with parametric methods, comparative approaches have the advantage of being able to detect old HGT events despite the process of sequence homogenization undergone by old GI regions. They however require the availability of closely related genomes or a reliable species tree for the input genomes. Because neither closely related genomes for all genomes of the *Corynebacteriales* data set nor any reliable input species tree was available, GIs were detected with the parametric method IslandPath-DIMOB v1.0.0, which is currently the most accurate stand-alone method for GI prediction (Bertelli and Brinkman 2018) (recall rate of 46.9% and high precision rate of 87.4%). The default parameters of IslandPath-DIMOB were used. The genes contained in the detected GI regions were classified as PTGs. Note that, because of the recall rate of the method, there may be horizontally transferred genes located in GIs that were not detected by the method. They may also be transferred genes not detected by the method because they are not located in GIs. These undetected horizontally transferred genes are considered in Step 6 during the species trees construction.

Step 3: Clustering of Genes into Homology Groups

The CDSs extracted in Step 1 were translated into protein sequences and clustered using Orthofinder1 (Emms and Kelly 2015). This protein clustering tool was chosen because of its high accuracy compared with other currently available gene clustering methods (Emms and Kelly 2015). Orthofinder solves gene length bias before constructing gene groups. An all-against-all BlastP with a stringent cutoff *e*-value of 10^{-4} was applied between and within proteomes, and the result was used as input in Orthofinder to compute gene clusters. The resulting clusters of genes are called homology groups. The default parameters of Orthofinder were used. Note that a new version of Orthofinder, Orthofinder2 (Emms and Kelly 2019), was released after the completion of the present study. The results presented in this study were obtained using Orthofinder1. However, the pipeline provided to reproduce the analysis on other data sets has been updated to include Orthofinder2.

Step 4: Preliminary Species Tree Construction Using Single-Copy Homology Groups

Single-copy homology groups were selected from the homology groups computed in Step 3. Single-copy homology groups are gene clusters containing exactly one gene from each genome. Such homology groups are considered as putative orthology groups that have evolved from a common ancestral gene without any gene duplication events. Thus, they can be used to infer a preliminary species tree using an alignment-based phylogeny estimation method. Due to HGT events, however, they may contain PTGs that should be removed before using the groups for estimating the species tree. Still, PTGs were removed from single-copy homology groups, and the remaining sequences in each group were aligned using the multiple sequence alignment software MAFFT (Katoh et al. 2002). The resulting alignments were concatenated, and the concatenated multiple alignment was used as input to the phylogeny construction method RAxML (Stamatakis 2014) to compute an initial phylogenetic tree with the set of *Corynebacteriales* genomes. The default parameters of MAFFT were used. RAxML was used with the following parameters: raxml -s alignmentfile -p 123456 -m PROTGAMMAAUTO -b 123456 -N 100 –o Nostoc_punctiforme –asc-corr lewis. The tree was rooted using the genome of *Nostoc punctiforme*—(GenBank ID: NC 010628), a symbiotic nitrogen-fixing cyanobacteria—as outgroup.

Step 5: Gene Tree Construction and Discarding of Confirmed Transferred Genes

Among the homology groups computed at Step 3, those that contained at most one gene per genome were extracted. The set of homology groups was restricted to this set because the gene tree-based methods for species tree estimation require that gene trees contain at most one gene per genome (Mirarab and Warnow 2015; Vachaspati and Warnow 2015). For each of the 9,161 homology groups selected, a gene tree was built using the sequence alignment tool MAFFT (Katoh et al. 2002), and the phylogeny inference tool FastTree (Price et al. 2010). FastTree is a ML method that only implements partially the ML approach. It was shown to be more accurate and faster than other ML approaches for applications on large data sets. FastTree was chosen for gene tree construction because of its effectiveness in computing trees on large data sets. The default parameters of FastTree were used. The gene trees were then rooted with homologous CDSs from *N. punctiforme*, as in Step 4. A total of 631 homology groups without any homolog in *N. punctiforme* were discarded, leaving 8,530 gene trees for the analysis. Each gene tree was compared with the preliminary species tree built in Step 4 in order to double-check the classification of PTGs detected in Step 2 and correct false positives. The comparison method is as follows (see supplementary fig. S7, Supplementary Material online, for an illustration). Given any maximum complete subtree T1 of a gene tree G such that the leaves of T1 were all PTGs, we considered T2, the sibling subtree of T1 in G. The sets of species corresponding to the genes at leaves of T1 and T2 are denoted SA1 and SA2, respectively. The PTGs in T1 were reclassified as vertically inherited genes if the lowest common ancestor (lca) node of SA1 and the lca node of SA2 in the species tree S were the same node or sibling nodes. The rationale is that if T1 is the result of a HGT event from a donor branch (a, b) to an acceptor branch (a′, b′) of the species tree such that b and b′ are not sibling nodes, then the lca node of SA1 should be the node b′ and the lca node of SA2 should be the node b. Thus, in the case where lca(SA1) and lca(SA2) are the same node or sibling nodes, the hypothesis that T1 is the result of a HGT event can be discarded. The PTGs that were not reclassified were confirmed as transferred genes and removed from the homology groups and the corresponding gene trees.

Step 6: Species Tree Construction Using Gene Tree-Based Methods

The gene trees obtained at the end of Step 5 were categorized into five collections of trees according to the maximum proportion of missing genomes in the gene trees: 20%, 40%, 50%, 60%, or 80% of missing genomes. Using each collection of trees, two species trees were constructed using the gene tree-based summary methods ASTRID (Vachaspati and Warnow 2015) and ASTRAL (Mirarab and Warnow 2015).

ASTRID and ASTRAL were used in order to account for remaining horizontally transferred genes that were not detected in Step 2, either because they were missed by the GI detection method, or because they are not located in GIs. The default parameters of ASTRID and ASTRAL were used. Subsequently, each set of five trees estimated using the same gene tree-based method (ASTRID or ASTRAL) was reduced to a single consensus tree following the majority-rule consensus algorithm in CONSENSE (Felsenstein 1993). Lastly, the ASTRID consensus tree, the ASTRAL consensus tree, and the RAxML preliminary species tree from Step 4 were reduced to single overall consensus tree.

Results

A New Gene Tree-Based Method Applied to Estimate *Corynebacteriales* Phylogeny

We present a gene tree-based method that includes the detection and correction of putative horizontally transferred genes to estimate bacterial phylogenies using complete genome sequences (for an overview of the method, see fig. 1; for a detailed description of the six steps, see the Materials and Methods section).

The phylogeny of *Corynebacteriales* was estimated using 360 records from NCBI Reference Sequence Database, release 81 (Step 1). The 360 genomes cover 101 species and 11 genera of *Corynebacteriales*, as presented in table 1 and additional files A1 and A2.

Table 1
Input Data Set for *Corynebacteriales* Phylogenetic Tree Estimation

Genus	Number of Species	Number of Genomes
Lawsonella	1	2
Hoyosella	1	1
Rhodococcus	7	22
Mycobacterium	32	169
Dietzia	1	1
Tsukamurella	1	1
Corynebacterium	46	150
Brevibacterium	1	2
Nocardia	6	6
Gordonia	4	5
Total	101	360

Using parametric methods for HGT detection, 168,724 PTGs located into 2,874 GIs were detected (Step 2). Additional file A3 presents the number of GIs and PTGs detected per genome.

The gene clustering step resulted in the clustering of 1,356,782 genes (99.2% of genes) into 17,821 nonsingleton homology groups (Step 3). Additional file A4 presents the details on the composition of the homology groups.

The homology groups containing exactly one gene from each of the 360 genomes were considered as putative orthology groups. After the PTGs were removed from these groups, they were used to build a preliminary species tree using the RAxML ML phylogenetic method (Stamatakis 2014) (Step 4). Supplementary table S1, Supplementary Material online, presents the 13 putative orthology groups used in this step. Supplementary figure S1, Supplementary Material online, shows the preliminary species tree.

The preliminary species tree was then used to check the PTGs in the homology groups using a phylogenetic approach that consists in comparing the gene trees of homology groups with the species tree. Using this approach, 13,966 PTGs (8.29% of PTGs) were reclassified as vertically inherited genes (Step 5).

The gene trees corresponding to homology groups with, at most, one gene per genome were clustered into five collections of trees according to the maximum proportion of genomes without any gene in the homology group: 20%, 40%, 50%, 60%, and 80%. For instance, gene trees in which 45% of genomes did not have a gene were included in the 50%, 60%, or 80% groups. Supplementary table S2, Supplementary Material online, provides the number of trees in the five resulting collections.

Two gene tree-based phylogenetic methods—ASTRID (Vachaspati and Warnow 2015) and ASTRAL II (Mirarab and Warnow 2015)—were applied to the 5 collections to generate 10 phylogenies on the 360 input genomes. ASTRID and ASTRAL are methods motivated by, and statistically consistent with, the multispecies coalescent model such that there is free recombination between, but not within, loci. The use of ASTRID and ASTRAL is motivated by the presence of horizontally transferred genes in the data, not detected in Step 2. The trees obtained using ASTRID and ASTRAL were rooted with the outgroup method by including a homologous CDS from the species *N. punctiforme*. In order to evaluate the similarity between the estimated phylogenies, the percentage of conserved clades between each pair of trees was computed (see table 2). The average pairwise similarity between ASTRID trees is 75.89% with values ranging from 67.22% to 87.78%. ASTRAL trees display a higher average pairwise similarity of 80.14% with values ranging from 74.17% to 89.17%. The average pairwise similarity between ASTRID trees and ASTRAL trees is 67.89% with values ranging from 65.56% to 73.89%.

Considering the high similarity between the five phylogenies estimated using each of the two methods, the trees estimated with ASTRID, on one side, and with ASTRAL, on the other, were reduced to two consensus trees with the CONSENSE majority-rule consensus tool (Felsenstein 1993). Supplementary table S3, Supplementary Material online, presents the similarity between the two resulting consensus trees and the ten initially estimated phylogenies. The two consensus trees for ASTRID and ASTRAL have a high percentage of conserved clades (78.27%), and 82.14% and 80.61%, respectively, of conserved clades with the preliminary species tree from Step 4 obtained with RAxML. Therefore, a final

Table 2

Square Matrix of Percentage of Conserved Clades between Phylogenies Estimated Using RAxML (Step 4), and ASTRID and ASTRAL (Step 5)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
RAxML (1)	100	82.65	82.14	83.16	82.65	82.14	81.12	81.12	80.10	81.12	81.63
ASTRID20 (2)		100	87.78	74.17	74.44	67.22	68.61	70.28	66.67	66.67	66.11
ASTRID40 (3)		—	100	76.11	75.56	68.61	67.5	69.17	67.22	67.78	66.67
ASTRID50 (4)		—	—	100	83.61	72.5	65.56	67.22	68.61	69.17	68.06
ASTRID60 (5)		—	—	—	100	78.89	66.11	67.22	68.06	70.56	70.28
ASTRID80 (6)		—	—	—	—	100	65.83	65.83	66.39	68.89	73.89
ASTRAL20 (7)		—	—	—	—	—	100	87.22	77.78	75.0	74.17
ASTRAL40 (8)		—	—	—	—	—	—	100	78.89	77.78	75.83
ASTRAL50 (9)		—	—	—	—	—	—	—	100	89.17	80.0
ASTRAL60 (10)		—	—	—	—	—	—	—	—	100	85.56
ASTRAL80 (11)		—	—	—	—	—	—	—	—	—	100

reduction of the three phylogenies—ASTRID consensus, ASTRAL consensus and RAxML—to a single consensus tree (referred to as overall consensus) was made. Supplementary figure S2, Supplementary Material online, depicts the four trees viewed at the genus level. All 14 trees generated in this research are available on the iTOL webserver (Letunic and Bork 2019), at https://itol.embl.de/shared/cobius_udes (last accessed April 2, 2020).

ASTRAL was used to compute the quartet supports of branches in the four trees. The quartet support of a branch is computed using the percentage of quartets in input gene trees that agree or disagree with the branch (Sayyari and Mirarab 2016). For each of the ASTRID, ASTRAL, and RAxML trees, the internal branches (nontrivial clades) were divided into two groups: those conserved in the overall consensus tree, and those that were not conserved. Figure 2 (bottom-left) presents the number of branches in each of the six groups, and figure 2 (top-left) presents boxplots of the quartet supports of branches in the six groups. For RAxML, the boxplots of the bootstrap supports of branches are also depicted. We observe that the clades from ASTRID, ASTRAL, and RAxML included in the overall consensus show high quartet support values, whereas the clades not included in the overall consensus show low quartet support values. The same observation holds for the RAxML bootstrap values. This means that the overall consensus tree is effective at retaining the clades of the three input trees which present the highest support values.

Table 3 presents the similarity measures between the four trees. The overall consensus has, respectively, 98.06%, 96.90%, and 84.69% of conserved clades with the ASTRID, ASTRAL, and RAxML trees. The internal branches of the overall consensus tree were divided into four groups depending on their presence in all three input trees (ASTRID, ASTRAL, and RAxML) or in only two of the input trees. Figure 2 (bottom-right) presents the number of internal branches in each of the four groups, and figure 2 (top-right) presents boxplots of the quartet supports of branches in the four groups. The branches present in all three input trees (ASTRID–ASTRAL–RAxML) constitute the largest group with the highest support values, followed by the ASTRID–ASTRAL branches, a few ASTRID–RAxML branches, and finally a few ASTRAL–RAxML branches. This means that the largest contribution comes from the consensus between the three trees or between ASTRID and ASTRAL trees, and further branches are added thanks to the consensus with the RAxML tree.

Analysis at the Genus Level

All the trees estimated in our study place the genus *Brevibacterium* (BV) within *Corynebacterium* (CR), which supports reclassifying *Brevibacterium* as *Corynebacterium*, as proposed in a recent study (Yang and Yang 2017). The *Corynebacteriales* phylogeny is still under debate. Past studies have reported various topologies for the phylogeny of this order. Gao and Gupta (2012) used the sequence alignments of 35 proteins with neighbor-joining methods to estimate the phylogeny of *Actinobacteria* that includes *Corynebacteriales*. Sen et al. (2014) used 54 protein sequences aligned with RAxML to infer a phylogeny on 100 actinobacterial strains. They also reported a second phylogeny on the 100 actinobacterial strains obtained by applying RAxML to the alignment of 5 conserved genes identified with a multilocus sequence analysis. We compared the overall consensus tree obtained in our study with the *Corynebacteriales* phylogenies from Gao and Gupta (2012) and Sen et al. (2014). Figure 3 depicts the compared phylogenies at the genus level. A strong consensus can be seen between the trees for the clade grouping *Nocardia* and *Rhodococcus* (6/6) and the clade regrouping *Gordonia* and *Tsukamurella* (5/6). We also observed a majority-rule consensus for a clade grouping *Nocardia*, *Rhodococcus*, and *Hoyosella* (3/5), and for placing *Corynebacterium* as the outgroup (3/6).

Systematic Analysis of the Phylogeny Reported inside Genera

We analyzed the phylogenies for each of the 5 genera with more than 1 species in the data set: *Corynebacterium* (46 species), *Mycobacterium* (32 species), *Rhodoccoccus* (7

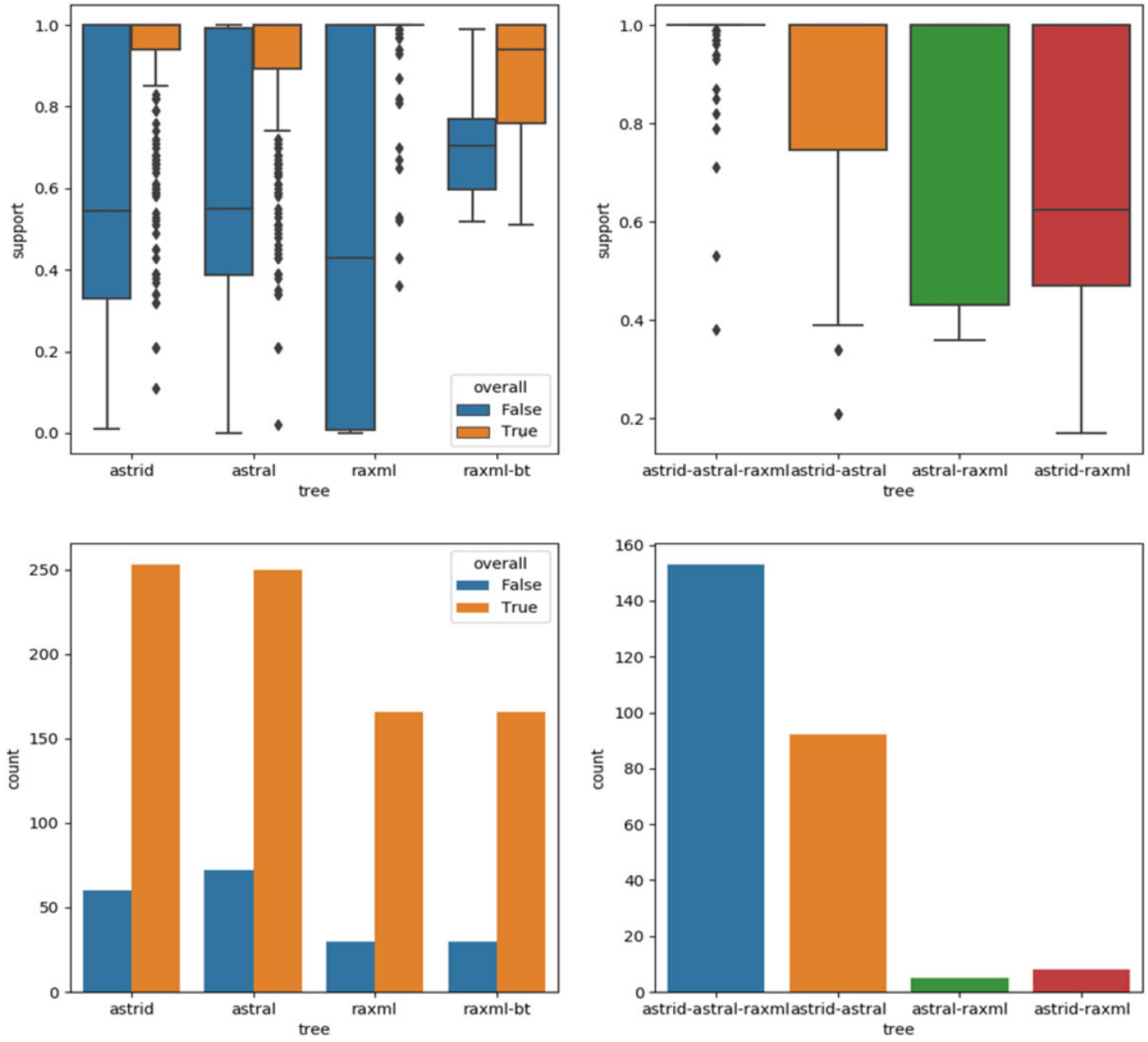

Fig. 2.—Support values of internal branches of the ASTRID, ASTRAL, RAxML, and Overall consensus trees. Bottom-left: Numbers of branches in the ASTRID, ASTRAL, and RAxML trees conserved in the overall consensus tree and numbers of branches not conserved. Top-left: Boxplots of the quartet support values of branches in each group and the bootstrap support values for the RAxML tree. Bottom-right: Numbers of branches in the overall consensus tree present in the ASTRID, ASTRAL, and RAxML trees, or in only two of the three trees. Top-right: Boxplots of the quartet support values of branches in each group.

Table 3
Square Matrix of the Percentage of Conserved Clades between the Consensus Trees and the Preliminary RAxML Tree

	(1)	(2)	(3)	(4)
ASTRID consensus (1)	100	78.27	82.14	98.06
ASTRAL consensus (2)		100	80.61	96.90
RAxML (3)		—	100	84.69
Overall consensus (4)		—	—	100

species), *Nocardia* (6 species), and *Gordonia* (4 species). In all the trees estimated in our study, the species of the same genus were grouped into monophyletic groups. We extracted the complete subtree corresponding to each genus in the overall consensus tree.

Corynebacterium

The genus *Corynebacterium* comprises a variety of bacterial species that includes potential pathogens for human and animals, as well as pathogens for normal microbiota (Von Graevenitz and Bernard 2006). Most of the mechanisms underlying diseases caused by these species are still unclear; a few phylogenies of the genus have been reconstructed (Pascual et al. 1995; Baek et al. 2018; Dangel et al. 2019).

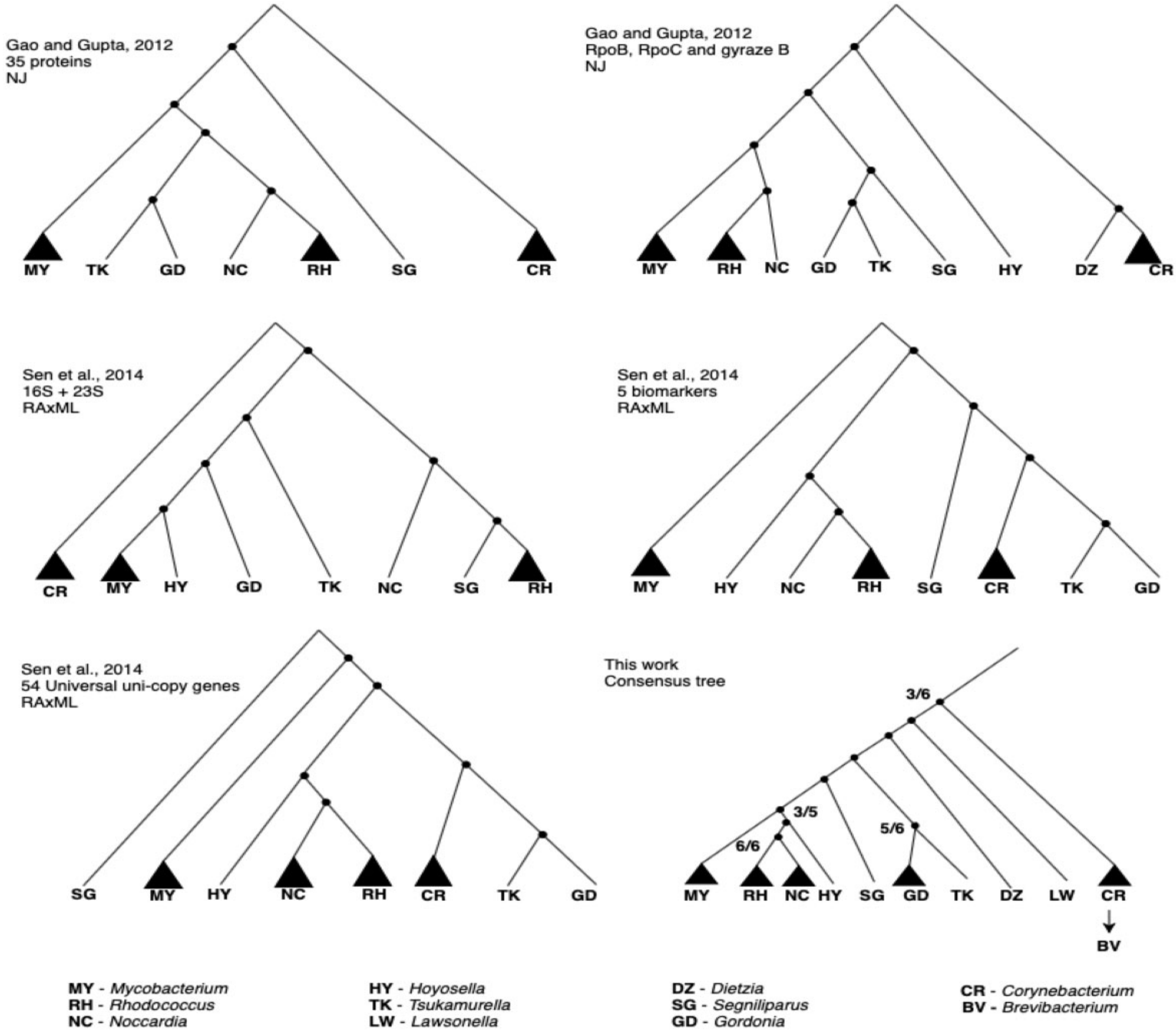

FIG. 3.—Illustration at the genus level of the overall consensus phylogeny reconstructed in our work, two phylogenies reconstructed in Gao and Gupta (2012), and three phylogenies reconstructed in Sen et al. (2014). For each tree, the genera for which the data set contains more than one species are represented as triangles. Note that the sets of genera differ between trees. The ratio of trees displaying the clade is indicated for each conserved clade.

In all the trees estimated in our study, we observed a division into two categories: nonpathogenic genomes and pathogenic genomes forming a monophyletic group (fig. 4). We noted a single exception: the classification of *Corynebacterium jeikeium* among nonpathogenic genomes. *Corynebacterium jeikeium* is a pathogen isolated from immunosuppressive patients highly exposed to antibiotic treatments (Tauch et al. 2005). The positioning of this pathogen species among nonpathogens is surprising and might be related to HGT for the acquisition of antibiotic-resistance genes.

In fact, as it was only isolated from immunosuppressive patients, this is a robust biological clue suggesting that this species does not act as a pathogen in healthy organisms (in this case, humans). Therefore, it is not a primary pathogen, which would corroborate our findings. More detailed studies are needed to refine this assumption. In the nonpathogenic group, *Corynebacterium glutamicum* is the most thoroughly studied species due to its biotechnological applications in producing amino acids such as L-arginine, L-histidine, L-carnitine, L-lysine, and L-valine (Keilhauer et al. 1993). In all the trees estimated in our study, the genomes of *Brevibacterium flavum* strain ATCC 15168 (RefSeq. CP011309) and *B. flavum* ZL 1 (RefSeq. CP004046) always appear in the same clade of the genomes of *C. glutamicum* (supplementary fig. S3, Supplementary Material online). This classification was recently proposed in the literature (Yang and Yang 2017).

In the pathogenic group, we observe a clustering of *Corynebacterium diphteriae* and *Corynebacterium ulcerans.*

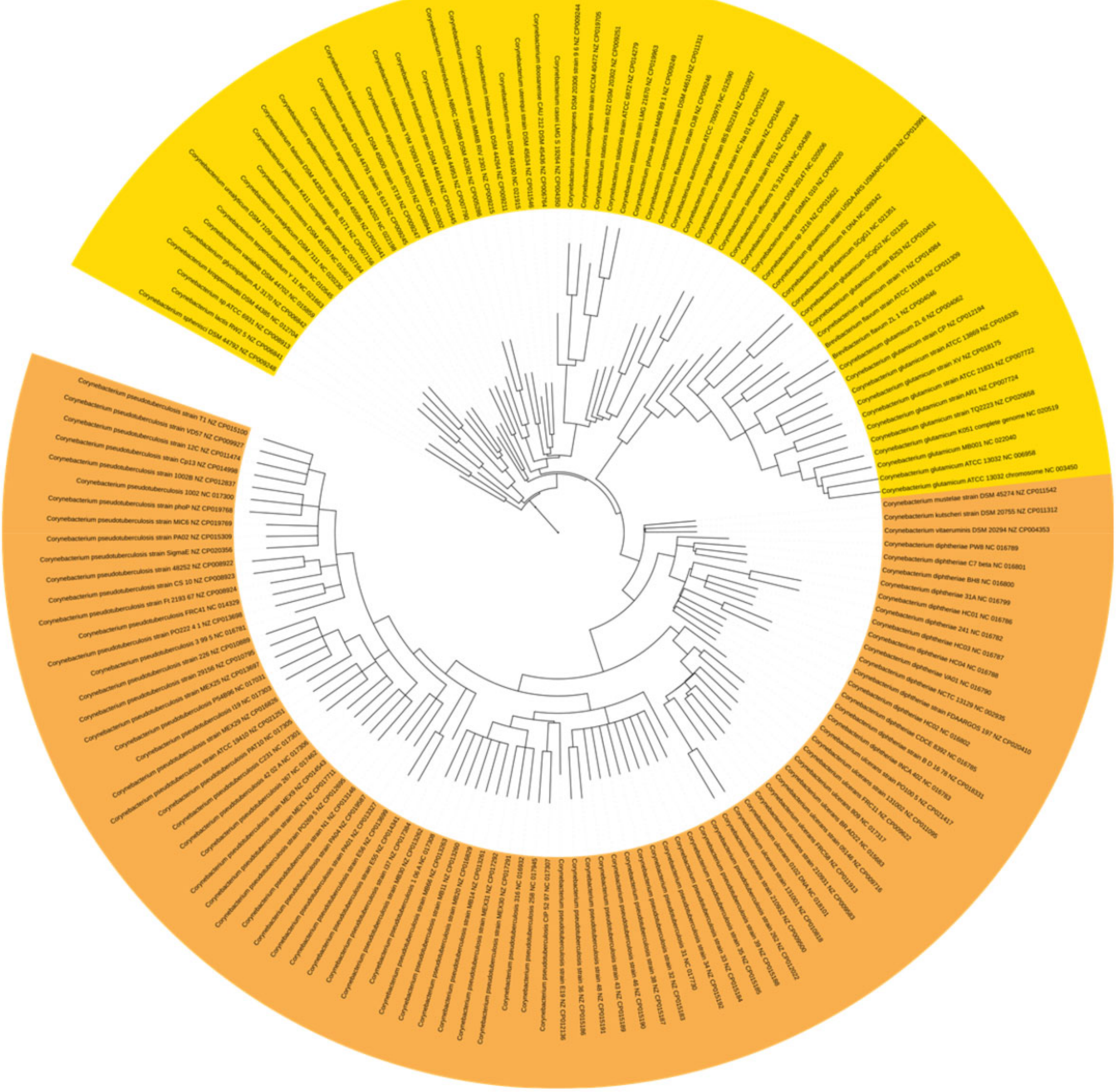

Fig. 4.—Partition of *Corynebacterium* into two categories: nonpathogenic in yellow and pathogenic in orange. Detailed method used to estimate this phylogeny (overall consensus tree: RAxML + ASTRID consensus + ASTRAL consensus).

Corynebacterium diphteriae is the etiological agent of diphtheria in humans, an infectious disease caused by the exotoxin produced by this pathogen (Cerdeno-Tarraga et al. 2003). *Corynebacterium ulcerans* is primarily reported for causing mastitis in cattle and humans due to the consumption of raw milk or unpasteurized dairy products in rural populations (Hommez et al. 1999). In the literature, *C. ulcerans* has been closely related to *C. diphtheriae*, which produces a toxin causing symptoms similar to those caused by *C. ulcerans* (Riegel et al. 1995).

We also detected the recently proposed anagenesis of *Corynebacterium pseudotuberculosis* (Oliveira et al. 2016). In this model, two biovars are described: equi and ovis. They mainly differ by the presence of the nitrate reductase enzyme present in biovar equi, which results in 1% of the nucleotide differences between biovars (Soares et al. 2013) (fig. 5). *Corynebacterium pseudotuberculosis* is the etiological agent of caseous lymphadenitis (CLA), a highly prevalent chronic disease affecting sheep and goats. It is difficult to control and causes significant economic losses to farmers

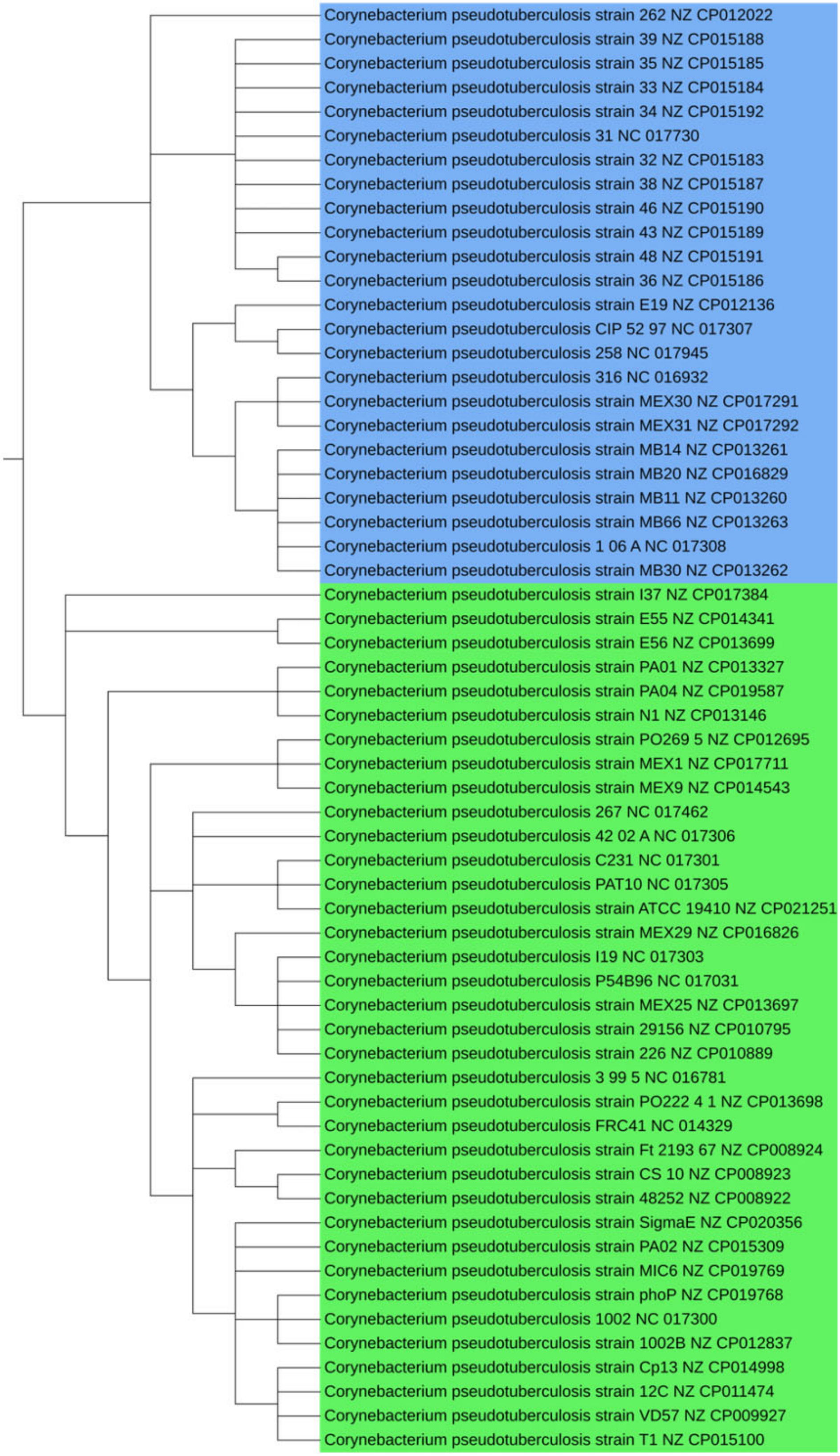
Corynebacterium pseudotuberculosis strain 262 NZ CP012022
Corynebacterium pseudotuberculosis strain 39 NZ CP015188
Corynebacterium pseudotuberculosis strain 35 NZ CP015185
Corynebacterium pseudotuberculosis strain 33 NZ CP015184
Corynebacterium pseudotuberculosis strain 34 NZ CP015192
Corynebacterium pseudotuberculosis 31 NC 017730
Corynebacterium pseudotuberculosis strain 32 NZ CP015183
Corynebacterium pseudotuberculosis strain 38 NZ CP015187
Corynebacterium pseudotuberculosis strain 46 NZ CP015190
Corynebacterium pseudotuberculosis strain 43 NZ CP015189
Corynebacterium pseudotuberculosis strain 48 NZ CP015191
Corynebacterium pseudotuberculosis strain 36 NZ CP015186
Corynebacterium pseudotuberculosis strain E19 NZ CP012136
Corynebacterium pseudotuberculosis CIP 52 97 NC 017307
Corynebacterium pseudotuberculosis 258 NC 017945
Corynebacterium pseudotuberculosis 316 NC 016932
Corynebacterium pseudotuberculosis strain MEX30 NZ CP017291
Corynebacterium pseudotuberculosis strain MEX31 NZ CP017292
Corynebacterium pseudotuberculosis strain MB14 NZ CP013261
Corynebacterium pseudotuberculosis strain MB20 NZ CP016829
Corynebacterium pseudotuberculosis strain MB11 NZ CP013260
Corynebacterium pseudotuberculosis strain MB66 NZ CP013263
Corynebacterium pseudotuberculosis 1 06 A NC 017308
Corynebacterium pseudotuberculosis strain MB30 NZ CP013262
Corynebacterium pseudotuberculosis strain I37 NZ CP017384
Corynebacterium pseudotuberculosis strain E55 NZ CP014341
Corynebacterium pseudotuberculosis strain E56 NZ CP013699
Corynebacterium pseudotuberculosis strain PA01 NZ CP013327
Corynebacterium pseudotuberculosis strain PA04 NZ CP019587
Corynebacterium pseudotuberculosis strain N1 NZ CP013146
Corynebacterium pseudotuberculosis strain PO269 5 NZ CP012695
Corynebacterium pseudotuberculosis strain MEX1 NZ CP017711
Corynebacterium pseudotuberculosis strain MEX9 NZ CP014543
Corynebacterium pseudotuberculosis 267 NC 017462
Corynebacterium pseudotuberculosis 42 02 A NC 017306
Corynebacterium pseudotuberculosis C231 NC 017301
Corynebacterium pseudotuberculosis PAT10 NC 017305
Corynebacterium pseudotuberculosis strain ATCC 19410 NZ CP021251
Corynebacterium pseudotuberculosis strain MEX29 NZ CP016826
Corynebacterium pseudotuberculosis I19 NC 017303
Corynebacterium pseudotuberculosis P54B96 NC 017031
Corynebacterium pseudotuberculosis strain MEX25 NZ CP013697
Corynebacterium pseudotuberculosis strain 29156 NZ CP010795
Corynebacterium pseudotuberculosis strain 226 NZ CP010889
Corynebacterium pseudotuberculosis 3 99 5 NC 016781
Corynebacterium pseudotuberculosis strain PO222 4 1 NZ CP013698
Corynebacterium pseudotuberculosis FRC41 NC 014329
Corynebacterium pseudotuberculosis strain Ft 2193 67 NZ CP008924
Corynebacterium pseudotuberculosis strain CS 10 NZ CP008923
Corynebacterium pseudotuberculosis strain 48252 NZ CP008922
Corynebacterium pseudotuberculosis strain SigmaE NZ CP020356
Corynebacterium pseudotuberculosis strain PA02 NZ CP015309
Corynebacterium pseudotuberculosis strain MIC6 NZ CP019769
Corynebacterium pseudotuberculosis strain phoP NZ CP019768
Corynebacterium pseudotuberculosis 1002 NC 017300
Corynebacterium pseudotuberculosis strain 1002B NZ CP012837
Corynebacterium pseudotuberculosis strain Cp13 NZ CP014998
Corynebacterium pseudotuberculosis strain 12C NZ CP011474
Corynebacterium pseudotuberculosis strain VD57 NZ CP009927
Corynebacterium pseudotuberculosis strain T1 NZ CP015100

FIG. 5.—Partition of *C. pseudotuberculosis* into two biovars. Biovars equi and ovis are shown in green and blue, respectively. Detailed method used to estimate this phylogeny (overall consensus tree: RAxML + ASTRID consensus+ ASTRAL consensus).

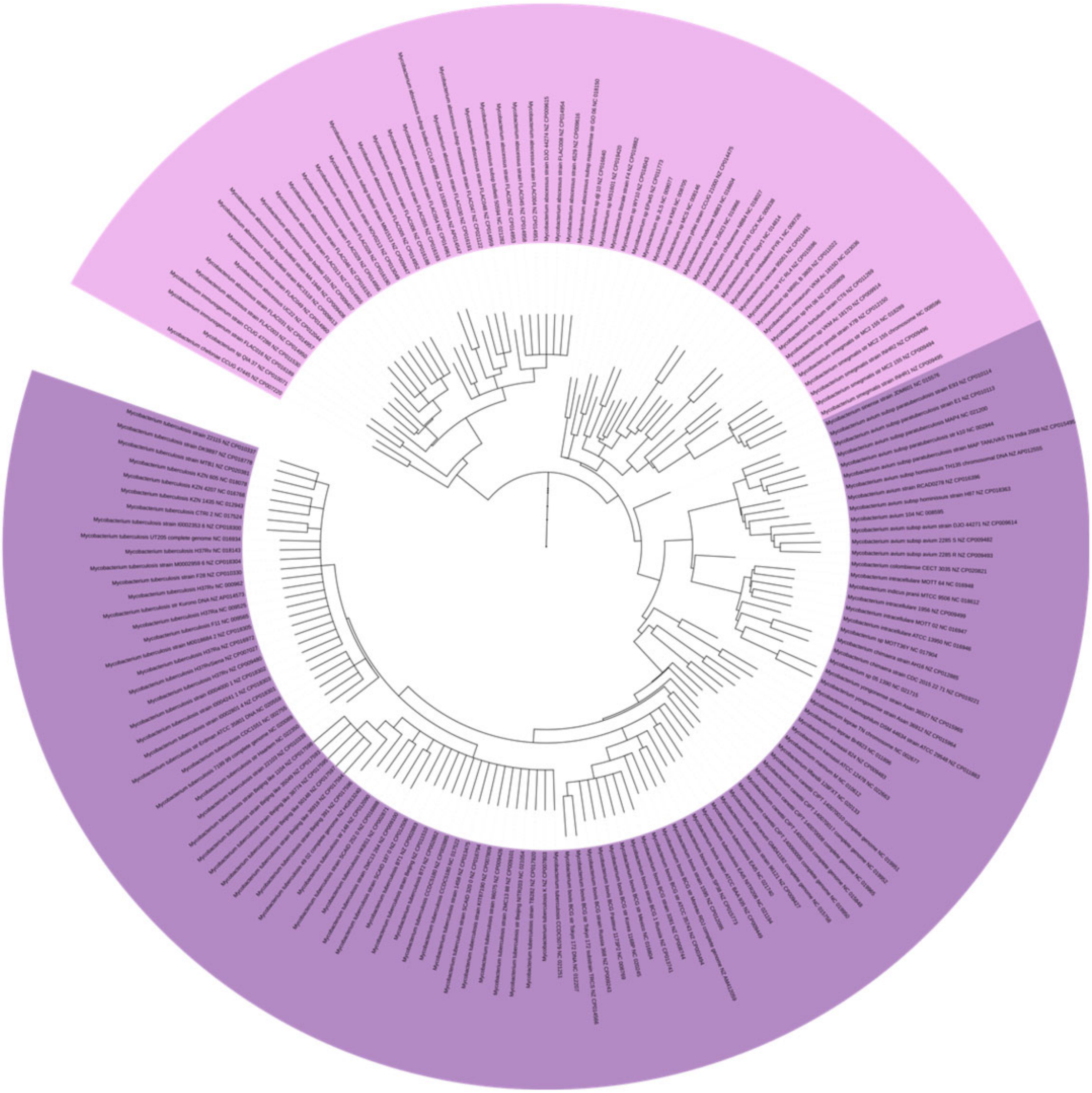

FIG. 6.—Partition of *Mycobacterium* into two categories: slow growers in purple and fast growers in pale pink. Detailed method used to estimate this phylogeny (overall consensus tree: RAxML + ASTRID consensus+ ASTRAL consensus).

(Baird and Fontaine 2007). Human infections caused by *C. pseudotuberculosis* are rare, but it has been reported as the agent of necrotizing lymphadenitis in human (Mills et al. 1997). Lastly, we observed that our phylogeny for *Corynebacterium* species is consistent with the phylogeny proposed by Gao and Gupta (2012).

Mycobacterium

The genus *Mycobacterium* comprises one the most dangerous human pathogens—*Mycobacterium tuberculosis*—which causes tuberculosis (Gagneux 2018; Koch and Mizrahi 2018). This genus also comprises others important animal pathogens such as *Mycobacterium leprae*, *Mycobacterium bovis*, and *Mycobacterium avium* (Frothingham and Wilson 1993). The taxonomy of *Mycobacterium* solely relies in two categories: slow growers and fast growers. This poorly detailed taxonomy is due to the lack of descriptive features for taxonomic classification. A more detailed classification would help with global monitoring of disease outbreaks caused by species of this genus (Rogall et al. 1990; Stahl and Urbance 1990). All the estimated trees display a division into the 2 categories: 61

genomes of slow growers forming a monophyletic group, and 108 genomes of fast growers forming another monophyletic group (fig. 6).

Rhodococcus

The *Rhodococcus* species are used as versatile genetic tools in the biotechnological industry because of their capacity for remediation, biotransformation and biocatalysis, biodegradation of diverse metabolic compounds, adaptation and tolerance to solvents, and interactions with metals (Sangal et al. 2019). *Rhodococcus* species are distributed in soil, water, and marine sediments (Larkin et al. 2005). Some of them are also pathogens for humans, animals, and plants (Prescott 1991). Although new *Rhodococcus* genomes are still being sequenced because of their important biotechnological applications, the current phylogenies of *Rhodococcus* are only estimated for closely related species using few biomarkers (Anastasi et al. 2016; Duquesne et al. 2017). We report a phylogeny of 22 *Rhodococcus* genomes corresponding to 6 species and 11 unclassified genomes, divided into 6 clusters (supplementary fig. S4, Supplementary Material online). The lack of estimated phylogenies for *Rhodococcus* at the species level in the literature makes it hard to conduct a proper comparison with past studies. We however observed partial agreement of the estimated phylogeny for the six *Rhodococcus* species with the phylogeny from (Anastasi et al. 2016): (*R. fascians*,(*R. pyridinovorans*, (*R. erythropolis*,(*R. opacus*, *R. jostii*)))).

Nocardia

Nocardia species are a complex group of organisms that cause serious human infections, especially in immunocompromised patients. Like *Rhodococcus* and the other genera of *Cornynebacteriales*, the taxonomy and phylogeny of *Nocardia* species are subject to open debate (Conville et al. 2018). The data set contains six complete genomes of *Nocardia* corresponding to six species. All the estimated trees display the same phylogeny for the six species (supplementary fig. S5, Supplementary Material online). In an estimation of the phylogeny of *Nocardia* species, Conville et al. (2018) described the complex history behind the taxonomy of the genus by reconstructing a phylogenetic tree using the 16S rRNA gene of 59 genomes of *Nocardia*. The intersection between their data set and our data set consists of only four species. The estimated phylogeny for these four species in (Conville et al. 2018) is ((*N. farcinica*, *N. brasilensis*), (*N. cyriacigeorgica*, *N. nova*)), which differs from the phylogeny estimated in this report.

Gordonia

Gordonia species have attracted interest from the biotechnological industry in recent years because of their ability to degrade environmental pollutants as well as natural polymers and compounds, making them potentially useful for environmental and industrial biotechnology (Arenskötter et al. 2004). Some species of *Gordonia* are reported to cause infections in humans (Ramanan et al. 2013; Sowani et al. 2017). Previous phylogenies of *Gordonia* were estimated with 16S rRNA genes, and the phylogeny is still under debate (Blaschke et al. 2007). Kang et al. (2009) studied the phylogeny of 23 species using *gyrB*, *secA1*, and 16S rRNA genes. In our study, we collected five *Gordonia* genomes, corresponding to three species and two unclassified genomes. All the estimated trees display the same phylogeny for the five genomes (supplementary fig. S6, Supplementary Material online). The induced phylogeny for the three species included in our data set—*G. polyisoprenivorans*, *G. bronchialis*, and *G. terrae*—agrees with the induced phylogeny from Kang et al. (2009).

Discussion

As new bacterial genomes are still being sequenced, one of the major problems lies with identifying the main bacteria groups and recovering the phylogenetic relationships between these groups (Larson 1998). Current modern molecular-biology techniques are still being redesigned to identify new species because the classical approaches based on sequence analysis are inefficient for discrimination (Glaeser and Kämpfer 2015). Furthermore, characterizing the differences between closely related species remains challenging (Christensen and Olsen 2018). Sequence-based phylogenies have been an active research field since the beginning of the 2000s. They grounded the current knowledge about the diversity of organisms on the Earth. Estimating bacterial phylogenies is not, however, a trivial problem. This is mainly because bacterial genomes are highly affected by the swapping of genetic material between genomes via HGT processes (Soucy et al. 2015). Through this mechanism, bacterial genomes acquire and spread genes that confer adaptive advantages, such as antibiotic-resistance genes leading to the rise of multidrug-resistant bacteria (Van Duin and Paterson 2016). Thus, accounting for horizontally transferred genes is necessary to accurately estimate bacterial phylogenies when using sequence-based phylogenetic methods. Nonetheless, current sequence-based phylogenetic methods do not include a step back to audit the data sets in order to identify and remove transferred genes. An alternative to identifying and removing transferred genes before estimating phylogenies is to infer ancestral recombination graphs that record of all coalescence and recombination events in the evolution of a set of homologous sequences (D O'Fallon 2013; Rasmussen et al. 2014). However, existing methods for ancestral recombination graphs inference are computationally intensive and limited to small numbers of sequences. Herein, the phylogeny of *Corynebacteriales* was estimated while accounting for HGTs, by detecting and removing a part of the HGT located in GIs using a parametric GI detection method, and by relying on phylogenetic reconstruction methods which are consistent the multispecies coalescent model with recombination within loci. The result is a species tree that displays all the genera as monophyletic clades. The estimated trees display several phylogenetic relationships proposed by previous studies: 1) the classification of *B. flavum* inside *C. glutamicum* (Yang and Yang 2017), 2) the monophyletic group composed of pathogens *C. ulcerans* and *C. diphteriae*, 3) the biovar speciation inside *C. pseudotuberculosis*, and 4) the division between slow growers and fast growers in *Mycobacterium*. Finally, it is important to recall that the phylogenomics method devised in this article presents the same limit as most phylogenetics and comparative genomics methods which reduce biological processes such as HGT to patterns, and thus investigate patterns (Nelson 1970). One should always remember that phylogenetics methods are consistent under the hypothesis that there is a one-to-one correspondence between the target biological processes and the patterns investigated.

Data Availability

All information to retrieve the data and the scripts used for the analysis are available on the CoBIUS lab GitHub (https://github.com/UdeS-CoBIUS/EXECT).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Brazilian funding agencies CNPq and FAPEMIG (24780/RTR/PRPG/ICB/DBG/BDS), grants from the Canada Research Chair (CRC Tier 2 Grant 950-230577), and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant RGPIN-2017-05552).

Literature Cited

Anastasi E, et al. 2016. Pangenome and phylogenomic analysis of the pathogenic actinobacterium *Rhodococcus equi*. Genome Biol Evol. 8(10):3140–3148.

Arenskötter M, Bröker D, Steinbüchel A. 2004. Biology of the metabolically diverse genus *Gordonia*. Appl Environ Microbiol. 70(6):3195–3204.

Baek I, et al. 2018. Phylogeny trumps chemotaxonomy: a case study involving *Turicella otitidis*. Front Microbiol. 9:834.

Baird G, Fontaine M. 2007. *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. J Comp Pathol. 137(4):179–210.

Belda E, Moya A, Silva FJ. 2005. Genome rearrangement distances and gene order phylogeny in γ-proteobacteria. Mol Biol Evol. 22(6):1456–1467.

Bertelli C, Brinkman FS. 2018. Improved genomic island predictions with IslandPath-DIMOB. Bioinformatics 34(13):2161–2167.

Bister B, et al. 2004. Abyssomicin C–A polycyclic antibiotic from a marine *Verrucosispora* strain as an inhibitor of the *p*-aminobenzoic acid/tetrahydrofolate biosynthesis pathway. Angew Chem Int Ed. 43(19):2574–2576.

Blaschke AJ, et al. 2007. *Gordonia* species: emerging pathogens in pediatric patients that are identified by 16S ribosomal RNA gene sequencing. Clin Infect Dis. 45(4):483–486.

Bourque G, Pevzner PA, Tesler G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Res. 14(4):507–516.

Castillo-Ramírez S, et al. 2012. Phylogeographic variation in recombination rates within a global clone of methicillin-resistant *Staphylococcus aureus*. Genome Biol. 13(12):R126.

Cerdeno-Tarraga A, et al. 2003. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res. 31(22):6516–6523.

Chan CX, Darling AE, Beiko RG, Ragan MA. 2009. Are protein domains modules of lateral genetic transfer? PLoS One 4(2):e4524.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 7(3–4):429–447.

Christensen H, Olsen JE. 2018. Sequence-based classification and identification of prokaryotes. In: Introduction to bioinformatics in microbiology. New York: Springer. p. 121–134.

Ciccarelli FD, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science 311(5765):1283–1287.

Comas I, et al. 2013. Out-of-Africa migration and Neolithic coexpansion of *Mycobacterium tuberculosis* with modern humans. Nat Genet. 45(10):1176–1182.

Conville PS, Brown-Elliott BA, Smith T, Zelazny AM. 2017. The complexities of *Nocardia* taxonomy and identification. J Clin Microbiol. 56(1):pii: e01419–17.

D O'Fallon B. 2013. ACG: rapid inference of population history from recombining nucleotide sequences. BMC Bioinformatics 14(1):40.

Dangel A, Berger A, Konrad R, Sing A. 2019. NGS-based phylogeny of diphtheria-related pathogenicity factors in different *Corynebacterium* spp. implies species-specific virulence transmission. BMC Microbiol. 19(1):28.

Duquesne F, et al. 2017. Development of a multilocus sequence typing scheme for *Rhodococcus equi*. Vet Microbiol. 210:64–70.

Emms DM, Kelly S. 2015. Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1):157.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):1–14.

Fair RJ, Tor Y. 2014. Antibiotics and bacterial resistance in the 21st century. Perspect Med Chem. 6:PMC.S14459.

Felsenstein J. 1985. Phylogenies and the comparative method. Am Nat. 125(1):1–15.

Felsenstein J. 1993. PHYLIP (phylogeny inference package), version 3.5 c. Seattle (WA): Joseph Felsenstein.

Fischbach MA, Walsh CT. 2009. Antibiotics for emerging pathogens. Science 325(5944):1089–1093.

Frothingham R, Wilson KH. 1993. Sequence-based differentiation of strains in the *Mycobacterium avium* complex. J Bacteriol. 175(10):2818–2825.

Gagneux S. 2018. Ecology and evolution of *Mycobacterium tuberculosis*. Nat Rev Microbiol. 16(4):202–213.

Gao B, Gupta RS. 2012. Phylogenetic framework and molecular signatures for the main clades of the phylum *Actinobacteria*. Microbiol Mol Biol Rev. 76(1):66–112.

Glaeser SP, Kämpfer P. 2015. Multilocus sequence analysis (MLSA) in prokaryotic taxonomy. Syst Appl Microbiol. 38(4):237–245.

Gogarten JP, Doolittle WF, Lawrence JG. 2002. Prokaryotic evolution in light of gene transfer. Mol Biol Evol. 19(12):2226–2238.

Hommez J, et al. 1999. Identification of nonlipophilic corynebacteria isolated from dairy cows with mastitis. J Clin Microbiol. 37(4):954–957.

Jeong H, Arif B, Caetano-Anollés G, Kim KM, Nasir A. 2019. Horizontal gene transfer in human-associated microorganisms inferred by phylogenetic reconstruction and reconciliation. Sci Rep. 9(1):1–18.

Kang Y, Takeda K, Yazawa K, Mikami Y. 2009. Phylogenetic studies of *Gordonia* species based on *gyrB* and *secA1* gene analyses. Mycopathologia 167(2):95–105.

Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30(14):3059–3066.

Keilhauer C, Eggeling L, Sahm H. 1993. Isoleucine synthesis in *Corynebacterium glutamicum*: molecular analysis of the ilvB-ilvN-ilvC operon. J Bacteriol. 175(17):5595–5603.

Kitahara K, Miyazaki K. 2013. Revisiting bacterial phylogeny: natural and experimental evidence for horizontal gene transfer of 16S rRNA. Mob Genet Elements. 3(1):e24210.

Klappenbach JA, Dunbar JM, Schmidt TM. 2000. rRNA operon copy number reflects ecological strategies of bacteria. Appl Environ Microbiol. 66(4):1328–1333.

Koch A, Mizrahi V. 2018. *Mycobacterium tuberculosis*. Trends Microbiol. 26(6):555–556.

Langille MG, Hsiao WW, Brinkman FS. 2010. Detecting genomic islands using bioinformatics approaches. Nat Rev Microbiol. 8(5):373–382.

Larkin MJ, Kulakov LA, Allen CC. 2005. Biodegradation and *Rhodococcus*—masters of catabolic versatility. Curr Opin Biotechnol. 16(3):282–290.

Larson A. 1998. The comparison of morphological and molecular data in phylogenetic systematics. In: Molecular approaches to ecology and evolution. New York: Springer. p. 275–296.

Lasek-Nesselquist E, Gogarten JP, Lapierre P. 2014. The impact of HGT on phylogenomic reconstruction methods. Brief Bioinform. 15(1):79–90.

Lerat E, Daubin V, Ochman H, Moran NA. 2005. Evolutionary origins of genomic repertoires in bacteria. PLoS Biol. 3(5):e130.

Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47(W1):W256–W259.

Maglott DR, Pruitt KD, Tatusova T. 2004. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33(Database issue):D501–D504.

Mills AE, Mitchell RD, Lim EK. 1997. *Corynebacterium pseudotuberculosis* is a cause of human necrotising granulomatous lymphadenitis. Pathology 29(2):231–233.

Mirarab S, Warnow T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31(12):i44–i52.

Miyazaki K, Sato M, Tsukuda M. 2017. PCR primer design for 16S rRNAs for experimental horizontal gene transfer test in *Escherichia coli*. Front Bioeng Biotechnol. 5:14.

Moret BM, Lin Y, Tang J. 2013. Rearrangements in phylogenetic inference: compare, model, or encode? In: Models and algorithms for genome evolution. New York: Springer. p. 147–171.

Moret BM, Wang L-S, Warnow T, Wyman SK. 2001. New approaches for reconstructing phylogenies from gene order data. Bioinformatics 17(Suppl 1):S165–S173.

Nelson GJ. 1970. Outline of a theory of comparative biology. Syst Zool. 19(4):373–384.

Oliveira A, et al. 2016. *Corynebacterium pseudotuberculosis* may be under anagenesis and biovar Equi forms biovar Ovis: a phylogenic inference from sequence and structural analysis. BMC Microbiol. 16(1):100.

Ou H-Y, et al. 2007. MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic Acids Res. 35(Web Server):W97–W104.

Pascual C, Lawson PA, Farrow JA, Gimenez MN, Collins MD. 1995. Phylogenetic analysis of the genus *Corynebacterium* based on 16S rRNA gene sequences. Int J Syst Evol Microbiol. 45(4):724–728.

Prescott JF. 1991. *Rhodococcus equi*: an animal and human pathogen. Clin Microbiol Rev. 4(1):20–34.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5(3):e9490.

Rajendhran J, Gunasekaran P. 2011. Microbial phylogeny and diversity: small subunit ribosomal RNA sequence analysis and beyond. Microbiol Res. 166(2):99–110.

Ramanan P, Deziel PJ, Wengenack NL. 2013. *Gordonia bacteremia*. J Clin Microbiol. 51(10):3443–3447.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. 2014. Genome-wide inference of ancestral recombination graphs. PLoS Genet. 10(5):e1004342.

Ravenhall M, Kunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. PLoS Comput Biol. 11(5):e1004095–16.

Retief JD. 2000. Phylogenetic analysis using PHYLIP. In: Bioinformatics methods and protocols. New York: Springer. p. 243–258.

Riegel P, et al. 1995. Taxonomy of *Corynebacterium diphtheriae* and related taxa, with recognition of *Corynebacterium ulcerans* sp. nov. nom. rev. FEMS Microbiol Lett. 126(3):271–276.

Rogall T, Wolters J, Flohr T, Bottger EC. 1990. Towards a phylogeny and definition of species at the molecular level within the genus *Mycobacterium*. Int J Syst Evol Microbiol. 40(4):323–330.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425(6960):798–804.

Ruiz JC, et al. 2011. Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. PLoS One 6(4):e18551.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4(4):406–425.

Sangal V, Goodfellow M, Jones AL, Seviour RJ, Sutcliffe IC. 2019. Refined systematics of the genus *Rhodococcus* based on whole genome analyses. In: Biology of Rhodococcus. New York: Springer. p. 1–21.

Sankoff D, Blanchette M. 1998. Multiple genome rearrangement and breakpoint phylogeny. J Comput Biol. 5(3):555–570.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7):1654–1668.

Schouls LM, Schot CS, Jacobs JA. 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. J Bacteriol. 185(24):7241–7246.

Sen A, et al. 2014. Phylogeny of the class *Actinobacteria* revisited in the light of complete genomes. The orders '*Frankiales*' and *Micrococcales* should be split into coherent entities: proposal of *Frankiales* ord. nov., *Geodermatophilales* ord. nov., *Acidothermales* ord. nov. and *Nakamurellales* ord. nov. Int J Syst Evol Microbiol. 64(Pt 11):3821–3832.

Soares SC, et al. 2013. The pan-genome of the animal pathogen *Corynebacterium pseudotuberculosis* reveals differences in genome plasticity between the biovar *ovis* and *equi* strains. PLoS One 8(1):e53818.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. Nat Rev Genet. 16(8):472–482.

Sowani H, Kulkarni M, Zinjarde S, Javdekar V. 2017. *Gordonia* and related genera as opportunistic human pathogens causing infections of skin, soft tissues, and bones. In: The microbiology of skin, soft tissue, bone and joint infections. Kidlington (United Kingdom): Elsevier. p. 105–121

Stahl DA, Urbance J. 1990. The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. J Bacteriol. 172(1):116–124.

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

Suyama M, Bork P. 2001. Evolution of prokaryotic gene order: genome rearrangements in closely related species. Trends Genet. 17(1):10–13.

Tauch A, et al. 2005. Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. J Bacteriol. 187(13):4671–4682.

Vachaspati P, Warnow T. 2015. ASTRID: accurate species trees from internode distances. BMC Genomics. 16(S10):S3.

Van Duin D, Paterson DL. 2016. Multidrug-resistant bacteria in the community: trends and lessons learned. Infect Dis Clin. 30(2):377–390.

Von Graevenitz A, Bernard K. 2006. The genus *Corynebacterium*—medical. In: The prokaryotes: volume archaea. Bacteria: *Firmicutes*, actinomycetes. New York: Springer. p. 819–842.

Wolf YI, Rogozin IB, Grishin NV, Koonin EV. 2002. Genome trees and the tree of life. Trends Genet. 18(9):472–479.

Yang J, Yang S. 2017. Comparative analysis of *Corynebacterium glutamicum* genomes: a new perspective for the industrial production of amino acids. BMC Genomics. 18(S1):940.

Yap WH, Zhang Z, Wang Y. 1999. Distinct types of rRNA operons exist in the genome of the actinomycete *Thermomonospora chromogena* and evidence for horizontal transfer of an entire rRNA operon. J Bacteriol. 181(17):5201–5209.

Zhi X-Y, Jiang Z, Yang L-L, Huang Y. 2017. The underlying mechanisms of genetic innovation and speciation in the family Corynebacteriaceae: a phylogenomics approach. Mol Phylogenet Evol. 107:246–255.

Associate editor: Brian Golding