Infection, Genetics and Evolution 56 (2017) 19–25

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

The new phylogeny of the genus *Mycobacterium*: The old and the news

Enrico Tortoli[a,*,1], Tarcisio Fedrizzi[b,1], Conor J. Meehan[c,1], Alberto Trovato[a], Antonella Grottola[d], Elisabetta Giacobazzi[b], Giulia Fregni Serpini[d], Sara Tagliazucchi[d], Anna Fabio[d], Clotilde Bettua[b], Roberto Bertorelli[e], Francesca Frascaro[d], Veronica De Sanctis[e], Monica Pecorari[d], Olivier Jousson[b], Nicola Segata[b,2], Daniela M. Cirillo[a,2]

[a] *Emerging Bacterial Pathogens Unit, IRCCS San Raffaele Scientific Institute, Milano, Italy*
[b] *Centre for Integrative Biology, University of Trento, Trento, Italy*
[c] *Mycobacteriology unit, Department of Biomedical Science, Institute of Tropical Medicine, Antwerp, Belgium*
[d] *Microbiology and Virology Unit, University Hospital Polyclinic, Modena, Italy*
[e] *NGS Facility, Laboratory of Biomolecular Sequence and Structure Analysis for Health, Centre for Integrative Biology, University of Trento, Italy*

ARTICLE INFO

Keywords:
Mycobacterium
Whole genome sequencing
Phylogeny
Average nucleotide identity

ABSTRACT

Background: Phylogenetic studies of bacteria have been based so far either on a single gene (usually the 16S rRNA) or on concatenated housekeeping genes. For what concerns the genus *Mycobacterium* these approaches support the separation of rapidly and slowly growing species and the clustering of most species in well-defined phylogenetic groups. The advent of high-throughput shotgun sequencing leads us to revise conventional taxonomy of mycobacteria on the light of genomic data. For this purpose we investigated 88 newly sequenced species in addition to 60 retrieved from GenBank and used the Average Nucleotide Identity pairwise scores to reconstruct phylogenetic relationships within this genus.

Results: Our analysis confirmed the separation of slow and rapid growers and the intermediate position occupied by the *M. terrae* complex. Among the rapid growers, the species of the *M. chelonae-abscessus* complex belonged to the most ancestral cluster. Other major clades of rapid growers included the species related to *M. fortuitum* and *M. smegmatis* and a large grouping containing mostly environmental species rarely isolated from humans. The members of the *M. terrae* complex appeared as the most ancestral slow growers. Among slow growers two deep branches led to the clusters of species related to *M. celatum* and *M. xenopi* and to a large group harboring most of the species more frequently responsible of disease in humans, including the major pathogenic mycobacteria (*M. tuberculosis*, *M. leprae*, *M. ulcerans*). The species previously grouped in the *M. simiae* complex were allocated in a number of sub-clades; of them, only the one including the species *M. simiae* identified the real members of this complex. The other clades included also species previously not considered related to *M. simiae*. The ANI analysis, in most cases supported by Genome to Genome Distance and by Genomic Signature-Delta Difference, showed that a number of species with standing in literature were indeed synonymous.

Conclusions: Genomic data revealed to be much more informative in comparison with phenotype. We believe that the genomic revolution enabled by high-throughput shotgun sequencing should now be considered in order to revise the conservative approaches still informing taxonomic sciences.

1. Background

Distinct ages have characterized systematic sciences. Since the first attempts of taxonomic classification (Linnaeus, 1735), comparative analysis of phenotypic traits has, with minor innovations (*e.g.* numerical taxonomy (Sneath and Sokal, 1973)), dominated for a long time. In the last half century, the role of phenotype has been progressively eroded by the unrestrainable progress of genetics. The availability of

* Corresponding author.
E-mail addresses: tortoli.enrico@hsr.it (E. Tortoli), t.fedrizzi@unitn.it (T. Fedrizzi), cmeehan@itg.be (C.J. Meehan), trovato.alberto@hsr.it (A. Trovato), grottola.antonella@policlinico.mo.it (A. Grottola), tagliazucchi.sara@policlinico.mo.it (S. Tagliazucchi), fabio.anna@policlinico.mo.it (A. Fabio), clotilde.bettua@unitn.it (C. Bettua), roberto.bertorelli@unitn.it (R. Bertorelli), francescafrascaro@libero.it (F. Frascaro), veronica.desanctis@unitn.it (V. De Sanctis), pecorari.monica@policlinico.mo.it (M. Pecorari), olivier.jousson@unitn.it (O. Jousson), nicola.segata@unitn.it (N. Segata), cirillo.daniela@hsr.it (D.M. Cirillo).
[1] These authors contributed equally.
[2] These authors contributed equally.

http://dx.doi.org/10.1016/j.meegid.2017.10.013
Received 4 August 2017; Received in revised form 9 October 2017; Accepted 10 October 2017
Available online 11 October 2017

sequencing and bioinformatics tools suitable to produce reliable sequences of the whole genomes in short time is definitively changing the scenario. Prokaryotes, better than others, can take advantage from this revolution as it is relatively well established how to perform whole-genome comparisons and phylogenies with a large number of strains.

With the aim of updating the taxonomic relationships between the members of the genus *Mycobacterium*, we sequenced the whole genomes of 88 strains of *Mycobacterium* species (including 41 sequenced in a recent study (Fedrizzi et al., 2017)) and, combined with 60 already present in GenBank, undertook a genus-wide comparative genomic study.

2. Methods

We sequenced the strains reported in Supplementary Table 1 on Illumina platforms (HiSeq 2000 and MiniSeq); 41 of them were part of a previous study (Fedrizzi et al., 2017). Reads were initially trimmed using Trim Galore (–phred33 –nextera –stringency 2 –fastqc –paired –retain_unpaired) and assembled with SPAdes ver. 3.9.1 software (Bankevich et al., 2012) (–carefull –cov-cutoff auto). The assembled genomes were quality controlled with QUAST ver. 4.2 (Gurevich et al., 2013). The annotation of genomes was performed by either resorting to the NCBI Prokaryotic Genome Annotation Pipeline (Angiuoli et al., 2008) (47 strains) or with PROKKA (Seemann, 2014) (41 strains). Quality estimates of the genomes sequenced in this study are given in Table 1.

All strains selected for sequencing but one (*Mycobacterium elephantis*), were type strains of different *Mycobacterium* species obtained from various international culture collections.

Of the 60 genomes retrieved from GenBank (Supplementary Table 2), 50 were from type strains. The identification at the species level of strains other than type was carefully verified and the ones with 16S rRNA gene identity < 99.4% in comparison with the sequence of respective type strain (including several genomes evidently mislabeled), were excluded from the study.

To determine the relatedness of the strains, the Average Nucleotide Identity (ANI) was employed using the assembled genomes as the representative data for each species. The ANI among respective genomes was calculated using the OrthoANI calculator (Lee et al., 2016a) choosing the "both directions" option. A distance matrix of pairwise ANI values converted to ANI-divergence (defined as complement to 100 of ANI data) (Chan et al., 2012) was used to reconstruct, with MEGA ver. 6.06 software (Tamura et al., 2013), the phylogeny. Two different phylogenetic trees were constructed using UPGMA (Sokal and Michener, 1958) and Neighbor Joining (Saitou and Nei, 1987) algorithms. In both analyses *Hoyosella altamirensis*, a member of *Mycobacteriaceae* family not belonging to the genus *Mycobacterium* (Hamada et al., 2016), was included as outgroup.

The ANI threshold of 95–96% is universally accepted for species demarcation (Goris et al., 2007; Kim et al., 2014; Richter and Rosselló-Mora, 2009). Whenever the ANI between two or more strains classified as belonging to different species scored ≥ 95%, the taxonomic attribution was checked using two other computational approaches. The Genome to Genome Distance (GGD) (Auch et al., 2010) was determined from assembled genomes using the GGD calculator (http://ggdc.dsmz.de/) according to formula 2, as recommended. The Genomic Signature-Delta Difference (GS-DD) (Karlin et al., 1997) was measured on respective annotated genomes using the software available at http://www.cmbl.uga.edu/software/delta-differences.html.

The genome of *Mycobacterium tuberculosis*^{T} was chosen as representative of the whole *M. tuberculosis* complex due to the ANI values close to 100% among all its members.

One genome retrieved from GenBank, *Mycobacterium hodleri*^{T} (BBGO00000000) produced 39 out of 147 invalid ANI pairwise results and was excluded from the analysis.

Table 1
WGS data of the strains sequenced in present study.

Species	Contigs no.	Genome size (bp)	G + C%	CDS	Mean coverage
M. algericum^T	107	4,619,277	68.33	4380	70.5
M. agri^T	385	7,002,017	66.48	6724	35.6
M. alsense^T	177	5,695,622	69.30	5277	77.1
M. angelicum^T	543	6,674,259	66.25	5804	20.2
M. aquaticum^T	111	7,927,623	66.48	7614	48.1
M. arosiense^T	497	5,984,789	66.83	5637	18.7
M. arupense^T	157	4,442,590	67.38	4126	15.0
M. asiaticum^T	137	5,978,015	66.39	5535	85.3
M. bacteremicum^T	98	5,953,791	68.11	5669	39.9
M. boenickei^T	101	6,546,474	66.84	6262	45.9
M. abscessus subsp. *bolletii*^{T}	30	5,061,466	64.06	4933	38.7
M. bouchedurhonnense	233	5,900,850	68.58	5630	18.1
M. branderi^T	64	5,903,581	66.51	5748	44.4
M. brumae^T	124	3,882,015	69.26	3663	98.2
M. celeriflavum^T	129	4,950,353	66.89	4786	166.9
M. chubuense^T	121	5,948,704	69.20	5637	48.7
M. diernhoferi^T	125	5,978,092	67.90	5674	37.0
"*M. decipiens*^{T}"	165	5,228,890	65.52	4676	44.9
M. duvalii^T	111	5,605,497	68.30	5330	51.6
M. elephantis	84	5,313,693	67.64	5094	58.2
M. franklinii^T	68	5,442,959	64.14	5402	27.6
M. heidelbergense^T	126	5,001,088	67.97	4615	53.4
M. insubricum^T	159	4,553,954	68.85	4252	32.3
M. intermedium^T	387	6,818,427	65.78	6076	27.3
M. koreense^T	153	4,088,537	69.33	3979	145.1
M. kumamotonense^T	148	4,818,368	68.12	4547	81.3
M. malmoense^T	134	5,301,070	66.99	4884	56.5
M. mantenii^T	157	6,122,108	66.89	5680	40.8
M. marseillense	112	5,456,935	67.71	5059	12.5
M. abscessus subsp. *massiliense*^{T}	30	4,967,937	64.09	4893	60.3
M. minnesotense^T	73	4,189,414	67.12	3852	79.8
M. monacense^T	126	6,003,959	68.41	5751	33.2
M. moriokaense^T	232	6,217,973	66.02	6095	126.0
M. noviomagense^T	249	4,744,606	65.73	4556	9.4
M. parafortuitum^T	88	6,136,108	68.53	5747	41.1
M. paraintracellulare^T	136	5,477,232	68.08	5153	122.2
M. paraseoulense^T	282	6,080,533	67.88	5691	23.5
M. persicum^T	389	6,172,652	66.20	5621	11.0
M. porcinum^T	102	6,960,672	66.74	6651	43.6
M. rhodesiae^T	72	6,011,218	66.59	5758	43.8
M. saopaulense^T	103	5,200,319	64.58	5225	49.3
M. scrofulaceum^T	224	6,175,554	68.42	5806	43.0
M. shinjukuense^T	181	4,409,975	67.76	3981	86.7
M. sphagni^T	105	6,070,288	65.92	5842	39.7
M. timonense^T	1222	6,013,961	68.51	5654	10.5
M. tusciae^T	71	6,163,099	65.69	6035	65.6
M. vulneris^T	158	6,270,094	66.77	5868	45.9

3. Results

The UPGMA tree built on 10,878 (148 × 147/2) ANI-divergence scores presented two major branches clearly differentiating rapidly and slowly growing *Mycobacterium* species (Fig. 1). Genetic diversity within these two subtrees showed the presence of well-defined, monophyletic sub-clades that included the large majority of the considered species.

A similar topology was obtained by the Neighbor Joining analysis. Among the rapid growers, a deep branch separated the species of the *Mycobacterium abscessus-chelonae* complex (Fig. 2), which appeared to be the most ancestral, from all others. This complex comprised nine taxa, including three for which the pairwise ANI scores, ranging from 97.2 to 97.4%, did not support the status of independent species and suggested instead that they represent subspecies of *Mycobacterium abscessus* (Tortoli et al., 2016). GGD and GS-DD lead to identical conclusions (Table 2), further supported by previously determined DDH (Leao et al., 2009).

Another major cluster within rapid growers encompassed the *Mycobacterium fortuitum-smegmatis* group (Fig. 2). In this cluster of 17

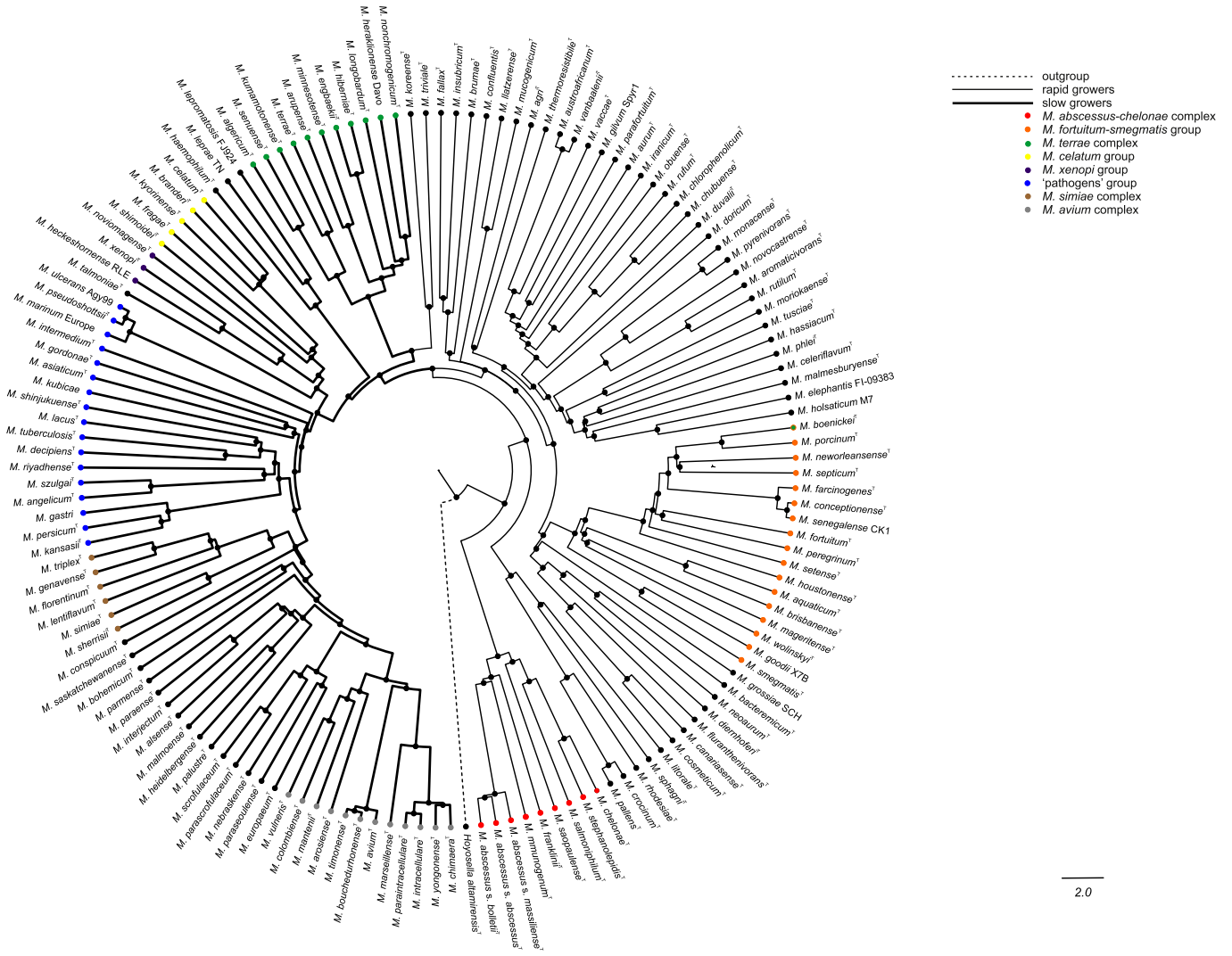

Fig. 1. Phylogenetic tree of the genus *Mycobacterium* constructed using the UPGMA algorithm, from the distance matrix of 10,878 ANI-divergence scores. Bar, 2 units difference in ANI-divergence value.

taxa, *Mycobacterium conceptionense*, *Mycobacterium farcinogenes* and *Mycobacterium senegalense* were very closely related and their pairwise ANI values (range 98.3–99.4) suggested they could represent variants of a single species. This possibility was confirmed by GGD for the three strains and by GS-DD for *M. conceptionense* and *M. senegalense* only (Table 2).

The remaining rapidly growing mycobacteria were characterized by deep branches which led to 46 species (Fig. 2) of mainly environmental mycobacteria only rarely, if ever, isolated from humans or animals. Among them most are known to possess aromatic hydrocarbon degradation activities (Kweon et al., 2015). Two species presenting slow growing phenotype, *Mycobacterium doricum* and *Mycobacterium tusciae*, belonged to this subtree. Pairwise ANI values between *Mycobacterium austroafricanum* and *Mycobacterium vanbaalenii* (98.7%) did not support their status of independent species as confirmed also by GGD and GS-DD (Table 2).

From an early ramification at the root of the branch leading to the evolution of slow growers originated *Mycobacterium triviale* and *Mycobacterium koreense*, the first of them has been assigned to the *Mycobacterium terrae* complex in the past (Wayne and Kubica, 1986). A parallel group included 11 taxa corresponding to the number of members of *M. terrae* complex investigated in the present study (Fig. 3). A clear-cut separation characterized the mycobacteria of this grouping with the exception of *Mycobacterium engbaekii* and *Mycobacterium hiberniae* presenting an ANI of 95.1%, that is inconclusive in determining whether or not they belong to the same species. For the latter two species, GS-DD and GGD supported however their fusion in a single taxon (Table 4).

The slow growers other than *M. terrae* complex were distributed among three major groups. The eight taxa present in the first one were subdivided in two subgroups presenting, as more representative species, *Mycobacterium celatum* and *Mycobacterium xenopi* (Fig. 3). In a larger group were included, in adjunct to pathogens (*M. tuberculosis* complex, *Mycobacterium leprae*, "*Mycobacterium lepromatosis*", *Mycobacterium ulcerans*), most of the species frequently isolated from humans and often associated to disease. Three strains, *Mycobacterium marinum*, *Mycobacterium pseudoshottsii and Mycobacterium ulcerans* presented pairwise ANI values supporting their possible inclusion in a single species (range 98.0–99.1). Both GGD and GD-DD confirmed this conclusion (Table 2).

The 33 members of the most-derived, large cluster of slow growers were organized in a series of sub-clusters. One of them included six species (Fig. 3) closely related to *Mycobacterium simiae*, while 12 taxa constituting the most marginal sub-clade (Fig. 3) belonged to the presently acknowledged *Mycobacterium avium* complex (Lee et al., 2016b). Pairwise ANI percentages insufficient to warrant the status of independent species were obtained for two groups, including four and three taxa, respectively. ANI values ranged from 97.6 to 97.8% for

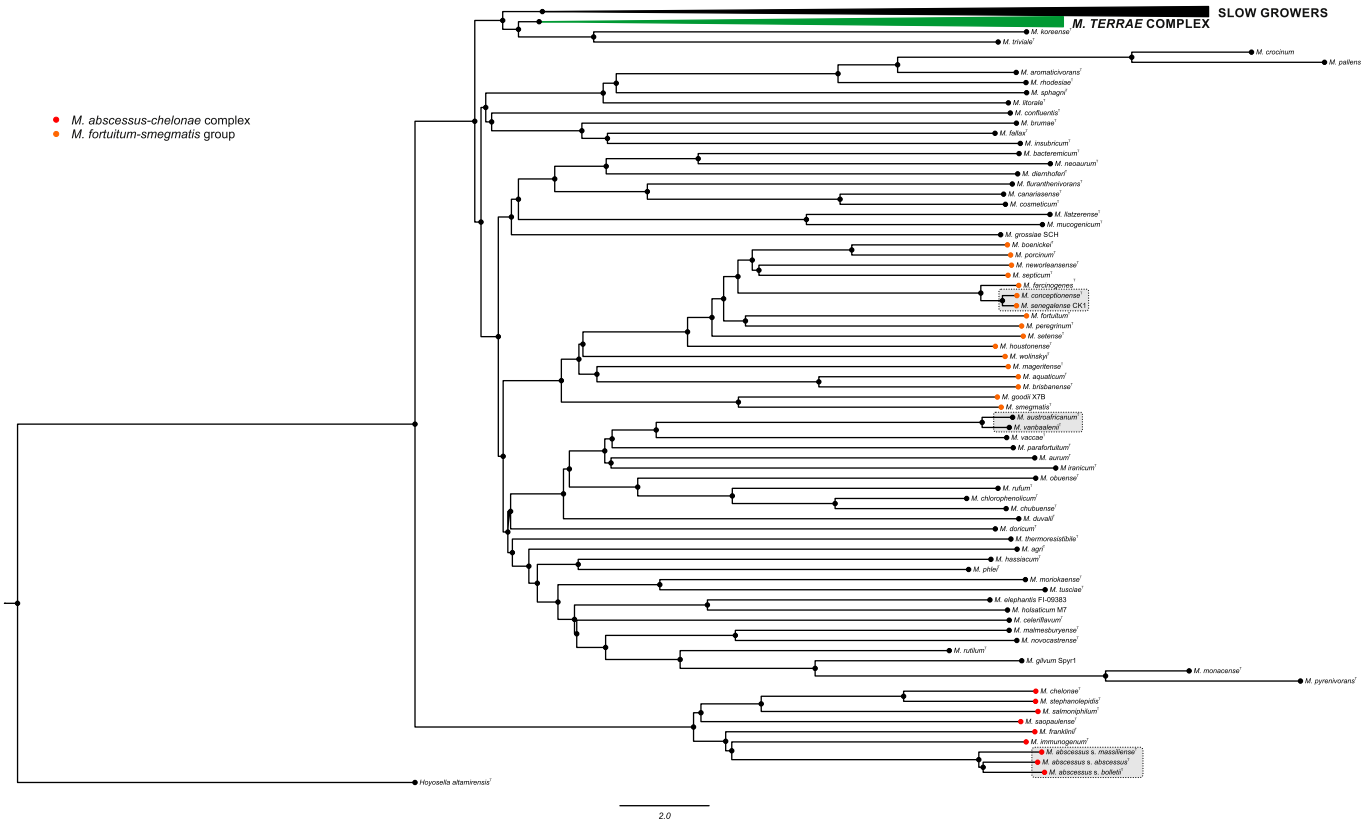

Fig. 2. Neighbor-joining tree constructed on 10,878 ANI-divergence scores, rooted using *H. altamirensis* as outgroup. Bar, 2 units difference in ANI-divergence value.

Table 2
Comparison of Average Nucleotide Identity, Genome to Genome Distance and Genomic Signature Delta Difference data of closely related *Mycobacterium* species. ANI values > 96 support the belonging to the same species. GGD values > 70% support the belonging to the same species. δ < δ* values support the belonging to the same species.

Species		ANI%	GGD%	GS-DD	Score[a]
M. conceptionense	*M. farcinogenes*	98.3	86	δ > δ*	2/3
M. conceptionense	*M. senegalense*	99.4	84	δ < δ*	3/3
M. farcinogenes	*M. senegalense*	98.3	83	δ > δ*	2/3
M. aromaticivorans	*M. pallens*	96.5	13	δ > δ*	1/3
M. pallens	*M. crocinum*	98.3	13	δ > δ*	1/3
M. abscessus s. abscessus	*M. abscessus s. bolletii*	97.4	86	δ < δ*	3/3
M. abscessus s. abscessus	*M. abscessus s. massiliense*	97.2	85	δ < δ*	3/3
M. abscessus s. massiliense	*M. abscessus s. bolletii*	97.2	88	δ < δ*	3/3
M. chimaera	*M. intracellulare*	97.6	77	δ < δ*	3/3
M. chimaera	*M. yongonense*	97.8	79	δ < δ*	3/3
M. yongonense	*M. intracellulare*	97.6	85	δ < δ*	3/3
M. engbaekii	*M. hiberniae*	95.1	89	δ < δ*	2/3
M. austroafricanum	*M. vanbaalenii*	98.7	80	δ < δ*	3/3
M. gilvum	*M. pyrenivorans*	96.6	13	δ > δ*	1/3
M. novocastrense	*M. pyrenivorans*	96.8	13	δ > δ*	1/3
M. pyrenivorans	*M. vanbaalenii*	97.1	13	δ > δ*	1/3
M. marinum	*M. pseudoshottsii*	98.2	82	δ < δ*	3/3
M. marinum	*M. ulcerans*	98.0	80	δ < δ*	3/3
M. ulcerans	*M. pseudoshottsii*	99.1	90	δ < δ*	3/3
M. monacense	*M. pyrenivorans*	97.5	84	δ > δ*	2/3

[a] Proportion of tests supporting belonging to the same species.

Mycobacterium chimaera, *Mycobacterium intracellulare*, *Mycobacterium yongonense* and *Mycobacterium paraintracellulare*, and were > 98.8 for *M. avium*, *Mycobacterium bouchedurhonense* and *Mycobacterium timonense*.

4. Discussion

The consistency of trees inferred using different algorithms is considered a reliable measure of robustness of phylogenetic reconstructions. In the present study the topologies obtained with UPGMA (Fig. 1) and Neighbor Joining (Figs. 2 and 3) were characterized by very high similarity. Furthermore the phylogeny of a smaller number of species reconstructed in a recent study (Fedrizzi et al., 2017), either from the genetic sequence of 243 concatenated conserved genes, or from gene presence/absence clustering, was fully in agreement with ours. Rapid growers were more ancestral than slow growers, the *M. terrae* complex occupied an intermediate position, and the species composition of different groupings overlapped.

Defining the taxonomic groupings within a genus historically relies upon phenotypic characteristics or portions of the genome such as 16S rRNA or genes shared among all involved species. With the advent of rapid whole genome sequencing approaches, integration of such data into taxonomic definitions is becoming crucial. Here we utilized ANI, a whole genome comparison approach, on the genus *Mycobacterium*, and found strong support for many, but not all, existing taxonomic groups and species.

The phylogeny of the genus *Mycobacterium* has been so far based either on 16S rRNA sequences (Rogall et al., 1990; Stahl and Urbance, 1990; Tortoli, 2012) or on concatenated sequences of a number of housekeeping genes (Devulder et al., 2005; Mignard and Flandrois, 2008; Tortoli, 2012). The one based on 16S rRNA has been frequently an object of criticisms because of the unusually high intraspecies identity (ranging from 94 to 100%) which makes impossible the distinction of a number of *Mycobacterium* species (Table 3). Beyond this limitation, it is substantially in agreement with the reconstructions emerging from the multilocus approaches and benefits from the presence, in the helix 18 of the 16S rRNA gene, of unique, phylogenetic signatures (Kirschner et al., 1993). In this helix: the rapidly growing

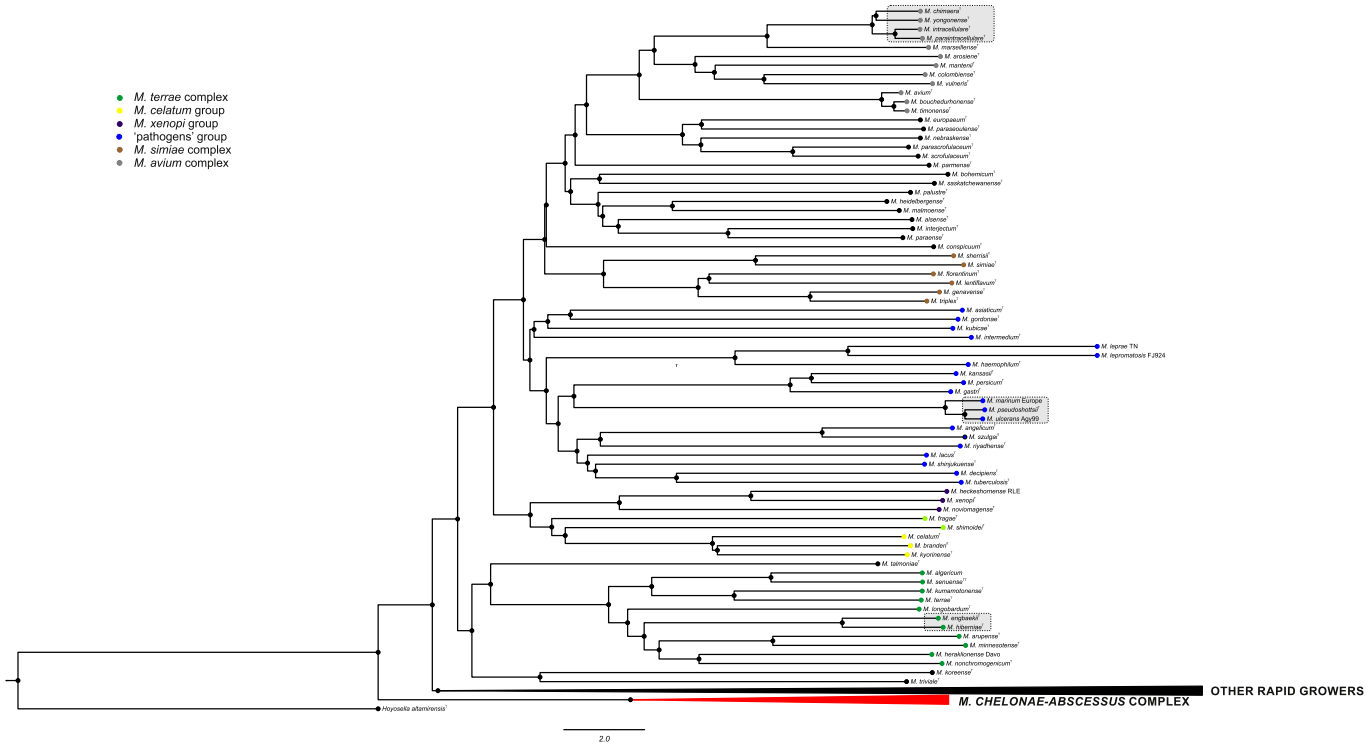

Fig. 3. Neighbor-joining tree constructed on 10,878 ANI-divergence values rooted using *H. altamirensis* as outgroup. Bar, 2 units difference in ANI-divergence value.

Table 3
Species presenting identical 16 rRNA sequences in the first 500 bp and the in the whole gene.

Species		500 bp	Total
M. abscessus all subsp.	*M. chelonae*	100%	99,7%
M. alvei	*M. setense*	100%	99,1%
M. angelicum	*M. szulgai*	100%	100%
M. conceptionense	*M. houstonense*	100%	99,7%
M. conceptionense	*M. senegalense*	100%	100%
M. farcinogenes	*M. fortuitum*	100%	100%
M. farcinogenes	*M. houstonense*	100%	100%
M. farcinogenes	*M. senegalense*	100%	100%
M. fortuitum	*M. houstonense*	100%	100%
M. fortuitum	*M. senegalense*	100%	100%
M. gastri	*M. kansasii*	100%	100%
M. houstonense	*M. senegalense*	100%	100%
M. intracellulare	*M. paraintracellulare*	100%	100%
M. marinum	*M. ulcerans*	100%	99,8%
M. marseillense	*M. yongonense*	100%	100%
M. mucogenicum	*M. phocaicum*	100%	100%
M. murale	*M. tokaiense*	100%	100%
M. paraseoulense	*M. seoulense*	100%	100%
M. peregrinum	*M. septicum*	100%	99,7%
M. vaccae	*M. vanbaalenii*	100%	99,3%

species do not present insertions; the majority of slow growers have a 12 nucleotide insertion; the members of the *M. terrae* complex have a 14 nucleotide insertion; a cluster of slow growers, known as *M. simiae* complex, are devoid of insertion but are characterized by a specific sequence not found in other mycobacteria. Exceptions to these features are extremely rare.

In general, our survey covering more than 80% of the officially recognized species (*M. tuberculosis* complex is regarded here as a single species) was largely in agreement with conventional phylogeny but allowed us to suggest additional aspects not highlighted by previous phylogenetic approaches.

The rapid growers were confirmed to be the most ancestral mycobacteria with the *M. chelonae-abscessus* complex being the closest to the root of the genus (Figs. 1 and 2). All the species expected to cluster in one of the major groups of rapid growers fulfilled expectations, with the only unpredicted finding being the position of *Mycobacterium wolinskyi* closer to *M. fortuitum* than to *M. smegmatis*. Even the "anomalous" location of the slowly growing species *M. doricum* and *M. tusciae* within the evolutionary lineages of rapid growers (Devulder et al., 2005; Mignard and Flandrois, 2008; Tortoli, 2012) was confirmed. In contrast *Mycobacterium holsaticum*, placed within slow growers by conventional phylogeny (Devulder et al., 2005), was repositioned, according to its phenotype, among rapid growers in our study. Once more, we confirmed the groundlessness of the inclusion of *M. triviale* within the *M. terrae* complex. This species, along with *M. koreense*, remained in our analysis as the most likely evolutionary link between rapid and slow growers.

The attribution of the *M. terrae* complex either to rapidly or slowly growing species, controversial according to the multilocus phylogeny (Mignard and Flandrois, 2008), was settled in agreement with 16S based reconstructions locating this group at the root of slow growers.

The phylogeny of slowly growing species highlighted major inconsistencies in the conventional knowledge. The traditional *M. simiae* complex, as defined by the specific signatures within the helix 18 of 16S rRNA, was redefined by our analysis. Only seven species (including *M. simiae*), clustered clearly together while the others were included among sister clades which comprised also other mycobacteria (*Mycobacterium alsense*, *Mycobacterium malmoense*, *Mycobacterium bohemicum*) not presenting the specific helix 18 signature above (Fig. 3). *Mycobacterium europaeum* and *Mycobacterium parascrofulaceum* in particular were very divergent from *M. simiae* and more closely related to the *M. avium* complex. The signature in the helix 18, previously considered a reliable marker for belonging to the *M. simiae* complex, came out deeply challenged by our findings and proved insufficiently specific, suggesting it can be a consequence of convergent evolution.

Although the genome-based phylogeny has produced only minor changes in comparison with the conventional reconstructions, it is now evident that phylogenetic studies disregarding the whole genome are outdated.

Table 4
Emendation proposals for the species proven belonging to the same taxon.

Species			Suggested action	New name
M. austroafricanum	*M. vanbaalenii*		Fusion	*M. austroafricanum*
M. conceptionense	*M. senegalense*		Fusion	*M. senegalense*
M. chimaera	*M. intracellulare*	*M. yongonense*	Subspeciation	*M. intracellulare* subsp. *intracellulare* subsp. *chimaera* subsp. *yongonense*
M. hiberniae	*M. engbaekii*		Subspeciation	*M. hiberniae* subsp. *hiberniae* subsp. *engbaekii*
M. intracellulare	*M. paraintracellulare*		Fusion	*M. intracellulare*
M. marinum	*M. pseudoshottsii*		Fusion	*M. marinum*

For defining the boundaries of bacterial species the DNA-DNA hybridization (DDH) method (Wayne et al., 1987) is still considered the gold standard. Such approach necessarily uses hard cut-offs to define whether two strains belong to the same species or not, and the threshold of 70% DDH, although inevitably arbitrary, has proved acceptable. More recent approaches based on whole genome similarity are also suitable to this purpose; in particular the ANI, the GGD (the *in silico* equivalent of DDH) and the GS-DD are widely used scores. All of them have been compared with DNA-DNA hybridization and *ad hoc* cut-offs have been defined (Auch et al., 2010; Goris et al., 2007; Karlin et al., 1997; Kim et al., 2014). Among the four tests mentioned above, the genomic ones are by far more reproducible than the DDH, with results of the latter presenting often inconsistent and dependent upon the methodology adopted. With three alternative tests available, and probably others coming in the future, DDH seems destined to remain relevant at historical level only. It is however also important to point out that the three genomic approaches do not produce fully overlapping results. This implies that, in the case of borderline values, two genomes of closely related strains can be regarded, according to the test used, as belonging either to a single, or to different, species. In this study, out of 20 cases in which pairwise ANI values between closely related species were supportive for belonging to a single taxon, the use of GGD and GS-DD confirmed the finding in eleven. In four such cases ANI and GGD were in agreement while in the remaining five cases the possibility of representing synonymous species was supported by ANI alone. ANI appeared therefore to be the most sensitive of the three measures in avoiding inappropriate classification to independent taxa of strains representing with high probability simple variants. To avoid the duplication of already recognized species in the future, we thus recommend the ANI score as the test of choice. For the past, a more conservative approach (consisting in emending the species confirmed illegitimate by agreeing results of ANI, GGD and GS-DD) could be adopted. Our proposals for the species investigated in this study are reported in Table 4. For the species *M. marinum*, *M. pseudoshottsii* and *M. ulcerans* an exception was made. Although their pairwise ANI scores are similar, we suggest, because of the specificity of disease produced in humans and the different host specificity, that *M. marinum* and *M. pseudoshottsii* are combined while leaving *M. ulcerans* as independent species. This approach is not different from the one universally accepted for the species of the *M. tuberculosis* complex and is furthermore supported by the massive difference in their genomes size.

In this study, we also found seven mislabeled *Mycobacterium* genomes deposited in GenBank (Table 5), which is alarming considering that public genomes are likely to be extensively used for downstream genomic and clinical investigations. A strict policy to prevent deposition of misleading sequences should be established and effort should be put in identifying and correcting such inconsistencies in public repositories.

A special case was the discrepancy we detected with *M. bouchedurhonense* and *M. timonense*. Their 16S rRNA genes were unexpectedly most closely related to *M. avium*^{T} (100% identity and one mismatch only, respectively) while differed of 8 and 13 bp from the sequences deposited in GenBank for *M. bouchedurhonense*^{T} and *M. timonense*^{T}, respectively. This implies that the two strains deposited in the culture collections do not match the ones reported in the sp. nov. descriptions (Ben Salah et al., 2009), probably due to a strain exchange. This finding was confirmed, upon our information, by respective culture collections.

The easy and affordable availability of reliable whole-genome sequences raises doubts about the real added value of investigating phenotypic traits when a new species is described. Actually, different taxonomists use their own panels of tests, often not standardized, to produce results of no use for colleagues and absolutely incomprehensible to the community of mycobacteriologists who have dismissed such approach since the '90s. For the genus *Mycobacterium* the major phenotypic traits that cannot be disregarded should include growth rate and pigmentation of colonies, while the classical investigation of biochemical activities is clearly obsolete. In contrast, because of the universality and completeness of their results, genotypic investigations need to be expanded and the determination of the whole genome sequence of every newly described species should be added to the presently required analysis of a number of housekeeping genes (Stackebrandt et al., 2002). If not already available in public domain databases, the genome of the species showing, at level of the 16S rRNA gene, the closest similarity should be sequenced as well; excluding the cases in which the level of such similarity (< 98.7%) (Kim et al., 2014; Stackebrandt and Ebers, 2006) indicate this analysis as not mandatory. It is probably time to modernize the policies still at the basis of taxonomic studies which are still too often regarded as "dogmas".

Table 5
Genomes proven mislabeled, present in GenBank.

Accession n.	Label	Corrected identification	ANI *vs* the type strain of the same species	Closest known species
CVTC00000000	*M. nonchromogenicum*	*M. fortuitum*	75.2	
CCBG00000000	*M. vulneris*	*M. porcinum*	75.7	
AGIQ00000000	*M. rhodesiae*	undescribed species	83.7	*M. rhodesiae*
CP003169	*M. rhodesiae*	undescribed species	76.7	*M. aurum*
LZLW00000000	*M. szulgai*	*M. kubicae*	79.8	
AGJJ00000000	*M. tusciae*	undescribed species	84.7	*M. austroafricanum*
CP003078	*M. smegmatis*	undescribed species	77.0	*M. moriokaense*

Declarations

Ethics approval and consent to participate. NOT APPLICABLE.

Consent for publication. NOT APPLICABLE.

Availability of data and material. The datasets generated and analyzed during the current study are available in the GenBank repository (see Supplementary Table 1).

Competing interests. The authors declare that they have no competing interests.

Funding. No funding for present study.

Author contribution. ET conceived the study and wrote the manuscript; TF and MP analyzed genomic data; CJM and NS analyzed genomic data and critically revised the manuscript; AG, OJ and DMC critically revised the manuscript; AT, EG, GFS, ST, AF, CB, RB, FF, and DS performed the isolation, cultivation, and sequencing tasks. All authors read and approved the final version of manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2017.10.013.

References

Angiuoli, S.V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G., Kodira, C.D., Kyrpides, N., Madupu, R., Markowitz, V., Tatusova, T., Thomson, N., White, O., 2008. Toward an online repository of Standard Operating Procedures (SOPs) for (meta) genomic annotation. OMICS 12, 137–141.

Auch, A.F., von Jan, M., Klenk, H.P., Goker, M., 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand. Genomic Sci. 2, 117–134.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Ben Salah, I., Cayrou, C., Raoult, D., Drancourt, M., 2009. *Mycobacterium marseillense* sp. nov., *Mycobacterium timonense* sp. nov. and *Mycobacterium bouchedurhonense* sp. nov., novel species in the *Mycobacterium avium* complex. Int. J. Syst. Evol. Microbiol. 59, 2803–2808.

Chan, J.Z., Halachev, M.R., Loman, N.J., Constantinidou, C., Pallen, M.J., 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. BMC Microbiol. 12, 302.

Devulder, G., Perouse, d.M., Flandrois, J.P., 2005. A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. Int. J. Syst. Evol. Microbiol. 55, 293–302.

Fedrizzi, T., Meehan, C.J., Grottola, A., Giacobazzi, E., Fregni Serpini, G., Tagliazucchi, S., Fabio, A., Bettua, C., Bertorelli, R., De Sanctis, V., Rumpianesi, F., Pecorari, M., Jousson, O., Tortoli, E., Segata, N., 2017. Genomic characterization of non-tuberculous mycobacteria. Sci Rep 7, 45258.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. 57, 81–91.

Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29, 1072–1075.

Hamada, M., Shibata, C., Sakurai, K., Hosoyama, A., Oji, S., Teramoto, K., Tamura, T., 2016. Reclassification of *Amycolicicoccus subflavus* as *Hoyosella subflava* comb. nov. and emended descriptions of the genus *Hoyosella* and *Hoyosella altamirensis*. Int. J. Syst. Evol. Microbiol. 66, 4711–4715.

Karlin, S., Mrazek, J., Campbell, A.M., 1997. Compositional biases of bacterial genomes and evolutionary implications. J. Bacteriol. 179, 3899–3913.

Kim, M., Oh, H.S., Park, S.C., Chun, J., 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int. J. Syst. Evol. Microbiol. 64, 346–351.

Kirschner, P., Springer, B., Vogel, U., Meier, A., Wrede, A., Kiekenbeck, M., Bange, F.C., Böttger, E.C., 1993. Genotypic identification of mycobacteria by nucleic acid sequence determination: report of a 2-year experience in a clinical laboratory. J. Clin. Microbiol. 31, 2882–2889.

Kweon, O., Kim, S.J., Blom, J., Kim, S.K., Kim, B.S., Baek, D.H., Park, S.I., Sutherland, J.B., Cerniglia, C.E., 2015. Comparative functional pan-genome analyses to build connections between genomic dynamics and phenotypic evolution in polycyclic aromatic hydrocarbon metabolism in the genus *Mycobacterium*. BMC Evol. Biol. 15, 21.

Leao, S.C., Tortoli, E., Viana-Niero, C., Ueki, S.Y., Lima, K.V., Lopes, M.L., Yubero, J., Menendez, M.C., Garcia, M.J., 2009. Characterization of mycobacteria from a major Brazilian outbreak suggests a revision of the taxonomic status of members of the *Mycobacterium chelonae-abscessus* group. J. Clin. Microbiol. 47, 2691–2698.

Lee, I., Kim, Y.O., Park, S.C., Chun, J., 2016a. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int. J. Syst. Evol. Microbiol. 66, 1100–1103.

Lee, S.Y., Kim, B.J., Kim, H., Won, Y.S., Jeon, C.O., Jeong, J., Lee, S.H., Lim, J.H., Lee, S.H., Kim, C.K., Kook, Y.H., Kim, B.J., 2016b. *Mycobacterium paraintracellulare* sp. nov., for the genotype INT-1 of *Mycobacterium intracellulare*. Int. J. Syst. Evol. Microbiol. 66, 3132–3141.

Linnaeus, C., 1735. Systema naturae, sive regna tria naturae systematice proposita per classes, ordines, genera et species. Apud Theodorum Haak, Leiden.

Mignard, S., Flandrois, J.P., 2008. A seven-gene, multilocus, genus-wide approach to the phylogeny of mycobacteria using supertrees. Int. J. Syst. Evol. Microbiol. 58, 1432–1441.

Richter, M., Rosselló-Mora, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc. Natl. Acad. Sci. U. S. A. 106, 19126–19131.

Rogall, T., Wolters, J., Flohr, T., Böttger, E.C., 1990. Towards a phylogeny and definition of species at the molecular level within the genus *Mycobacterium*. Int. J. Syst. Bacteriol. 40, 323–330.

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

Seemann, T., 2014. PROKKA: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.

Sneath, P.H.A., Sokal, R.R., 1973. Numerical Taxonomy: the Principles and Practice of Numerical Classification. W.H. Freeman & Co., San Francisco, Ca.

Sokal, R., Michener, C., 1958. A statistical method for evaluating systematic relationships. Univers. Kansas. Sci. Bull. 38, 1409–1438.

Stackebrandt, E., Ebers, J., 2006. Taxonomic parameters revised: tarnished gold standards. Microbiol. Today 33, 152–155.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C., Whitman, W.B., 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 52, 1043–1047.

Stahl, D.A., Urbance, J.W., 1990. The division between fast- and slow-growing species corresponds to natural relationship among the mycobacteria. J. Bacteriol. 172, 116–124.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 30, 2725–2729.

Tortoli, E., 2012. Phylogeny of the genus *Mycobacterium*: many doubts, few certainties. Infect. Genet. Evol. 12, 827–831.

Tortoli, E., Kohl, T.A., Brown-Elliott, B.A., Trovato, A., Leão, S.C., Garcia, M.J., Vasireddy, S., Turenne, C.Y., Griffith, D.E., Philley, J.V., Baldan, R., Campana, S., Cariani, L., Colombo, C., Taccetti, G., Teri, A., Niemann, S., Wallace Jr., R.J., Cirillo, D.M., 2016. Emended description of *Mycobacterium abscessus*, *Mycobacterium abscessus* subsp. *abscessus* and *Mycobacterium abscessus* subsp. *bolletii* and designation of *Mycobacterium abscessus* subsp. *massiliense* comb. nov. Int. J. Syst. Evol. Microbiol. 66, 4471–4479.

Wayne, L.G., Kubica, G.P., 1986. Family *Mycobacteriaceae* CHESTER 1897, 63^{AL}. In: Sneath, P.H.A., Mair, N.S., Sharpe, M.E., Holt, J.G. (Eds.), Bergey's Manual of Systematic Bacteriology. The Williams & Wilkins Co., Baltimore, pp. 1435–1457.

Wayne, L.G., Brenner, D.J., Colwell, R.R., Grimont, P.A.D., Kandler, O., Krichevsky, M.I., Moore, L.H., Moore, W.E.C., Murray, R.G.E., Stackebrandt, E., Starr, M.P., Trüper, H.G., 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int. J. Syst. Bacteriol. 37, 463–464.