Check for updates

RESEARCH ARTICLE

REVISED Typing and classification of non-tuberculous mycobacteria isolates [version 2; peer review: 1 approved, 1 approved with reservations]

Thomas Clarke ¹, Lauren Brinkac ^{1,2}, Joanna Manoranjan³, Alberto García-Basteiro^{4,5}, Harleen Grewal⁶, Anthony Kiyimba⁷, Elisa Lopez^{4,5}, Ragini Macaden^{6,8}, Durval Respeito^{4,5}, Willy Ssengooba ^{7,9}, Michele Tameris ¹⁰, Granger Sutton ¹

¹J Craig Venter Institute, Rockville, MD, 20850, USA

²Noblis, Reston, VA, 20191, USA

³Aeras, Rockville, MD, USA

⁴Centro de Investigação em Saúde de Manhiça, Maputo, Mozambique

⁵ISGlobal, University of Barcelona, Barcelona, Catalonia, Spain

⁶Infectious Diseases Unit, St. John's Research Institute, Bangalore, Karnataka, India

⁷College of Health Sciences, Makerere University, Kampala, Uganda

⁸St. John's Research Institute, Bangalore, India

⁹Makerere University, Kampala, Uganda

¹⁰Department of Pathology, University of Cape Town, Cape Town, South Africa

v2 First published: 09 Apr 2020, 9:249
<https://doi.org/10.12688/f1000research.22620.1>

Latest published: 28 May 2020, 9:249
<https://doi.org/10.12688/f1000research.22620.2>

Abstract

Background: There are a large and growing number of non-tuberculous mycobacteria (NTM) species that have been isolated, identified, and described in the literature, yet there are many clinical isolates which are not assignable to known species even when the genome has been sequenced. Additionally, a recent manuscript has proposed the reclassification of the *Mycobacterium* genus into five distinct genera.

Methods: We describe using a community standard fast average nucleotide identity (ANI) approximation method, MASH, for classifying NTM genomes by comparison to a resource of type strain genomes and proxy genomes. We evaluate the genus reclassification proposal in light of our ANI, MLST, and pan-genome work.

Results: We describe here a sequencing study of hundreds of clinical NTM isolates. To aid in characterizing these isolates we defined a multi-locus sequence typing (MLST) schema for NTMs which can differentiate strains at the species and subspecies level using eight ribosomal protein genes. We determined and deposited the allele profiles for 2,802 NTM and *Mycobacterium tuberculosis* complex strains in PubMLST.

Open Peer Review

Approval Status

	1	2
version 2		
(revision)		
28 May 2020	view	view
version 1		
09 Apr 2020	view	

1. **Charles Warden** , City of Hope National Medical Center, Duarte, USA

2. **Radhey S. Gupta** , McMaster University, Hamilton, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Conclusions: The MLST schema and our pan-genome analysis of Mycobacteria can help inform the design of marker-gene diagnostics. The ANI comparisons likewise can assist in the classification of unknown genomes, even from previously unknown species.

Keywords

non-tuberculous mycobacteria, NTM, taxonomic classification, MASH ANI, MLST

Corresponding author: Granger Sutton (GSutton@jcv.org)

Author roles: **Clarke T:** Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Brinkac L:** Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Manoranjan J:** Investigation, Resources; **García-Basteiro A:** Investigation, Resources, Writing – Review & Editing; **Grewal H:** Investigation, Resources, Writing – Review & Editing; **Kiyimba A:** Investigation, Resources, Writing – Review & Editing; **Lopez E:** Investigation, Resources, Writing – Review & Editing; **Macaden R:** Investigation, Resources, Writing – Review & Editing; **Respeito D:** Investigation, Resources, Writing – Review & Editing; **Ssengooba W:** Investigation, Resources, Writing – Review & Editing; **Tameris M:** Investigation, Resources, Writing – Review & Editing; **Sutton G:** Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work has been funded in whole or in part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under award number U19AI110819 and by the Research Council of Norway [grants 179342 and 192534]; the University of Bergen (Norway); Aeras (USA) and St. John's Research Institute (India).

Copyright: © 2020 Clarke T *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Clarke T, Brinkac L, Manoranjan J *et al.* **Typing and classification of non-tuberculous mycobacteria isolates [version 2; peer review: 1 approved, 1 approved with reservations]** F1000Research 2020, 9:249 <https://doi.org/10.12688/f1000research.22620.2>

First published: 09 Apr 2020, 9:249 <https://doi.org/10.12688/f1000research.22620.1>

REVISED Amendments from Version 1

The changes in the new version were made in response to the first reviewer. For a detailed description of the suggestions and comments and our responses, I will refer to the response to the reviewer we posted in responses, including descriptions for the changes in the manuscript, which are listed below:

Page 4: Abstract: Added “community standard” for MASH to clarify that the program does not originate with us, but instead is a widely used rapid genomic ANI calculator

Introduction: Made suggested syntax changes: italicized *Mycobacterium* as it is a genus name, removed a period, and added a comma.

Page 5: Results: Added first, second, third and finally, plus semi-colons to help separate the different facets and make the sentence more readable. Also added that the ANI mentioned in the second facet is related to the clustering in the first.

Page 7: Added clarification that disagreements in the NCBI names with our assignments are possibly due to divergent naming regimes and do not necessarily indicate different classifications.

Page 8: Added clarification that our genus assignments still do need organismal trait observations for full confirmation per a suggestion by the reviewer

Figure 1 Legend. The added sentence is to add similar clarification as in page 7.

Page 11: Added a period as suggested by the reviewer.

Page 12: We have added direct comparisons of our ANI and MLST results with Matsumoto *et al.* (reference #44), specifically in Figure 1C, that were previously only in Supplementary Table S13. We have highlighted two divergent claims of the placement of individual genomes (*M. avium*) and of genus clades. We additionally have added a clause where the Matsumoto placement of a species confirms our placement.

Any further responses from the reviewers can be found at the end of the article

Introduction

Non-tuberculous mycobacteria (NTM) are a ubiquitous group of diverse environmental mycobacteria related to *Mycobacterium tuberculosis*, the bacterium that causes tuberculosis (TB)^{1–7}. A significant number of NTMs cause clinical disease. Diseases caused by NTM infections are important sources of morbidity and mortality throughout the world, with a resurgence of infections in the United States, especially in immunocompromised individuals. The host-pathogen interactions of NTMs remain poorly characterized. Importantly, the background prevalence of NTMs has been shown to interfere with important aspects of TB disease control and management, including the efficacy of new TB vaccines^{8,9}, established immuno-diagnostic methods, and the specificity of diagnostic and immune biomarkers¹⁰. Data on NTM prevalence in TB-endemic countries, however, is limited and often under-reported. Possible factors contributing to under-reporting include overlapping clinical manifestations of TB and NTM disease, lack of awareness of NTMs, and insufficient laboratory infrastructure to identify and culture NTMs¹¹. The NTM-NET published a survey of NTM patient isolates¹²: “A total of 20 182 patients had NTM species cultured from pulmonary samples in these centres in 2008; 91.3% (n=18 418) of the isolates were identified to species/complex level; the remaining 1764 isolates (8.7%)

were not identified beyond *Mycobacterium* species other than *M. tuberculosis* complex. A total of 91 different NTM species were encountered in this survey.” While this was an important survey of NTM, geographical distribution the low number of NTM species identified and the inability to classify some samples emphasizes the need for more classification resources. Gupta *et al.* proposed the *Mycobacterium* genus be divided into five distinct genera: *Mycobacterium*, *Mycolicibacter* gen., *Mycolicibacterium*, *Mycolicibacillus*, and *Mycobacteroides*¹³ in the *Mycobacteriaceae* family. There are over 197 *Mycobacteriaceae* species from these five genera according to the [List of Prokaryotic Names with Standing in Nomenclature \(LPSN\)](#)^{14,15} and elsewhere.

We undertook to sequence, analyze and characterize NTM strains representing geographically diverse TB-endemic areas with high NTM prevalence. Ancillary findings from adolescent and neonatal cohort studies conducted by Aeras and their collaborators to determine the incidence and prevalence of TB in Cambodia, India, Mozambique, South Africa, and Uganda included the prevalence and typing of NTM infections. The sites that participated in these studies collected NTM samples which were sequenced by JCVI. *Extended data, Supplemental table S1*¹⁶ lists the samples sequenced from each site.

Our experience elucidated some of the issues which confound clinicians. Current commercial methods for typing NTMs are limited to a small number of the better known NTMs and even then are often not sensitive or specific enough^{2,17}. Connecting the published literature for NTM type strains and related characterization of the species to a clinical isolate is often difficult even given the genome of the isolate due to non-specific marker gene sequences being available. This is being addressed to some extent as many more type strains are having their genomes sequenced^{18–20}. We propose a fast average nucleotide identity (ANI) method when whole genome sequencing is practical and a multi locus sequence typing (MLST) approach when whole genome sequencing is not feasible.

Results

There are four major facets of our analysis: first, clustering of *Mycobacteriaceae* family genomes using ANI; second, a pan-genome of representative genomes from each clade determined by ANI mentioned above; third, an evaluation of our MLST scheme; and finally, a comparison of our pan-genome to the work of Gupta in terms of genera assignment and genera specific genes. ANI-based clustering has been shown to be well correlated with known species and subspecies definitions^{21–30}. We show how automated ANI clustering and ANI comparisons to representative genomes can inform classification of a novel genome. Pan-genome analysis can be used to evaluate genome relatedness across two major comparators: ANI as previously mentioned across all genes but also across subsets of genes, and gene content similarity. We discuss aspects of both pan-genome ANI and gene content. MLST can be used in the absence of genome sequencing data if the loci used in the MLST scheme are acquired by other methods such as PCR. We

propose a distance metric based on the concatenated alignment of the MLST loci and evaluate how this can be used in the absence of ANI for classification. Gupta recently proposed a reclassification of five genera within the Mycobacteriaceae family¹³. We evaluate this proposal compared to our pan-genome analysis. Gupta also proposed genera specific genes to help delineate the genera and we evaluate those as well.

ANI-based clade generation and representative genome

A total of 2,802 genome assemblies in the Mycobacteriaceae family in NCBI RefSeq were downloaded on 8 August 2019 (to reduce the number of genomes analyzed we only used complete *M. tuberculosis* genome assemblies). This represents the complete set as at 30 April 2019. We added six Mycobacteriaceae genome assemblies available in GenBank that had been excluded from RefSeq either for no specified reason or for being untrustworthy type strains, though the latter were not used to assign names to other genomes. An additional 11 genomes from other Corynebacteriales families that were sequenced as part of the project (*Extended data*, Supplemental table S1¹⁶) were included in the analysis for a total of 2819 genomes. We generated species- and sub-species-level clusters of the Mycobacteriaceae genomes using an all versus all ANI comparison with MASH^{21,31}. We generated species and sub-species clusters with average linkage and a cutoff value of 95% ANI for species and 98% for subspecies (*Extended data*, Supplemental tables S2 and S3³²) using GGRaSP³³. This has been shown to be a reasonable way to computationally compute tentative species/subspecies clusters^{18,21–30}. For the species-level 95% clustering, there were 269 clusters of which 263 are from *Mycobacteriaceae*, 126 of the *Mycobacteriaceae* clusters contained a type strain, of which 13 contained more than one type strain of different species (*Extended data*, Supplemental tables S2 and S3³⁴). For the subspecies-level 98% clustering, there were 336 clusters of which 329 are from *Mycobacteriaceae*, 137 of the *Mycobacteriaceae* clusters contained a type strain, of which 11 contained more than one type strain, and 1 which contained multiple subspecies (*Extended data*, Supplemental table S3³²). To represent as much of the taxonomic space as succinctly as possible, we selected representative genomes using clades generated from the two cluster levels. By default, we used the 95% ANI clusters, but for the nine 95% clusters that contained two or more 98% clusters containing type strains, we used all of the 98% clusters in place of the 95% clusters. This meant that 26 98% clusters replaced these nine 95% clusters, resulting in 280 *Mycobacteriaceae* clusters out of 286 total clusters, 131 of which contained a type strain, 12 of which contained more than one type strain of different subspecies (*Extended data*, Supplemental table S4³⁵). For each of the final 280 *Mycobacteriaceae* clusters, a representative genome was determined using GGRaSP based on the minimum of summed pairwise MASH ANI distances within the cluster which is called the cluster medoid genome. For the clusters with a RefSeq-included type strain, only the RefSeq-included type strains were considered for the representative genomes, whereas in the other clades, all genomes were considered.

Pan-genome

We used the J. Craig Venter Institute (JCVI) pan-genome pipeline to generate a pan-genome using 279 of the 280 representative

Mycobacteriaceae genomes plus a *Norcadia* genome (*N. nova* SAMN02900753, cluster 190) as an outgroup^{36,37}. We were required to exclude one of the clusters (180B) as the sole cluster member (SAMN04216932) is not available in RefSeq, whose consistent annotation we use for the pan-genome pipeline. Using gene presence in at least 95% of genomes to determine core genes, we found 1,680 core genes (723 at 100% including the outgroup genome another 407 missing in one genome only) out of 118,867 orthologous gene clusters including 71,559 gene clusters present in only a single genome. A previously reported NTM pan-genome¹⁹ using only 99 genomes struggled to identify many core genes (243), possibly due in part to using a 100% threshold for genome presence and generated a lot of orthologous gene clusters (150,000), possibly due to the 80% nucleotide identity cutoff used. The pangenome results in multiple similarity measures which aid classification of the genomes. For example, identifying core genes is crucial for developing diagnostics, especially for identifying potential MLST loci with the least diversity to allow for primer design. Additionally, non-core genes which are restricted to only a segment of the phylogeny can be used for classification purposes, such as genus-specific genes for genus classification. Finally, the JCVI pan-genome pipeline produces pairwise genome ANI results based on the orthologous clusters determined from which genome ANI trees can be created.

Genus and species classification of NTM genomes using MASH and pan-genome ANI

We believe the ANI based clusters (*Extended data*, Supplemental table S4³⁵) and their representative genomes can be used for novel genome classification. How to classify clusters which contain more than one differently named type strain or contain no type strain at all is complicated. As Tortoli *et al.*¹⁸ point out, historical species definitions made before genome sequencing may need to be revised when existing species are found to have very similar genomes. Species should not be redefined simply based on ANI clustering since other factors such as significant difference in genome size (gene content), significantly divergent host specificity, disease presentation, or significant presence in the literature must be taken into account. We leave it to those more familiar with Mycobacteriaceae species to make those determinations. Three other problems were pointed out by Tortoli *et al.*¹⁸ which we also observed: some type strains have been genome sequenced more than once where the genomes are clearly different species from each other³⁸; type strain genomes do not conform to species descriptions and previously sequenced marker gene sequences, and non-type strain genomes have been misnamed based on the first two problems. We identified two type strains: *M. phlei* (SAMEA4040484) and *M. chelonae* (SAMEA4040023) whose clustering with other type names suggests that their names have been swapped on submission to NCBI. We have used our corrected names for these samples in our tables and figures. As mentioned above, for clusters with more than one differently named type strain, we chose the medoid of the type strains as the cluster representative but in *Extended data*, Supplemental table S4³⁵ we give all of the type strains in the cluster and in *Extended data*, Supplemental table S5³⁹ we use an asterisk with the proposed species/subspecies classification name to indicate the uncertainty for members of these clusters. For species clusters

lacking a type strain, a cluster-specific “genomosp.” (this is NCBI’s preferred designation for proposed species clades without a type strain) such as genomsp. X was used as the species name for all the genomes in the cluster. If a subspecies cluster did not have a type strain, then only the species level name is given since NCBI does not support a genomsp. concept at the subspecies level.

For all the Mycobacteriaceae genomes, we assigned a genus to them using average-linkage clustering based on PanOCT

pan-genome pairwise ANI distances (Figure 1)³⁶. For the four new genera which Gupta proposed, we found the least common ancestor node which contained all the Gupta genus type strains for that genus. Since Gupta supplied only one type strain for the *Mycobacterium* genus, we also found the root node of the largest subtree that included this type strain and no additional Gupta genus type strains. The genomes descending from these five nodes were then assigned to the respective Gupta genus. Other than the single outgroup genome not from Mycobacteriaceae, there is only one clade which is not included

Figure 1. Phylogenetic tree of Mycobacteriaceae representative genomes. The tree was generated from the pan-genome ANI distance metric using average-linkage clustering. The branches are colored according to the assigned genera of the genomes from least-common-ancestors with the Gupta-defined type strains (shown on the tree as the colored circle on the ancestral nodes). The leaves are colored by the genus given by NCBI for the genome on 09/09/2019. Some of the NCBI genus names listed do not account for the Gupta *et al.* genus definitions, and thus branch and leaf color disagreements probably only reflect a difference in genus naming conventions. On the outside of the tree, the orange circles denote NCBI-defined type strains, whereas farther out circles indicate the type strains identified in Gupta *et al.* as belonging to a specified genus. The green star identifies the cluster containing two type strains for which the NCBI type strain genus identification and the Gupta assignment differs, and the orange star identifies the cluster containing a type strain previously suggested to be mis-identified.

in the five Gupta subtrees. In [Figure 1](#), this clade is indicated by black colored branches. By Gupta's criteria for NTM genera within the Mycobacteriaceae this novel clade should be designated as a new genus. This clade contains four representative genomes but only one type strain, *Mycobacterium haemophilum* ([SAMN01918491](#)). We labeled these genomes as Unclassified genera in *Extended data*, Supplemental table S4³⁵, pending assignment of an additional genus name. NCBI has accepted the Gupta proposal in naming NTM species except for the misidentified genome, *Mycolicibacterium vulneris* ([SAMN06651660](#))^{38,40}. For this genome, NCBI has accepted the genus reassignment from the Gupta corrigendum, *Mycobacterium vulneris* which matches our clade assignment, though they have not renamed the species. Neither Gupta nor NCBI has renamed all NTM genomes in GenBank based on some form of clustering or phylogeny. This means the current NCBI names are not consistent with our clade-based assignment, as seen in the occasional differences in leaf and branch colors in [Figure 1](#) and that these differences do not necessarily reveal divergent classification of the species using Gupta based genera. Of the representative type strains, four have NCBI genus names which disagree with our assignment. These are labeled as *Mycobacterium* (*neumannii*, *syngnathidarumi*, *dioxanotrophicus* and *aquaticum*), but are located in the *Mycolicibacterium* subtree. One of them (*neumannii*) even shares a cluster with a Gupta *Mycolicibacterium* type strain (*Mycolicibacterium flavescens*). There are also 58 non-type representative genomes whose NCBI genus assignment disagrees with ours, and we have highlighted these rows in *Extended data*, Supplemental table S4³⁵. Our pan-genome ANI clustering was done for the representative genomes, so we conferred the same genus as determined for each representative genome to genomes in that representative's species/subspecies cluster (*Extended data*, Supplemental table S5³⁹). This also identified four non-representative type strains which had a different genus than their NCBI name. All four were previously labeled as *Mycobacterium*, with two renamed as *Mycobacteriodes* (*chelonae* subsp. *gwanake* and subsp. *chelonae*), and two as *Mycolicibacterium* (*kyoganese* and *lehmannii*). However, these genus reassignments based on ANI should be confirmed with supporting growth and metabolic observations similar to the Gupta *et al.* results.

The species and sub-species classification of all the non-representative genomes was done based on cluster membership. For genomes within clusters containing a single named type strain, that species name was used for all genomes. For clusters lacking a type strain, a cluster-specific "genomosp." was used as the species name for all the genomes in the clade. For clades with multiple named type strains, the species name of the best ANI match to one of those type strains was used. We warn that the above caution about assigning species based solely on ANI holds and that further work might be necessary. The classification results are shown in *Extended data*, Supplemental table S5³⁹.

We propose classifying a novel genome by running MASH ANI of the novel genome compared to the set of representative genomes. To verify this approach, we classified each genome

using MASH pairwise ANI to the set of representative genomes compared against the cluster-based naming. For each genome the closest MASH ANI to the representative genomes were almost exclusively within the same cluster. Only for 26 genomes in clusters 28B as well as single genomes in cluster 47B and in cluster 165B was the closest ANI assigned match not in the same cluster. This only occurred for clusters discriminated at the 98% threshold (subspecies). For the 26 genomes in cluster 28B but with a best ANI match to the cluster 28A's representative genome, clusters 28B and 28A are two of the four subspecies clusters of *M. abscessus*. There are only three named *M. abscessus* subspecies and the fourth cluster (28A) is probably artifactual occurring right at the boundary of the 98% MASH ANI threshold. Clusters 28A and 28B should probably be combined based on the simple observation that all genomes in both clusters had their second-best MASH ANI match to the other's representative sequence and all these ANI distances were less than 2% which is within the clustering threshold. The same is basically true for clusters 47B/E and 165B/A where only a few second-best MASH ANI distance values are slightly above 2%. Regardless of whether these clusters are merged based on these criteria when a novel genome is being evaluated the second-best MASH ANI distance should also be considered to ascertain the certainty of an assignment. Any novel genome exceeding the 95/98% ANI threshold to this set of genomes can be assigned a species/subspecies name using MASH ANI to representative genomes (again with more careful scrutiny if the second-best MASH ANI distance is very close to the best MASH ANI distance). The set of MASH sketches will need to be updated both as more genomes are sequenced falling outside of the existing clusters and as real species/subspecies names are defined or determined. While recent NTM type strain sequencing projects (BioProjects: [PRJNA354248](#)¹⁸, [PRJDB4227](#)²⁰, [PRJNA299467](#)¹⁹) have greatly increased the number of sequenced NTM type genomes, many have still not been sequenced and could correspond to some of the proposed genomospecies. Genomes sequenced as part of our project (*Extended data*, Supplemental table S1¹⁶) have assisted in the expansion, contributing genomes to 88 of the 280 *Mycobacteriaceae* clades. This includes 58 clades for which they are the only representatives and a third of both the *Mycobacterium* and the *Mycolicibacillus* clades.

Genus and species classification using MLST and conserved signature proteins

One common diagnostic tool used for identifying species or strains is multi-locus sequence typing (MLST), where a small set of single copy conserved gene sequences are used to distinguish genotype. MLST can be done computationally using genome sequences or via primer-based sequencing. In either case it is important that the gene sequences be conserved enough to design primers or find the gene via homology search; and divergent enough to be able to distinguish different species or strains. We found that some marker genes that had been used in the past⁴¹ tended to be either fragmented in genome assemblies or too divergent to align easily. When working with an earlier pan-genome we chose the core genes with no variation in gene length and highest conservation to serve as marker genes which

all turned out to be ribosomal proteins (S7, S8, S12, S14Z, S19, L16, L19, L35). This MLST schema was submitted to [PubMLST Mycobacteria](#)⁴². These eight marker genes are also core genes in the 280 Mycobacteriaceae genome pan-genome with very low gene length variation and high conservation (*Extended data*, Supplemental file S6³⁴).

When genome sequence is not available for a clinical sample but MLST is a possibility, we believe that our or a comparable MLST schema could be used for diagnostic species/subspecies assignment, such as seen in *Extended data*, Supplemental table S7⁴³. Using our MLST schema an ST is never shared across species/subspecies clusters (*Extended data*, Supplemental table S5³⁹). For most of the genomes, the ST distance (the fraction of mismatches across the combined multiple sequence alignment of the MLST loci) of a genome to the representative genome of that genomes ANI cluster is smaller than the ST distance to any other representative genome (*Extended data*, Supplemental table S5³⁹). Only for 142 genomes was the closest ST different than the ANI assigned cluster (*Extended data*, Supplemental table S5³⁹). These genomes include 27 out of the 28 genomes whose best MASH hit to a representative were not in the same cluster as themselves, excluding only [SAMN07811434](#) from Cluster 165B. This difference mostly occurred for an ST when the incorrectly chosen ST was from a different subspecies cluster of the same species, though for 59 genomes they were from different species clusters. For the 59 genomes which had best ST hits to different species the ST distance was above

the cutoff for same species assignment explained below. For almost all 142 discrepant genomes the second closest ST was the correct one, with the exception of four cluster 47E genomes where the best ST hit was to 47D, and the second-best ST hit was to 47B. A remaining issue is whether a novel genome is within an existing cluster or a novel cluster. [Figure 2](#) shows statistics for ST distances versus species and subspecies ANI ranges. We can see from this data that ST distance for 95–98% ANI ranges from 0.004–0.02, 98–100% ANI from 0–0.01, and <95% ANI from 0.006–0.26. We devised cutoffs to separate out the three groups using two standard deviations from the mean ST distance for 95–98% ANI (0.007 ± 0.003) which separated the three ANI-based groups (Different Species > 0.01 ≥ Same Species ≥ 0.004 > Same Subspecies) with minimal misassignment ([Figure 2](#)). Diagnosticians should use this data to inform the likelihood that a novel ST is from a known or unknown species/subspecies cluster.

A recent article published after our analysis had been completed describes what is called an MLST approach to determine NTM species/subspecies classification⁴⁴. Matsumoto *et al.* do compare their method against several others including our pubMLST scheme. They do not use our LOCUST software and it is not clear exactly which software on the pubMLST website they do use to evaluate our MLST scheme. For their 29 test genomes, they find that our MLST scheme correctly assigns all genomes at the species level but does not give a subspecies classification as theirs does. This makes sense because the

Figure 2. ST pairwise distance distribution. The distribution of ST distances between pairs of *Mycobacteriaceae* genomes classified by ANI clustering. Three categories of pairwise relationships were examined: genomes of different species (MASH ANI < 95%), genomes of the same species (95% < MASH ANI < 98%), and genomes of the same sub-species (MASH ANI > 98%). ST distance cutoffs to divide the groups were calculated from the mean distances of genomes from the same species plus or minus two standard deviation (solid horizontal lines). These successfully classified nearly all of the data. The percentage of each category of ANI matches which fall into the ST distance bins given by the two thresholds (0.01 and 0.004) are shown directly on the figure.

profile we uploaded to pubMLST at the time of their analysis did not include subspecies level information. We have recently updated our pubMLST profile to include subspecies information. The MLST scheme presented by Matsumoto *et al.* is atypical. First, the set of loci used is very large (184), which is not consistent with performing efficient PCR. Second, only 53 of the loci are core genes; whereas, the remaining 131 are species specific genes. Third, because most of the loci are species specific it is not even clear how one defines a sequence type and what the authors discuss instead is a similarity score. Fourth, the msltverse software which is used (and is not compatible with standard MLST schemes) needs whole-genome read data as input obviating any advantage of MLST over whole-genome sequencing. As we have argued, whole-genome data for ANI comparison is preferable to an MLST approach but we also argue that the ANI data is sufficient for species categorization and that msltverse is not necessary. We have nothing against looking for the presence or absence of genes in whole-genome sequence data to help identify virulence or antibiotic resistance. We do caution that only WGS data which can be assembled into complete chromosomes and plasmids can reliably determine the absence of a gene. We also note that claiming a particular gene is species/subspecies specific is only possible if a large, diverse set of genomes is available for all species/subspecies. Finally, attempting to do species/subspecies classification based on gene content is problematic due to horizontal transfer, particularly across subspecies boundaries.

To further aid in the genus classification of novel *Mycobacteriaceae* genomes, it is also possible to look for the presence of conserved, genus-specific genes. Gupta *et al.* provide 89 conserved signature proteins which can be used to define the family and genera (Extended data, Supplemental table S8⁴⁵)¹³. These proteins were selected by Gupta *et al.* based on BlastP comparisons of the genera type strain genomes and either the non-*Mycobacteriaceae* protein databases or *Mycobacteriaceae* genomes outside of a given genus Using our *Mycobacteriaceae* pan-genome with the genus assignments discussed above, we identified genus specific gene clusters which were present in each representative genome in a genus but not in any of the other genomes. If no cluster met this criterion for a genus, we required the difference between a gene clusters presence in the genus genomes and the out of genus genomes to be greater than 90%. We used the medoids of these selected gene clusters as potential CSPs. This resulted in 648 pan-genome based CSPs (Extended data, Supplemental table S9⁴⁵). Only a small number²⁵ of the Gupta CSPs were mapped to the pan-genome's CSP clusters using either GenBank identifier matches or best blast matches probably due to different criteria for selection. Most⁴⁶ of the remaining 65 of the Gupta *et al.* CSPs were genus-specific, but only had a partial presence in the genomes associated with the genus.

To test the genus prediction success rate of BlastX searches of these gene sets, we compared them to the 2,802 RefSeq *Mycobacteriaceae* genomic sequences. We did not include 11 of the Gupta CSPs which have since been removed from RefSeq (Extended data, Supplemental table S8⁴⁵). Classification of the genomes with the CSPs was based on a simplified

assignment algorithm, where a genome was assigned to the genus with the greatest percentage of genus-specific genes that had a blast hit to a given genome with an e-value < 1e-5. The Gupta CSPs gave the same answer as the pan-genome based genus assignment for all the RefSeq *Mycobacteriaceae* genomes excluding the Unclassified clade 98.7% of the time, whereas the pan-genome CSPs exceeds this with a 99.9% match rate, even when accounting for the Unclassified genus. The one incorrect assignment was an Unclassified genome being assigned as *Mycobacteroides*. The assignment could potentially be improved by expanding the search beyond only a single medoid gene for each gene cluster. Overall using the CSPs to assign genus seems to have limited utility compared to MASH ANI or MLST. Researching why certain genes are genus specific and what phenotype(s) they confer might be of more value.

Performance on Novel Genomes

After our analysis was complete, some novel high-quality NTM genomes became available from Matsumoto *et al.*⁴⁴: "In this work, we expanded the NTM genomic data by 63 species, of which 27 species were newly determined and the remaining 36 species were re-sequenced by Illumina and Oxford Nanopore Technologies (ONT)." We compared these 63 genomes along with the genomic classifications in NCBI and the species tree in Figure 1C in Matsumoto *et al.*⁴⁴ to our cluster representatives using ANI and ST distance (Extended data, Supplemental table S13⁴⁷). The classification of the genomes using these strategies and the msltverse reported in Matsumoto are highly similar, though there were differences. Two dissimilarities include the imposition of the *Mycolicibacillus* and *Mycobacterioides* genome clades into the larger *Mycolicibacterium* clade in the Matsumoto *et al.* Figure 1C⁴⁴ as compared to the fully distinct clades in the pangenome tree (Figure 1), and the different placements of *M. avium* in the two trees, with the pangenome placement more reflective of earlier suggestions and the genera name. We also used our and Gupta's CSP proteins to predict genera (Extended data, Supplemental table S13⁴⁷). Of the genomes, 16 are not in any of the clusters by ANI (95% threshold). Of these genomes, 15 currently have genus names that agree with their closest MASH medoid (four *Mycobacterium* and eleven *Mycolicibacterium*). The single genus name disagreement is BioSample SAMD00153170, which is currently labeled *Mycobacterium gallinarum*, but which is closest to a *Mycolicibacterium* clade with a MASH distance well within the generally accepted ANI range (94.19%) of being in the same genus and is further located in a *Mycolicibacterium* clade in the tree shown Matsumoto *et al.* Figure 1C⁴⁴. Additionally, seven genomes are in previously untyped clusters, thus enabling the clades to be named with a species name (Extended data, Supplemental table S4³⁶). For 47 of the 63 genomes the best ANI match to a cluster representative is greater than 95%. Of these 47, 33 have the same species name as the representative, 7 as noted previously do not have a real species name, and 7 have different species names. *Mycobacteroides abscessus* subsp. *massiliense* has its best ANI match to cluster 28A which is one of the four subspecies clusters for *Mycobacteroides abscessus* but the previously sequenced type strain for *Mycobacteroides abscessus* subsp. *massiliense* is in cluster 28B. We had already observed that clusters 28A and 28B

should probably be merged based on ANI and ST mapping and this only reinforces that observation. For the 7 with different species names, there are four categories: split the existing cluster into two or more clusters at 98% ANI to represent different species (three genomes with 95.1–96% ANI to cluster representative), split the existing cluster into two or more clusters at 98% ANI to represent different subspecies (two genomes with 97.6–98.3% ANI to cluster representative), consider declaring two type species to be the same species and one of them to be a heterotypic synonym (1 genome with 99.0% ANI to cluster representative), and the genome type strain is already in the cluster but the chosen medoid between the two species is not the same (paraintracellular and intracellular in cluster 47D). This leaves 16 genomes with a best match to a cluster representative below 95% ANI. These 16 genomes would found new clusters where each would be a singleton cluster and be the representative for that cluster. Of the 63 genomes 25 had novel ST types which we added to the pubMLST schema. The ST distances were generally well correlated with the ANI values with only 4 genomes with ANI below 96.1% having ST distances above the same species threshold. For the CSPs neither Gupta (five discrepant) nor ours (one discrepant) performed perfectly for genus assignment. *Mycolicibacterium aubagnense* was classified by both methods as *Mycobacterium*, and four additional genomes had discrepant Gupta CSP based genus assignments. This confirms our skepticism about this approach.

Discussion

Our analysis shows that the universe of NTM species/subspecies is still far from complete: 142 of our clusters do not even have an associated type strain, many NTM species still do not have a genome sequence, and there are probably many NTM clades which are not represented in our clusters. Our suspicion of the sparseness of the NTM family tree is reinforced with the addition of the Matsumoto genomes, a quarter of which were not assigned to our species-level clusters. The incompleteness of NTM genomic data could easily lead to misdiagnosed or undiagnosed NTM infections. We have proposed an efficient structure based on ANI and species/subspecies cluster representative genomes to classify any patient sample with genomic data. This includes recognizing that a patient sample is from a novel NTM clade. As more NTM genomic data becomes available, our representation is easily extended with new species/subspecies cluster representative genomes. Many NTM species are slow or very slow growing making isolation of DNA for genomic sequencing challenging. When genomic data is not available, we show that our pubMLST scheme is sufficient for the classification of most patient samples. Diagnostic companies may not prefer our choice of MLST loci, so we have included all of our data on the pan-genome clusters (*Extended data*, Supplemental file S6³⁴) from which to choose for PCR or hybridization-based methods. We chose our loci based on being more highly conserved at the nucleotide level for PCR primer design, being present in all genomes, being full length in all genomes, and having low length variability so that gaps would not complicate multiple sequence alignment-based analysis (*Extended data*, Supplemental file S6³⁴). When more extensive analysis is desired, for example for the presence of noncore

antibiotic or virulence genes a pan-genome analysis can be helpful. We showed that the pan-genome can be used to determine core and genus-specific genes.

Methods

Sample collection and preparation

A total of 288 samples of genomic DNA were collected from five different geographical sites (*Extended data*, Supplemental Table S1¹⁶). The genomic DNA was isolated from cultures grown from samples of sputum or gastric aspirates that were originally collected from TB presumptive patients with the exception of an adolescent cohort study (E-004-IN-A) from the St. John's Research Institute (SJRI), where the specimens were collected from healthy children. Samples were initially treated with a decontamination and digestion step using N-acetyl-L-cysteine-sodium hydroxide (NALC-NaOH) and then screened for acid-fast bacilli (AFB) using smear microscopy with Ziehl-Neelsen (ZN) or Auramine O fluorescent staining (South African Tuberculosis Vaccine Initiative (SATVI) and Makerere University (MU). Samples obtained from SJRI, Centro de investigação de Saúde de Manhica (CISM), and MU were also examined directly by ZN and fluorescent microscopy prior to decontamination. After decontamination and digestion, samples were cultured on Lowenstein Jensen (LJ) media and also inoculated into Mycobacterial growth indicator tubes (MGIT) 960 BD, following manufacturer's instructions. Samples from SATVI that were MGIT positive for growth were then stained for AFB by ZN and Gram stain. Samples from all sites that were AFB-positive were further identified at the genus and species levels. Samples at Institut Pasteur – Cambodia (IPC) and SATVI were initially analyzed by AccuProbe, Gen-probe, Inc., while samples at MU were first typed by mycobacteria specific 16S rDNA PCR identification. All sites, with the exception of SATVI, used the Hain GenoType Mycobacterium CM/AM test system, according to manufacturer's instructions, to identify NTMs at the species level. Samples at SATVI that were Accuprobe negative were further identified by methods of Brunello *et al.*⁴⁸ and Telenti *et al.*⁴⁹ (Briefly, PCR amplification of hsp65 gene, digestion of product by 2–3 restriction enzymes, identification by RFLP – examination of banding patterns on agarose gel electrophoresis); see Hatherill *et al.*⁵⁰ for further details.) If any other bacteria were detected during smear microscopy examination for AFB, the cultures were subject to a second round of NALC-NaOH decontamination and reinoculated into LJ and MGIT. The contaminated cultures were also plated onto blood agar to identify contaminants. All decontaminated specimens positive for AFB were saved and stored at either -20°C (CISM) or -80°C (IPC, SJRI, MU, SATVI). Positive growth cultures from the MGIT were also stored at -80°C in the MGIT media of Middlebrook 7H9.

Sample regrowth

In an attempt to provide broad coverage of the NTM strains both by species and geographic location, no more than 10 strains of the most frequent species from each site were chosen for regrowth. At the same time, no fewer than five strains from the same species were collected when possible. Additionally, a number of unidentifiable NTM cultures were also collected

from each of the sites. Samples within these parameters were chosen at random for regrowth; however, samples that failed to regrow were replaced with new samples. Sites noted that samples frozen at more recent dates were easier to regrow. Regrowth of cultures was done from MGIT Middlebrook 7H9 or decontaminated specimen frozen stocks by inoculating MGIT 7H9 media (CISM, IPC, MU, SJRI), LJ medium (CISM, SATVI/CSU), and 7H11 (SATVI/CSU, CISM). Subculturing on LJ media for single colony isolation was done for samples from CISM, IPC, SATVI/CSU, and SJRI. Regrown samples from MU were observed by ZN staining to re-confirm presence of AFB and also rule out contaminants by culturing on blood agar.

DNA extraction

NTM cultures were grown in 7H9 medium up to an OD >0.6 and then centrifuged for five minutes at 10,000 rpm. Supernatant was removed and pellets resuspended in TE buffer before incubating at 95°C for 10 minutes. Several sites grew cultures on solid media instead of 7H9 prior to DNA extraction. This media included 7H11 agar (SATVI) and LJ agar (SJRI, CISM, CISM). For these samples, single colonies were used for DNA extraction in TE buffer. After the incubation step, samples were centrifuged for 10 minutes at 10,000 rpm and supernatant removed and passed through a sterile syringe Filter (0.2 µm). Samples were stored at 4°C.

Sequencing and assembly

Illumina NextSeq library preparation and sequencing was used. Genome assembly was performed using SPAdes v 3.1.1⁵¹, which produced good quality assemblies as detailed in supplementary materials (*Extended data*, Supplemental Table S1¹⁶).

Genome annotation

The NTM genomes sequenced for this project were annotated by NCBI's prokaryotic annotation pipeline, PGAP⁵².

ANI calculation and clustering

A total of 2,802 NTM and TB complex strains were downloaded from NCBI RefSeq on 08/08/2019 and their PGAAP annotation used for comparative analysis. 6 Mycobacteriaceae strains excluded from RefSeq for either being untrustworthy as type or for no given reason and 11 Corynebacteriaceae sequenced by as part of this project were also downloaded. ANI between genomes was calculated using MASH v2.0, a very fast tool for determining approximate pairwise ANI values based on min-Hashing of kmer overlap³¹. Pairwise comparisons of all the genomes was run with 10,000 21-mers per genome. We used the GGRaSP R package v1.0 which generated a complete linkage tree from the distance matrix calculated by subtracting 100 from the MASH ANI results³³. The resulting tree was divided into clusters at 95% (species level) and 98% (sub-species) ANI cutoffs using R commands `ggrasp.cluster(threshold = 2)` and `ggrasp.cluster(threshold = 5)`. For 95% ANI clusters which contains at least two daughter 98% clusters with species or sub-species type strains, all the 98% clusters are retained; otherwise, only the 95% clusters are retained. All clusters are named by the GGRaSP 95% numeric order, and the 98% cluster having

a sequential additional letter. For each cluster, a representative was selected by picking the type strain that has the lowest combined ANI to all the other genomes in the cluster. For clusters with no type strain, the genome with the lowest combined ANI to all the other genomes was selected. If a cluster contained a single type strain or a Matsumoto genome was later assigned to a cluster with no type strain, then genomes in that cluster were named based on that type strain. For clusters with multiple type strains genomes were also individually named from MASH ANI comparisons with type strains (both those downloaded on 8/8/2019 and those subsequently sequenced by Matsumoto) using the name deriving from the type strain with the highest ANI value. Species names are only used when the ANI value is greater than or equal to 95% and sub-species for values greater than or equal to 98%.

Pan-genome generation and analysis

We used the JCVI pan-genome pipeline (default settings using nucleotide sequences) to generate a pan-genome using 279 of the 280 representative *Mycobacteriaceae* genomes plus a *Norcadia* genome (*N. nova* SAMN02900753, cluster 190) as an outgroup^{36,37}. We were required to exclude one of the clusters (180B) as the sole cluster member (SAMN04216932) is not available in RefSeq whose consistent annotation we use for the pan-genome pipeline. When representative assemblies were not available in RefSeq (two clusters: 141, 188A), but assemblies from other genomes within the cluster were, a secondary representative present in RefSeq was used.

Genus assignment

The genus of the 279 *mycobacteriaceae* medoids was assigned using the pan-genome ANI average-linkage clustering generated from the JCVI pan-genome pipeline^{36,37}. For four of the genera, the Least-Common Ancestor for all genus type strains designated by Gupta *et al.* (found either within the text or in Tables 8–11) was identified¹³. Since the genus *Mycobacterium* only had a single designated genus type strain, the most ancestral node containing the genus type strain and no other type strains was used as the genus ancestral node. All genomes descending from these nodes were assigned to that genus. Any remaining *Mycobacteriaceae* genomes were in a single cluster and labeled as unclassified. All non-representative genomes were assigned the same genus as the representative strain in their cluster.

MLST

Core genes identified by pan-genome analysis^{35,37} were evaluated to identify candidate marker genes for MLST analysis. Eight highly conserved ribosomal proteins (S7, S8, S12, S14Z, S19, L16, L19, L35) were selected. Distinct allelic profiles and sequence type definitions were generated by searching the consensus sequences of the eight ribosomal proteins against 2,802 NTM and TB complex strains using the novel schema function of LOCUST v3.0⁵³. The corresponding allele sequences and full MLST schema was submitted to <https://pubmlst.org/mycobacteria/>. Species and subspecies assignments were made for 2,802 NTM and TB complex strains using the cluster approximation function of LOCUST version 3.0 (manuscript in preparation) based on the 280 type and proxytype strains

discussed previously (*Extended data*, Supplemental table S4³⁵) with the command line `typer.pl -biosample_list biosample.list -download_schema "Mycobacteria" -cluster_approximation -cluster_approximation_list type_strain.list -t fasttree`

The programs used in this study are available through GitHub at the JCVI repository at <https://github.com/orgs/JCVenterInstitute>, including: JCVI pan-genome pipeline (PangenomePipeline), LOCUST, and GGRaSP.

Genus-specific gene assignment

Genus Conserved and Unique Proteins were selected for each of the six Mycobacteriaceae genera by identifying clusters which are present in all the genomes of a genus and absent in all other genomes. When there were no such clusters for a genus, clusters where the percentage difference between presence in the genus and presence outside the genus was greater than 90% were selected. The medoids for each of the selected clusters were used. Protein sequences of the cluster centroids were obtained using transeq in the EMBOSS suite. Protein identifications for all of the Gupta *et al.* genes were collected from tables 2, 5, and 7¹³ and the sequences were downloaded from NCBI. Comparisons of the presence and absence of the CSPs in the genomes was done using tblastn for the whole genome sequences and the CSPs. Blast matches with e-values < 1e-5 were considered a hit.

Data availability

Underlying data

NCBI BioProject: Sequence and Analysis of Non-tuberculous Mycobacteria (NTM) strains. Accession number [PRJNA305922](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA305922).

This BioProject contains all clinical NTM sample sequences; Individual biosample IDs of the genomes can be found in *Extended data*, Supplemental table 1¹⁶.

MLST allele sequences and corresponding schema is available at the PubMLST *Mycobacteria* genome database (<https://pubmlst.org/mycobacteria/>). All MLST profiles generated as part of these analysis contain the specific genome names in the organism column in *Extended data*, Supplemental table 7⁴³. Other MLST types have NA in the column.

Extended data

Figshare: Supplemental Table 1. JCVI NTM Genomes. <https://doi.org/10.6084/m9.figshare.12016290.v1>¹⁶.

Figshare: Supplemental Tables 2 & 3: NTM Genomes MASH ANI-based Clusters. <https://doi.org/10.6084/m9.figshare.12024051.v1>³².

Figshare: Supplemental Table 4: NTM Clusters. <https://doi.org/10.6084/m9.figshare.12024057.v1>³⁵.

Figshare: Supplemental Table 5. NTM Genomes. <https://doi.org/10.6084/m9.figshare.12024117.v1>³⁹.

Figshare: Supplemental File 6. Pangenome Matchtable. <https://doi.org/10.6084/m9.figshare.12024126.v1>³⁴.

Figshare: Supplemental Table 7. NTM ST Definitions. <https://doi.org/10.6084/m9.figshare.12024144.v1>⁴³.

Figshare: Supplemental Tables 8 and 9. Analysis of Gupta *et al.* 2018 and Pangenome-Derived Conserved Signature Proteins in NTM Genomes. <https://doi.org/10.6084/m9.figshare.12024159.v2>⁴⁵.

Figshare: Supplemental Table 10: NTM Pangenome-Based ANI All Versus All Table. <https://doi.org/10.6084/m9.figshare.12024024.v2>⁵⁴.

Figshare: Supplemental Table 11: NTM Genome MASH ANI Distances. <https://doi.org/10.6084/m9.figshare.12024030.v2>⁴⁶.

Figshare: Supplemental Table 12. NTM All versus All ST-based Distance Matrix. <https://doi.org/10.6084/m9.figshare.12024303.v1>⁵⁵.

Figshare: Supplemental Table 13. Analysis of Genomes published in Matsumoto *et al.* 2019. <https://doi.org/10.6084/m9.figshare.12024324.v1>⁴⁷.

Extended data are available under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/) (CC-BY 4.0).

Acknowledgements

The authors would like to thank Mario Vaz at the Saint John's Research Institute and Esther Lau at Aeras for their assistance with the collection of the NTM samples, and Karen Beerli, Karrie Goglin, and Indresh Singh at JCVI for their assistance with the sequencing and assembly of the genomes.

References

1. Jang MA, Koh WJ, Huh HJ, *et al.*: **Distribution of nontuberculous mycobacteria by multigene sequence-based typing and clinical significance of isolated strains.** *J Clin Microbiol.* 2014; 52(4): 1207–12. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
2. Tortoli E: **Microbiological features and clinical relevance of new species of the genus *Mycobacterium*.** *Clin Microbiol Rev.* 2014; 27(4): 727–52. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Tortoli E: **Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s.** *Clin Microbiol Rev.* 2003; 16(2): 319–54. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

4. Tortoli E: **The new mycobacteria: an update.** *FEMS Immunol Med Microbiol.* 2006; 48(2): 159–78.
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Wolinsky E: **Nontuberculous mycobacteria and associated diseases.** *Am Rev Respir Dis.* 1979; 119(1): 107–59.
[PubMed Abstract](#)
6. Wayne LG, Sramek HA: **Agents of newly recognized or infrequently encountered mycobacterial diseases.** *Clin Microbiol Rev.* 1992; 5(1): 1–25.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
7. Brown-Elliott BA, Wallace RJ Jr: **Clinical and taxonomic status of pathogenic nonpigmented or late-pigmenting rapidly growing mycobacteria.** *Clin Microbiol Rev.* 2002; 15(4): 716–46.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Poyntz HC, Stylianou E, Griffiths KL, et al.: **Non-tuberculous mycobacteria have diverse effects on BCG efficacy against *Mycobacterium tuberculosis*.** *Tuberculosis (Edinb).* 2014; 94(3): 226–37.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
9. Rowland R, McShane H: **Tuberculosis vaccines in clinical trials.** *Expert Rev Vaccines.* 2011; 10(5): 645–58.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
10. Dhanasekaran S, Jenum S, Stavrum R, et al.: **Effect of non-tuberculous Mycobacteria on host biomarkers potentially relevant for tuberculosis management.** *PLoS Negl Trop Dis.* 2014; 8(10): e3243.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Gopinath K, Singh S: **Non-tuberculous mycobacteria in TB-endemic countries: are we neglecting the danger?** *PLoS Negl Trop Dis.* 2010; 4(4): e615.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Hoefsloot W, van Ingen J, Andrejak C, et al.: **The geographic diversity of nontuberculous mycobacteria isolated from pulmonary samples: an NTM-NET collaborative study.** *Eur Respir J.* 2013; 42(6): 1604–13.
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Gupta RS, Lo B, Son J: **Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus *Mycobacterium* into an Emended Genus *Mycobacterium* and Four Novel Genera.** *Front Microbiol.* 2018; 9: 67.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
14. Euzéby JP: **List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet.** *Int J Syst Bacteriol.* 1997; 47(2): 590–2.
[PubMed Abstract](#) | [Publisher Full Text](#)
15. Parte AC: **LPSN—list of prokaryotic names with standing in nomenclature.** *Nucleic Acids Res.* 2014; 42(Database issue): D613–616.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Clarke T: **Supplemental Table 1. JCVI NTM Genomes.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12016290.v1>
17. Tortoli E, Pecorari M, Fabio G, et al.: **Commercial DNA probes for mycobacteria incorrectly identify a number of less frequently encountered species.** *J Clin Microbiol.* 2010; 48(1): 307–10.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
18. Tortoli E, Fedrizzi T, Meehan CJ, et al.: **The new phylogeny of the genus *Mycobacterium*: The old and the news.** *Infect Genet Evol.* 2017; 56: 19–25.
[PubMed Abstract](#) | [Publisher Full Text](#)
19. Fedrizzi T, Meehan CJ, Grottole A, et al.: **Genomic characterization of Nontuberculous Mycobacteria.** *Sci Rep.* 2017; 7: 45258.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
20. Katahira K, Ogura Y, Gotoh Y, et al.: **Draft Genome Sequences of Five Rapidly Growing *Mycobacterium* Species, *M. thermoresistibile*, *M. fortuitum* subsp. *acetamidolyticum*, *M. canariense*, *M. brisbanense*, and *M. novocastrense*.** *Genome Announc.* 2016; 4(3): pii: e00322-16.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
21. Sutton GG, Brinkac LM, Clarke TH, et al.: ***Enterobacter hormaechei* subsp. *hoffmannii* subsp. nov., *Enterobacter hormaechei* subsp. *xiangfangensis* comb. nov., *Enterobacter roggenkampii* sp. nov., and *Enterobacter muelleri* is a later heterotypic synonym of *Enterobacter asburiae* based on computational analysis of sequenced *Enterobacter* genomes [version 2; peer review: 2 approved].** *F1000Res.* 2018; 7: 521.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
22. Kim M, Oh HS, Park SC, et al.: **Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes.** *Int J Syst Evol Microbiol.* 2014; 64(Pt 2): 346–51.
[PubMed Abstract](#) | [Publisher Full Text](#)
23. Konstantinidis KT, Ramette A, Tiedje JM: **The bacterial species definition in the genomic era.** *Philos Trans R Soc Lond B Biol Sci.* 2006; 361(1475): 1929–40.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
24. Chan JZ, Halachev MR, Loman NJ, et al.: **Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*.** *BMC Microbiol.* 2012; 12(1): 302.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
25. Goris J, Konstantinidis KT, Klappenbach JA, et al.: **DNA-DNA hybridization values and their relationship to whole-genome sequence similarities.** *Int J Syst Evol Microbiol.* 2007; 57(1): 81–91.
[PubMed Abstract](#) | [Publisher Full Text](#)
26. Richter M, Rosselló-Móra R: **Shifting the genomic gold standard for the prokaryotic species definition.** *Proc Natl Acad Sci U S A.* 2009; 106(45): 19126–31.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
27. Zhang W, Du P, Zheng H, et al.: **Whole-genome sequence comparison as a method for improving bacterial species definition.** *J Gen Appl Microbiol.* 2014; 60(2): 75–8.
[PubMed Abstract](#) | [Publisher Full Text](#)
28. Varghese NJ, Mukherjee S, Ivanova N, et al.: **Microbial species delineation using whole genome sequences.** *Nucleic Acids Res.* 2015; 43(14): 6761–71.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
29. Meier-Kolthoff JP, Auch AF, Klenk HP, et al.: **Genome sequence-based species delimitation with confidence intervals and improved distance functions.** *BMC Bioinformatics.* 2013; 14(1): 60.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
30. Colston SM, Fullmer MS, Beka L, et al.: **Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using *Aeromonas* as a test case.** *mBio.* 2014; 5(6): e02136.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
31. Ondov BD, Treangen TJ, Melsted P, et al.: **Mash: fast genome and metagenome distance estimation using MinHash.** *Genome Biol.* 2016; 17(1): 132.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
32. Clarke T: **Supplemental Tables 2 & 3: NTM Genomes MASH ANI-based Clusters.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024051.v1>
33. Clarke TH, Brinkac LM, Sutton G, et al.: **GGRASP: A R-Package for Selecting Representative Genomes Using Gaussian Mixture Models.** *Bioinformatics.* 2018; 34(17): 3032–3034.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
34. Clarke T: **Supplemental Table 4: NTM Clusters.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024057.v1>
35. Fouts DE, Brinkac L, Beck E, et al.: **PanOCT: Automated Clustering of Orthologs Using Conserved Gene Neighborhood for Pan-Genomic Analysis of Bacterial Strains and Closely Related Species.** *Nucleic Acids Res.* 2012; 40(22): e172.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
36. Inman JM, Sutton GG, Beck E, et al.: **Large-scale Comparative Analysis of Microbial Pan-Genomes Using PanOCT.** *Bioinformatics.* 2019; 35(6): 1049–50.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
37. Clarke T: **Supplemental Table 5. NTM Genomes.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024117.v1>
38. Tortoli E: **Commentary: Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus *Mycobacterium* into an Emended Genus *Mycobacterium* and Four Novel Genera.** *Front Microbiol.* 2018; 9: 2065.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
39. Gupta RS, Lo B, Son J: **Corrigendum: Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus *Mycobacterium* into an Emended Genus *Mycobacterium* and Four Novel Genera.** *Front Microbiol.* 2019; 10: 714.
[PubMed Abstract](#) | [Publisher Full Text](#)
40. Mignard S, Flandrois J-P: **A seven-gene, multilocus, genus-wide approach to the phylogeny of mycobacteria using supertrees.** *Int J Syst Evol Microbiol.* 2008; 58(Pt 6): 1432–41.
[PubMed Abstract](#) | [Publisher Full Text](#)
41. Jolley KA, Bray JE, Maiden MCJ: **Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications.** *Wellcome Open Res.* 2018 [cited 2020 Jan 13]; 3: 124.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
42. Clarke T: **Supplemental File 6. Pangenome Matchtable.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024126.v1>
43. Clarke T: **Supplemental Table 7. NTM ST Definitions.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024144.v1>
44. Matsumoto Y, Kinjo T, Motooka D, et al.: **Comprehensive subspecies identification of 175 nontuberculous mycobacteria species based on 7547 genomic profiles.** *Emerg Microbes Infect.* 2019; 8(1): 1043–53.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
45. Clarke T: **Supplemental Tables 8 and 9. Analysis of Gupta et al. 2018 and Pangenome-Derived Conserved Signature Proteins in NTM Genomes.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024159.v2>
46. Clarke T: **Supplemental Table 11: NTM Genome MASH ANI Distances.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024030.v2>
47. Clarke T: **Supplemental Table 13. Analysis of Genomes published in Matsumoto et al. 2019.** *figshare.* Dataset. 2020. <http://www.doi.org/10.6084/m9.figshare.12024324.v1>
48. Brunello F, Ligozzi M, Cristelli E, et al.: **Identification of 54 mycobacterial species by PCR-restriction fragment length polymorphism analysis of the *hsp65* gene.** *J Clin Microbiol.* 2001; 39(8): 2799–806.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
49. Telenti A, Marchesi F, Balz M, et al.: **Rapid identification of mycobacteria to the species level by polymerase chain reaction and restriction enzyme analysis.** *J Clin Microbiol.* 1993; 31(2): 175–8.
[PubMed Abstract](#) | [Free Full Text](#)
50. Hatherill M, Hawkrige T, Whitelaw A, et al.: **Isolation of non-tuberculous mycobacteria in children investigated for pulmonary tuberculosis.** *PLoS One.*

- 2006; 1(1): e21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
51. Nurk S, Bankevich A, Antipov D, *et al.*: **Assembling single-cell genomes and mini-metagenomes from chimeric MDA products.** *J Comput Biol.* 2013; 20(10): 714–37.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
52. Tatusova T, Ciufo S, Fedorov B, *et al.*: **RefSeq microbial genomes database: new representation and annotation strategy.** *Nucleic Acids Res.* 2015; 43(7): 3872.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
53. Brinkac LM, Beck E, Inman J, *et al.*: **LOCUST: a custom sequence locus typer for classifying microbial isolates.** *Bioinformatics.* 2017; 33(11): 1725–6.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
54. Clarke T: **Supplemental Table 10: NTM Pangenome-Based ANI All Versus All Table.** *figshare.* Dataset. 2020.
<http://www.doi.org/10.6084/m9.figshare.12024024.v2>
55. Clarke T: **Supplemental Table 12. NTM All versus All ST-based Distance Matrix.** *figshare.* Dataset. 2020.
<http://www.doi.org/10.6084/m9.figshare.12024303.v1>

Open Peer Review

Current Peer Review Status:

Version 2

Reviewer Report 22 September 2020

<https://doi.org/10.5256/f1000research.26697.r66637>

© 2020 Gupta R. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Radhey S. Gupta

Department of Biochemistry and Biomedical Sciences, Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

This paper pertains to the typing and classification of non-tuberculous mycobacteria (NTM) isolates based on genome sequence information. The methods used include clustering of genomes based on approximate pairwise ANI values using MASH server, generation of pangenomes and using ANI distances from these for construction of a phylogenetic tree, MLST analysis based on sequences of some conserved proteins and identification of proteins that are indicated to be specific for different clusters. The results from these analyses regarding assignment of studied genomes into different *Mycobacteriaceae* genera are compared with the results published by Gupta *et al.* (2018). The authors have also compared the results of conserved signature proteins (CSPs) identified by their analysis with those described by Gupta *et al.* (2018).

This study uses genome sequences from a large number of mycobacterial-related isolates obtained from different geographical regions to understand and classify these genomes. While the objective of this study is important, **I have numerous significant concerns about the results described here and the inferences based on them.** These concerns are noted below and they need be addressed before this manuscript is approved for indexing.

1. In Figure 1, the authors present a phylogenetic tree for the *Mycobacteriaceae* genomes based on ANI distance metric. While the ANI is now widely used for determining overall genomic-relatedness and also for species/subspecies identification, the usefulness of this metric for constructing a phylogenetic tree for a group of species, which are genetically as diverse as the members of the family *Mycobacteriaceae*, is dubious. Hence, the usefulness or reliability of this method needs to be validated. For instance, Odnow *et al.* (2016), in their publication describing the ANI distance estimation by MASH, have emphasized that MASH is not explicitly designed for phylogeny reconstruction, especially for genomes with high genetic divergences or large size differences, as is the case for *Mycobacteriaceae* species. Furthermore, all phylogenetic trees contain information for the branching order and interrelationships among different genomes/species. These relationships are of much importance in evaluating the reliability of a tree. However, these relationships are not

discernible from this tree as no information is provided regarding the names of the type species that are part of different clades. The tree also does not indicate any statistical measure that indicates reliability of the observed relationships. In the absence of such information, the tree shown in Figure 1 is not very informative or useful.

2. The authors indicate that the tree shown in Figure 1 is based on representative genomes. However, it is unclear whether this tree includes a representative from all 269 observed clusters? Furthermore, does it contain information **for all named *Mycobacteriaceae* species** including those for which no genomic clusters were observed? The authors are using this tree to indicate that the clades corresponding to different genera can be distinguished based on ANI distances. Hence, for this claim to be valid it is essential that all named *Mycobacteriaceae* species as well as a representative from different observed clusters be included in this tree.
3. In the tree in Figure 1, the genomes from *Mycobacteriaceae* have been grouped into five genera based on the common ancestor containing the type strains from the five previously described genera. However, can the clades corresponding to these genera be demarcated from each other solely on the basis of ANI metric? If so, what is the range of AAI values (intragenera and intergenera) for the clades corresponding to the five genera and the cut off values for demarcation of these genera? Furthermore, in some claims and statements made in this manuscript, the authors are implying that a tree based on AAI distances is more reliable than a tree constructed based on different large datasets of concatenated protein sequences including the core genome proteins (supported by large numbers of molecular markers) and the phylogenomic tree that forms the basis of [Genome Taxonomy Database](#), which has now become an important resource for microbial phylogeny and taxonomy. In the tree shown in Figure 1 based on ANI distances, the authors have observed a clade (marked with black lines), which contains *Mycobacterium haemophilum* and three genomes from other isolates. The authors believe it be a distinct clade (genus?) and state (page 6 top) "By Gupta's criteria for NTM genera within the *Mycobacteriaceae* this novel clade should be designated as a new genus".

However, I am not sure what the authors mean by this statement. In our work, *Mycobacterium haemophilum* groups reliably and deeply within the genus *Mycobacterium* and it shares the CSIs and CSPs that are specific for this genus. Thus, by our criteria, there is no question or uncertainty regarding the placement of this species within the genus *Mycobacterium*. Thus, when the common ancestor of the genus *Mycobacterium* was identified in this study, this species should have been part of the genus *Mycobacterium* and the branching of this species (clade) outside of *Mycobacterium* points to the limitations/unreliability of ANI approach points for the assignment of species into different genera. Additionally, on page 9, the authors indicate that according to their results the species *Mycolicibacterium aubagnense*, instead of being part of the genus *Mycolicibacterium* should be a part of the genus *Mycobacterium*. However, the placement of *Mycolicibacterium aubagnense* within the genus *Mycolicibacterium* is supported by different phylogenetic analysis (see GTDB database and our unpublished results) and several CSIs and CSPs that it uniquely shares with other members of the genus *Mycolicibacterium*. With this said, based on the tree in Figure 1, it is unclear where *Mycolicibacterium aubagnense* is branching. Hence, the names of all typed species should be shown in this tree. However, these anomalies may represent only some of the many that may exist in the results presented in

Figure 1, but they are not apparent due to the inability to discern the names of different type species and their interrelationships from this tree.

4. There are also a number of other important concerns as well as incorrect statements relating to this tree that the authors have made in this manuscript. These presumably stem from their lack of understanding or unfamiliarity with the taxonomic conventions. For example, on page 5 and 10 (and in Figure 1), the authors state that “Since Gupta supplied only one type strain for the *Mycobacterium* genus, we also found the root node of the largest subtree that included this type strain and no additional Gupta genus type strains”. In this context, the authors should know that in our publication, we transferred many of the species from the genus *Mycobacterium* to other novel genera and for these species the type strains were specified. However, for the species which are retained in the genus *Mycobacterium*, **no change was proposed in their names**. Hence for the genus *Mycobacterium* we only specified the type species, which is *Mycobacterium tuberculosis*. However, for different other species that are part of the genus *Mycobacterium*, the type strain information has not changed and it remains the same as described in the original publications when these species were originally described. This information can be found on the [LPSN webserver](#). Marking of the type species for the genus *Mycobacterium* may affect the demarcation of this genus in Figure 1, as illustrated by the branching outside of this clade of *Mycobacterium haemophilum*.
5. The authors also state in multiple occurrences that in a number of observed genomic clusters, more than one type strain of a species or subspecies was present. However, for any validly published species/subspecies, **there is only one correct type strain**, which was identified when the species was initially described. The same type strain is deposited in diverse culture collections under different names. For example, the type strain for *Mycobacterium haemophilum* is deposited under the following names, (viz. ATCC 29548; CCUG 47452; CIP 105049; DSM 44634; JCM 15465; NCTC 11185), but they should all be same unless contaminated. Thus, in cases where more than one type strain is present in a cluster and they branch separately from each other, only one of these is the correct type strain and this should be identified by comparison (identity) of its 16S rRNA sequence with the 16S rRNA sequence of the type strain. It is incorrect to arbitrarily assign a particular strain, a medoid genome, or the best match, as the type strain for any cluster.
6. In Figure 1, all of the *Mycobacteriaceae* genera except *Mycobacterium* (besides the species from the indicated clades), contain other species that are denoted as *Mycobacterium*, which make these genera polyphyletic. This is seen most clearly for the genus *Mycolicibacterium*, which contains more species in comparison to the other genera. To explain this, the authors have indicated that this mix-up is due to the delay in the renaming of these species/strains according to the criteria used by Gupta *et al.* While this explanation is likely correct for many of the “*Mycobacterium*” species/strains grouping within these other genera, it is also possible that the ANI-based tree construction approach is perhaps misclassifying some of the genomes into these genera. Furthermore, the genus *Mycobacterium* is the only one that appear homogeneous in Figure 1. However, this homogeneity is solely due to the fact that any species/strain that is not assigned to the other genera is placed by the NCBI taxonomy into the default genus *Mycobacterium*. However, it is possible that several of the species/strains that are assigned to the genus *Mycobacterium* by the ANI method may be incorrect, as indicated by the examples of *Mycobacterium haemophilum* and

Mycolicibacterium aubagnense. Thus, to demonstrate that the ANI based analysis can correctly assign different genomes into the appropriate genera, **it is important to confirm the branching of these species/genomes into different genera by other independent means, where the reliability of the branching can be ascertained.**

7. For *Mycolicibacterium vulneris*, its placement into the genus *Mycolicibacterium* was based on a misannotated genome. This problem was corrected and its current classification is indicated on the NCBI database. However, this species does not need renaming, as the authors have indicated. Based on taxonomic rules/conventions, the original name of a species remains in standing even after it has been reclassified. Thus, we cannot reclassify *Mycolicibacterium vulneris* as *Mycobacterium vulneris*, as the name *Mycobacterium vulneris* already exists and it should be used for this species. Thus, I am not sure why the authors have highlighted this problem in their tree and in the text.
 8. The authors note that based on ANI analysis, there were 269 species-level clusters of which 263 are from *Mycobacteriaceae* and of these only 126 clusters contained a type strain. However, the assessment that all of the remaining 143 clusters are from the family *Mycobacteriaceae* appears to be based solely on the branching of these clusters in the tree shown in Figure 1 described. However, if based on ANI analysis, if the genus *Mycobacterium* cannot be reliably demarcated, then how one can be sure that all of these other clusters are part of the family *Mycobacteriaceae* and some of them do not branch outside of it? It should be mentioned that the NTM is a very broad term and it could include members from other bacterial families/orders outside of *Mycobacteriaceae*. Thus, many sequences deposited in the NCBI database as *Mycobacterium* spp. may not even belong to the *Mycobacteriaceae* family. The genomes from clinical samples that the authors have sequenced as part of this study are all assumed to belong to the family *Mycobacteriaceae* based on Acid-fast positive staining. However, the Acid fast positive response is not a unique property of the *Mycobacteriaceae* and it is shared by several other genera/families within the order *Corynebacteriales*. Thus, it is important to confirm by other means that all of the novel NTM clusters lacking any type strain are indeed a part of the family *Mycobacteriaceae*. A simple way to confirm this will be to carry out Blastp (or Blastn) searches on some conserved genes/proteins from each cluster and ensure that all of the top hits are from *Mycobacteriaceae* species and not from some unrelated families/genera.
 9. I also do not understand the following statement made by the authors: "These genus reassignments based on ANI should be confirmed with supporting growth and metabolic observations similar to the Gupta *et al.* results". The authors should know that our identification of genera was based on phylogenomics and comparative genomic studies identifying molecular markers specific for different genera and "not based on growth and metabolic observations"
 10. In multiple places throughout the manuscript, the authors refer to our earlier work as "Gupta recently proposed", "Gupta also proposed" "Gupta supplied", "Gupta-defined", "Gupta assignment" "Gupta genus", "work of Gupta" and many other instances. However, this is not the appropriate convention that should be used to refer to our work, which should be referred to as Gupta *et al.*^{Ref}
- Another form of analysis the authors have carried out is a MLST schema for differentiation of different NTM clusters identified on the basis of ANI clustering. This MLST analysis is based on

genes for 8 ribosomal proteins viz. S7, S8, S12, S14Z, S19, L16, L19 and L35. The authors chose these genes/proteins due to their showing minimal variation in gene length and highest conservation within the studied genomes. There are a number of questions concerning the design of these studies and their overall utility.

1. As stated by the authors at multiple places in this paper, the main objective of the MLST analysis is that this form of analysis or approach can be employed to differentiate among different species/strains **when whole genome sequencing is not feasible**. They argue that because these genes are highly conserved at the nucleotide level, PCR primers can be designed to amplify these genes to carry out such analysis. **However, I see many serious limitations in carrying out comparable analysis based on these gene sequences when the whole genome sequences are not available.** First, the authors have not provided any information regarding the primer sequences, so that specific sizes of gene fragments that are to be used for MLST can be amplified. Furthermore, as they have used full length sequences for these genes in their MLST analysis, I do not see how someone can amplify full length sequences for these genes, in the absence of genome sequences. Second, even if one undertakes the designing of PCR primers based on conserved regions within these genes, the overall sequence length of the gene fragments will be much smaller than used in the present analysis. Thus, the results obtained cannot be compared with those described here. It should be noted that all of the genes that the authors have chosen for these analyses are quite small and their sequence length ranges between 190-468 nucleotides (nt). Depending upon where in these genes the PCR primers can be designed, and after excluding the sequences for the primers in MLST analysis, a very small fraction of these gene sequences will be available for MLST analysis. In contrast to the results presented here, there are several detailed studies involving MLSA (see reference 40), where the authors have provided detailed information for the PCR primers for many highly conserved genes and have used them for differentiation of mycobacterial species. However, the authors have discounted these studies and instead suggest that their choice of gene sequences is superior. In view of the concerns noted here, it is important for the authors to justify in an unbiased manner their choice of the MLST genes and how they will be used for PCR-based analysis, when the genome sequences are not available.
2. The authors have also not indicated how many informative positions are available in their concatenated sequence alignment of the genes, which were used for MLST analysis. Given the very small sizes and high degree of conservation of these genes, the overall information content of this gene set is expected to be small and its resolving ability should be limited. In comparison, a single conserved gene such as RpoB or RpoC will have more genetic information (between 3500-4000 nt) than the combined sequences of all of these genes (approx. 2700 nt). Thus, it is unclear why the authors have chosen these particular genes and why anyone carrying out such analysis should use these gene sequences, whose information content is limited and which do not appear suitable for experimental analysis. In contrast to this study, a recent paper by Matsumoto, which the authors have referred to, has used 184 genes, including 53 core genes and 133 species-specific genes. The resolving ability of this MLST dataset should be much greater than the present dataset, if MLST analysis is carried out on the basis of whole genome sequences.
3. The MLST/MLSA analysis commonly involves construction of a phylogenetic tree based on the concatenated sequences. The tree contains much useful information that is not apparent from the ST definitions. Hence, it is important that a phylogenetic tree based on

this MLST dataset be included in this work.

4. The authors also indicate that using whole-genome data, ANI comparison is preferable to an MLST approach and it is sufficient for species categorization. Then, could the authors explain what is the advantage or utility of the MLST data based on the described sets of proteins?

Lastly, the authors have used the JCVI pan-genome pipeline to generate a pan-genome of the *Mycobacteriaceae* species. Based on this pangenome, the authors have Identified 648 CSPs that are indicated as genera specific. The authors have seemingly compared the results/characteristics of these CSPs with the CSPs described in our work. Based on their analyses, the authors have reached the inference "Overall using the CSPs to assign genus seems to have limited utility compared to MASH ANI or MLST". . Additionally, they have also expressed "skepticism" regarding our use of CSPs for identification of species from different genera. As this work and the authors inferences directly relates to our work, I have many questions regarding these analyses. However, before these questions can be asked, it is important that the authors clarify the results presented in Supplemental Tables 8 and 9 on the species distributions of various described CSPs, as I have found them difficult to understand. In this regard, it would be helpful if the authors could clarify the following points taking the example of CSPs specific for the genus *Mycobacteroides* (lines 186-642), which comprise >70% of all CSPs identified by the authors. First, for the CSPs that are specific for *Mycobacteroides*, some are indicated to be specific for this genus, whereas others are present in 100% (or a very large fraction) of the other *Mycobacteriaceae* genera. With this said, how do these two types of CSPs differ from each other and why are the CSPs that are present in other *Mycobacteriaceae* genera considered as CSPs for this genus? Second, besides the *Mycobacteroides*, are other significant Blast hits from other genera outside of *Mycobacteriaceae* observed for these CSPs?

In order to properly evaluate and compare the significance of the CSPs identified by the authors, I am providing in the attached [Table](#), the results of Blastp searches for two CSPs described in our work that were indicated as specific for the genus *Mycobacteroides*. *These blast searches were carried out against the nr database using default settings, except the total number of hits examined were 500.* These CSPs are representative of the CSPs described in our work and the results of blast searches for the other CSPs could be discussed at a later time, if necessary. In both these cases, all significant blast hits are observed for the species/strains from the genus *Mycobacteroides* (include all named species from this genus as well as many unnamed *Mycobacteriodes* spp.) However, in both these cases no significant hit is observed from other *Mycobactriaceae* genera or species from other bacterial phyla. Thus, these CSPs are unique characteristics of the members of *Mycobacteroides* genus and they exhibit high degree of specificity and predictive ability for the members of this genus.

In order to compare the results from this study with our CSPs, it will be helpful if the authors can provide the results of similar Blast searches (carried out against nr database) for some of the CSPs specific for the genus *Mycobacteroides* that they have identified. I suggest that they provide Blastp results for four of their *Mycobacteroides*-specific CSPs from Table 9 (YP_001701589.1, YP_001701270.1, WP_005083466.1 and YP_001701583.1) covering the two types of CSPs for this genus that they have described. The results should be shown for top 500 hits, as I have shown for the two CSPs from our work. The results from these studies will be very helpful in our critically evaluating the significance of the CSPs identified in the present work with those identified by our lab. Furthermore, it would allow us to respond accordingly to the inferences reached by the

authors either here or in a separate publication.

The authors should indicate what the term “**Noveigenus**” refers to in Table 9. If this is a “new genus”, why there is no mention of it in the entire manuscript? It is highly unusual **and inappropriate** to include the name of a “new genus” in the Supplementary material, without any mention of it in the text.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

Not applicable

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: This study extensively refers to our work and many of the inferences are based on our work. Additionally, this study also compares the results obtained here with our results. However, many of the inferences drawn from these comparisons are of dubious nature and they are not supported by the results presented here.

Reviewer Expertise: Phylogenomics, Comparative genomics, Prokaryotic and mycobacterial classification.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 June 2020

<https://doi.org/10.5256/f1000research.26697.r64044>

© 2020 Warden C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Charles Warden

Integrative Genomics Core, City of Hope National Medical Center, Duarte, CA, USA

Thank you very much for your detailed response.

I think there is at least one more thing that needs to be changed. However, since this is open peer-review, I chose a status of “**Approved**” this time, meeting the criteria of requiring “minor changes”. More specifically, I think some more information about the **critical assessment of the assemblies** is needed.

If some additional smaller number of changes are made, then I would be OK with the paper if everybody else is OK with the paper (such as the 2nd reviewer or other readers that can comment at any point).

More specifically, these are the points where I think more information is needed:

1) In general, for most bioinformatics applications, I would say some trial and error is required for each project. While I am most familiar with this in the RNA-Seq context, I have done some work with *de novo* assemblies (for DNA-Seq and RNA-Seq). For the *de novo* assemblies that I have worked with, I would say that it is not uncommon to find something missing in an assembly that was truly present in the genome for a sample. However, if I could assemble a sequence in at least 1 sample in a set of related samples, then I could use a direct alignment to quantify/test the presence of that sequence in samples that might miss the sequence from the *de novo* assembly.

So, I can think of a couple of ways in which this might be relevant (the first of which I think is important for a revision):

i) What is the **short read alignment rate** for the assembled genome for its representative assembly? I don’t know what the exact thresholds for quality flags should be, and maybe it is OK to just provide the continuous alignment rate for a given aligner. However, my understanding is that we agree that having some coverage metrics can be useful.

I am not sure if we are already on the same page, but I am talking about the sequence that you believe is the true genome sequence and that you deposited separately (probably the largest contig?) and not the full set of assembled contigs (where you could theoretically assemble contamination and/or contigs may not be robust).

ii) If there are any functionally important sequences, perhaps a direct alignment of short reads to the gene sequence could be tested? This may be less important and/or harder to implement than **i)**, but it is something that I could think of.

2) I think the following sentence was interesting: “*That being said[,] the NTM samples exhibited a higher than normal rate of mixed samples[,] possibly due to slow growth rate and the isolation of the DNA at the collaborating laboratories.*”. Can you please either point out where this is mentioned in the paper or add a paragraph in the next revision? It might be OK to simply provide a slightly differently formatted version of the response paragraph.

Without listing them all here, thank you for your detailed response to the assembly assessments.

These are excellent, and I think they will help the readers if they check the open peer-review.

Otherwise, there are the current minor points:

- Maybe you already have an answer to one of the above questions, but you mention **alignment coverage** as a QC metric. I think statistics like the mean coverage, standard deviation or quartiles of coverage, and/or percent of bases with X coverage are useful, but I am having a hard time finding them in the supplemental materials. For example, in Supplemental Table 1, I see “*Contig N50 (bp)*”, “*Number of Contigs*”, and “*Total Length (bp)*”, but I don’t see the coverage values. Can you please indicate where I can find those values, or add them to Supplemental Table 1?
- It was a good response that you can flag assemblies when the size of the SPAdes and Velvet assemblies considerably varies. I understand that the Velvet assembly may simply sometimes not be very good. However, perhaps it could help to add that to Supplemental Table 1?
- Part response to the review says “**Supplemental Table X and Y** for more detailed descriptions of the subspecies ANI ranges for the different sub-species”. I think you meant to fill these in with numbers? Supplemental Tables 2 and 3 have 95% and 98% MASH ANI clusters. Supplemental Table 11 has all of the MASH ANI distances. I see the top 2 matches for ANI and ST distances in Supplemental Table 5. So, I would guess you meant to refer to 2 of those tables?
- In terms of Figure 2, I think this means the ANI and ST measures are usually (but not perfectly) consistent, and perhaps this is more of what you wanted to emphasize. This would be a little different than plotting the ranges for pairwise distances within *Mycobacteroides*, *Mycobacterium*, *Mycolicibacterium*, *Mycolicibacter*, and *Mycolicibacillus* (showing the variation within each of the current 3 distance categories). I still think the later might be interesting, but this can be optional (since pointing out the outliers was probably more important).
- I am personally trying to think of what to recommend for reproducible and sustainable research in the future. I usually have to go back and revise code throughout a project, so I understand that this can make re-running the code more complicated than a simple wrapper. This also affects the total number of projects that can be worked on, if it is possible to accurately estimate time for post-publication support and review (probably with some over-estimation, since under-estimation of time/effort requirements can compromise quality/safety). I think this is important to consider for the future, but you have provided a lot of details and I think that is OK with current expectations.
- I almost felt bad when I thought you were moving revision #1 authors to the acknowledgments. While authorship needs to be viewed as a responsibility that can be either good or bad in the long-run, I see that you mean everybody currently in the Acknowledgements is OK with being in that section (in the original submission). While the type of contribution was noticeable, I would say this was a minor point, and I won’t worry about this (if I need to provide any additional responses).
- Finally, you don’t have to call me “Dr. Warden” (in part, because I don’t have a PhD or

MD). For example, I am just as happy if you call me "Charles."
Thank you very much for your hard work.

Is the work clearly and accurately presented and does it cite the current literature?

No

Is the study design appropriate and is the work technically sound?

No

Are sufficient details of methods and analysis provided to allow replication by others?

No

If applicable, is the statistical analysis and its interpretation appropriate?

No

Are all the source data underlying the results available to ensure full reproducibility?

No

Are the conclusions drawn adequately supported by the results?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics and Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 27 April 2020

<https://doi.org/10.5256/f1000research.24972.r62617>

© 2020 Warden C. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Charles Warden

Integrative Genomics Core, City of Hope National Medical Center, Duarte, CA, USA

Thank you very much for putting together this publication.

I think this is a useful paper where you have done a good job overall with being transparent, and I hope it can become a peer-reviewed version soon.

I think these may be my most important questions (related to clustering in Figure 1):

- It might already be described somewhere (or split among different locations), but I think some discussion is needed for the large cluster with mixed *Mycobacterium* and *Mycolicibacterium* (in the top-right of Figure 1). I have general genomics experience (including working with some pathogen data), but the biological meaning for this mixed portion of the dendrogram was not clear to me (since I don't have specific experience working with *Mycobacteria*). I would guess this could be true for other F1000Research readers. Is it possible to provide this in a way that is easy to find (such as a section in the results, or a paragraph in the discussion)?
- For example, am I correctly understanding Figure 1 and Figure 10 in the Gupta *et al.* 2018 paper show classifications that cluster together (which would be different than the Figure 1 in this paper)? I think this has something to do with whether or not there is a green *Mycolicibacterium* dot on the outside of the tree, but then the question would be what to think about the Gupta *Mycobacterium* genomes within that branch (in Figure 1 of this paper)?
- Likewise, it looks like the "subgenus" separates more clearly in Figure 1 of the Matsumoto *et al.* 2019 paper. You mention this paper multiple times, but I don't believe that I see a sentence that directly compares the results presented in that study to Figure 1 of this study. Unless I have missed something, I think that may be useful to add.

Otherwise, these are my questions related to the interpretation of the results:

- The separation in Figure 2 looks pretty good, but there are some outliers. Are these well-characterized examples, or could these be misclassifications?
- Is there a way to connect Figure 1 and Figure 2? Figure 2 was what I noticed first and felt like I could understand. However, could you do something like separate "Same sub-species" into separate clusters (such as the 5 clusters in Figure 1, if I understand the content)? So, if some clusters have more variation than others, then that would become easier to see. Maybe this should be a Figure 3, for reasons of space.
- If you have performed automated assembly for all of the samples, what sort of checks have you done to justify important differences are due to true biological sequences rather than assembly artifacts (which may or may not be random)? You do mention that "good assemblies" were used from SPAdes and provide some statistics in Supplemental Table 1. However, I think it might help if **i)** more information about the criteria for "good" can be provided and **ii)** if any additional factors can be considered.
- For example, there is noticeable variation in the N50 values. Are there any values that tend to be more common in assemblies that don't have a type strain or previous genome sequence (and/or that appear in the branch with mixed clusters in Figure 1, form small clusters, etc.)?
- Are there any features that vary with other expected (or known) genome metrics, such as the proportion or maximum length of repeat sequences?

- Additionally, in Supplemental Table 1, some samples are listed as “No Assembly.” Are there QC measurements that might be able to be used to flag a sample to be removed prior to analysis. This by itself provides a possible flag, but this and the earlier questions relate to whether some samples can generate an assembly but it may be less reliable than others (whether that is due to some limitation in the sample, a limitation in the total number of reads, etc.).

I also have some other minor notes that might help with communicating your results:

- It may help readability to add numbers for the items listed after “There are four major facets of our analysis” in the Results. While the list is smaller, something similar could be done for “genome relatedness across two major comparators”. This is already in the results section (versus the introduction section, for example), but maybe it would even help to split this into 2 paragraphs and provide more of an explanation? Or, may be a visualization can help with the analysis strategy as a figure? I had a hard time quickly and effectively understanding what was meant to be communicated here.
- It looks like you have done a good job of sharing the raw data and providing a lot of details in supplemental files/tables. I had to start reading the paper to find it, but you also mention sharing code on GitHub. However, am I correctly understanding that these are for the programs that you used (versus code to reproduce the results for this specific study)? If you essentially have 1 wrapper script, perhaps adding that within a .zip file (to see the versions and parameters used) could be sufficient?
- The abstract mentions using the method “MASH”. The methods provide a link (<https://github.com/marbl/Mash>) that matches the cited reference for a pre-existing method called Mash (Ondov *et al.* 2016, Genome Biol. 2016 Jun 20;17(1):132). When I re-read the abstract, I can see that this doesn’t mean you *introduced* a method called MASH. I am not sure what other readers think (and this should be obvious to those with experience using that program), but I think one thing that could make this more clear to other reads would be to say “*We used Mash to classify NTM genomes*” (versus “*We describe using a fast average nucleotide identity (ANI) approximation method, MASH, for classifying NTM genomes*”). Or, am I overlooking something?
- It seems like a lot of the authors are described as “Investigation, Resources, Writing – Review & Editing”, which kind of sounds like the contributions listed for those in the acknowledgments. Is there some sort of threshold for the amount of work used to identify somebody as an author versus being listed in the acknowledgments? Maybe this even means more authors should be added – I just thought the number of individuals with this exact set of contributions was noticeable.
- On the flip side, I noticed that you said “*We leave it to those more familiar with Mycobacteriaceae species to make those determinations*” – while I can understand that it might be difficult to provide all possible checks for all possible samples, is there at least one collaborator that might be able to list some exceptions/limitations to purely using the ANI strategy for assignments (which could be an acknowledgment, if we are talking about a contribution of a few points without revising any analysis)?
- I think there should only be one period in a sentence in the introduction (either before or

after the quote); different NTM species were encountered in this survey.”.

- I think there is a minor formatting issue where only part of the word is italicized in the Introduction: *Mycolicibacter*gen (“gen” is not in italics)
- I think there should be a comma in the following sentence in the introduction: “*While this was an important survey of NTM geographical distribution the low number of NTM species*” à “*While this was an important survey of NTM geographical distribution, the low number of NTM species*”.
- I think there is some unintended capitalization in the results: “*of a given genus Using our*” à “*of a given genus using our*” (this don’t look like the end of a sentence, so I don’t think the “U” should be capitalized)

I think changes are needed (beyond an optional “suggestion”), but I think they can probably be made so that the next revision can be accepted. This is what I was trying to communicate with a decision of “**Accept with Reservations**”.

Thanks Again!

Is the work clearly and accurately presented and does it cite the current literature?

Yes

Is the study design appropriate and is the work technically sound?

Yes

Are sufficient details of methods and analysis provided to allow replication by others?

Partly

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Yes

Are the conclusions drawn adequately supported by the results?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics and Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 May 2020

Thomas Clarke, J Craig Venter Institute, Rockville, USA

We would like to thank Dr. Warden for his thoughtful and critical comments for our manuscript. Below, we have listed the comments (bold font) and our responses (italics font).

I think these may be my most important questions (related to clustering in Figure 1):

- **It might already be described somewhere (or split among different locations), but I think some discussion is needed for the large cluster with mixed Mycobacterium and Mycolicibacterium (in the top-right of Figure 1). I have general genomics experience (including working with some pathogen data), but the biological meaning for this mixed portion of the dendrogram was not clear to me (since I don't have specific experience working with Mycobacteria). I would guess this could be true for other F1000Research readers. Is it possible to provide this in a way that is easy to find (such as a section in the results, or a paragraph in the discussion)?**

Thank you for bringing this up. It is our understanding that the genera mixture in the top-right clade is indicative solely of the delay in NCBI renaming species according to their Gupta-specified genus, and that the species labeled Mycobacterium are in fact Mycolicibacterium. We have added a sentence to the figure 1 legend clarifying this view ("Some of the NCBI genus names listed do not account for the Gupta et al genus definitions, and thus branch and leaf color disagreements probably only reflects a difference in genus naming conventions"), as well as added a clause to the sentence (shown in bold): "This means the current NCBI names are not consistent with our clade-based assignment, as seen in the occasional differences in leaf and branch colors in Figure 1 and that these differences do not necessarily reveal divergent classification of the species using Gupta based genera." However as we have not investigated the biological properties of each of the genomes in this tree, we can not conclusively state that they match the Gupta et al physical definitions of the genera as we clarify in an added sentence at the conclusion of the paragraph.

- **For example, am I correctly understanding Figure 1 and Figure 10 in the Gupta et al. 2018 paper show classifications that cluster together (which would be different than the Figure 1 in this paper)? I think this has something to do with whether or not there is a green Mycolicibacterium dot on the outside of the tree, but then the question would be what to think about the Gupta Mycobacterium genomes within that branch (in Figure 1 of this paper)?**

As specified in the response above, we do not feel that the genome labeled as Mycobacterium are in fact Gupta Mycobacterium but instead old style Mycobacterium where the name has not yet been updated. We found no instances of a genome classified using the Gupta definitions within a clade that was not assigned that genus. We have attempted to clarify the language describing this as noted above, including in the Figure 1 legend.

- **Likewise, it looks like the "subgenus" separates more clearly in Figure 1 of the Matsumoto et al. 2019 paper. You mention this paper multiple times, but I don't believe that I see a sentence that directly compares the results presented in that study to Figure 1 of this study. Unless I have missed something, I think that may be useful to add.**

We agree with this. Currently, the data in figure 1c in Matsumoto currently forms part of our analysis as described in the paragraph and in supplemental table 13, so we have added a clarifying clause in the text. We further agree that additional discussion comparing our findings

would be useful. We have added three sentences comparing our results with those of Matsumoto when discussing the placement of the Matsumoto et al genomes in Figure 1C, including our disagreement of the *M. avium* placement and the deeper resolution of our full pagenome tree in Figure 1 in our manuscript.

Otherwise, these are my questions related to the interpretation of the results:

- **The separation in Figure 2 looks pretty good, but there are some outliers. Are these well-characterized examples, or could these be misclassifications?**

The species and subspecies pairs in figure 2 was defined as noted using MASH-based ANI to generate as complete and consistent a dataset as possible for the comparison, since there are no sub-species pairs included in the Pan-Genome analysis, and the similarly named species could have included both pairs at both species and sub-species levels and thus given inconclusive results. As such, though the ANI values might elide or remove biologically defined subtypes, we feel it is the most consistent, though it is possible that genomes listed as sub-species are biologically more realistically defined as species

- **Is there a way to connect Figure 1 and Figure 2? Figure 2 was what I noticed first and felt like I could understand. However, could you do something like separate "Same sub-species" into separate clusters (such as the 5 clusters in Figure 1, if I understand the content)? So, if some clusters have more variation than others, then that would become easier to see. Maybe this should be a Figure 3, for reasons of space. We were able to separate the different sub-species into clusters using the 98% Mash ANI cutoff. The phylogeny of the full genome set including sub-species set was too large to show details of the spread of different sub-species groups, but we can point to Supplemental Table X and Y for more detailed descriptions of the subspecies ANI ranges for the different sub-species, though by definition all of our sub-species clusters are limited to a 2% ANI range**
- **If you have performed automated assembly for all of the samples, what sort of checks have you done to justify important differences are due to true biological sequences rather, than assembly artifacts (which may or may not be random)? You do mention that "good assemblies" were used from SPAdes and provide some statistics in Supplemental Table 1. However, I think it might help if i) more information about the criteria for "good" can be provided and ii) if any additional factors can be considered.**

At the time of assembly and submission the JCVI pipeline ran two assemblies: SPAdes and Velvet. SPAdes tended to give more contiguous assemblies and wasn't obviously inferior by other metrics. JCVI generated excel spreadsheets of a variety of statistics at the assembly level such as genome size and N50 but also produced data on a per contig basis. One flag of possible assembly issues was when the SPAdes and Velvet statistics such as genome length were not similar. Contig level stats included coverage levels of all contigs and best Blast matches against NCBI genomes. Built into the assembly process contigs with very low coverage versus the median coverage of large contigs were filtered out of the assembly but overall statistics as well as per contig statistics were retained for the filtered contigs. It was not uncommon to see small contigs with coverage depth 2-5x versus primary large contigs with coverage depth 100x or higher where the small contigs would match Ralstonia a common sequencing materials contaminant. Sometimes it was clear that low coverage contigs were a second NTM species but still low enough coverage to be effectively filtered out. Mixed samples where contigs were not at least in a 10 to 1 ratio were deemed inseparable even if 90% of contigs could be easily separated and were not submitted. Besides coverage levels, mixed samples could be seen by outlying genome sizes, varied Blast

matches, and more hard to interpret poor assembly (lots of contigs and low N50). We did not submit mixed samples which could not be corrected by filtering out very low coverage contigs. We did not submit assemblies with more than 500 contigs (after filtering out very short and very low coverage contigs) or an N50 < 15kp. While the assembly pipeline was automated as well as the gathering of data to assess the quality of the assembly and the chance of a sample being inseparably mixed the final decision was taken by Granger Sutton who has more than 20 years of experience evaluating assemblies following JCVI guidelines. That being said the NTM samples exhibited a higher than normal rate of mixed samples possibly due to slow growth rate and the isolation of the DNA at the collaborating laboratories. While the authors stand behind the quality of the 201 submitted samples for the BioProject, it is certainly possible that not all contamination/mixing was detected and removed.

- **For example, there is noticeable variation in the N50 values. Are there any values that tend to be more common in assemblies that don't have a type strain or previous genome sequence (and/or that appear in the branch with mixed clusters in Figure 1, form small clusters, etc.)?** *This is a very good question which we did not consider for either the JCVI genomes or non JCVI genomes. While genome assembly quality, particularly the non-removal of contamination or mixing of strains, could cause bad ANI values other factors could affect the likelihood that a genome is classified into a cluster with or without a type strain. For instance, genome studies targeting a single outbreak such as the case for many of the *M. abscessus* genomes which numerically dominate the genomes under study here would bias the results of any simple correlation. We did look at the correlation between N50 and number of contigs to whether a genome was placed into a cluster with a type strain. No significant logistic correlation between MASH mapping with existing type strain and quality of assembly was observed either using the number of contigs or the N50.*
- **Are there any features that vary with other expected (or known) genome metrics, such as the proportion or maximum length of repeat sequences?** *That is another good question that was not the focus of our work. We did not look at repeat content but it has been our experience that there are four major factors influencing N50 size: sequence quality (separate metrics are kept for this and usually flagged before assembly), repeat content, coverage depth (usually once above 30x it is not much of an issue), and the presence of multiple strains of the same species (this is not the same as mixed samples with different species which still tend to assemble well but can be difficult to separate entirely with confidence due to high copy repeats, low copy plasmids, etc). Lower than expected N50 and higher contigs counts in our experience can indicate multiple strains of the same species where variable regions between the conserved core genome regions can chop up the assembly – one call also see an increase in “SNPs” indicating population variation. We try to recognize these mixed strains and not submit them. Studying the actual repeat profile of different NTM species and how that affects assembly while interesting is beyond the scope of this paper.*
- **Additionally, in Supplemental Table 1, some samples are listed as “No Assembly.” Are there QC measurements that might be able to be used to flag a sample to be removed prior to analysis. This by itself provides a possible flag, but this and the earlier questions relate to whether some samples can generate an assembly but it may be less reliable than others (whether that is due to some limitation in the sample, a limitation in the total number of reads, etc.).**

The No Assembly samples are samples that we created BioSamples for based on collaborator input but for which no DNA ever arrived, library prep failed, or assembly was deemed low quality and not submitted. After that initial experience we waited to create BioSamples until a successful assembly existed. In total we received 273 samples of which we submitted assemblies for 201 samples.

I also have some other minor notes that might help with communicating your results:

- **It may help readability to add numbers for the items listed after “There are four major facets of our analysis” in the Results. While the list is smaller, something similar could be done for “genome relatedness across two major comparators”. This is already in the results section (versus the introduction section, for example), but maybe it would even help to split this into 2 paragraphs and provide more of an explanation? Or, maybe a visualization can help with the analysis strategy as a figure? I had a hard time quickly and effectively understanding what was meant to be communicated here.***We agree and have added “first”, “second”, “third”, and “final” and changed the commas to semicolons to the list for clarification, and a comma to the second list to also help accentuate the different type.*
- **It looks like you have done a good job of sharing the raw data and providing a lot of details in supplemental files/tables. I had to start reading the paper to find it, but you also mention sharing code on GitHub. However, am I correctly understanding that these are for the programs that you used (versus code to reproduce the results for this specific study)? If you essentially have 1 wrapper script, perhaps adding that within a .zip file (to see the versions and parameters used) could be sufficient?***All the programs listed above were run separately and in parallel so we feel the addition of a wrapper script would not add much to assist in their usage beyond what is listed here.*
- **The abstract mentions using the method “MASH”. The methods provide a link (<https://github.com/marbl/Mash>) that matches the cited reference for a pre-existing method called Mash (Ondov *et al.* 2016, Genome Biol. 2016 Jun 20;17(1):132). When I re-read the abstract, I can see that this doesn’t mean you *introduced* a method called MASH. I am not sure what other readers think (and this should be obvious to those with experience using that program), but I think one thing that could make this more clear to other reads would be to say “We used Mash to classify NTM genomes” (versus “We describe using a fast average nucleotide identity (ANI) approximation method, MASH, for classifying NTM genomes”). Or, am I overlooking something?**

The reviewer is correct that Mash is a previously released and highly used program. We have added a couple of descriptors to MASH in our abstract to more fully clarify this.

- **It seems like a lot of the authors are described as “Investigation, Resources, Writing – Review & Editing”, which kind of sounds like the contributions listed for those in the acknowledgments. Is there some sort of threshold for the amount of work used to identify somebody as an author versus being listed in the acknowledgments? Maybe this even means more authors should be added – I just thought the number of individuals with this exact set of contributions was noticeable.**

This manuscript is dependent on and describes novel NTM sequences which were collected from multiple sites around the globe as is seen in supplemental table 1 with JCVI providing the

sequencing and assembly. The authors listed were instrumental in both collecting and assisting in the collection of the samples that were used in the paper, and thus they had very similar contributions. We reached out to all the authors and some felt that their contributions were more consistent with an acknowledgement and those we have placed in that section.

- **On the flip side, I noticed that you said “We leave it to those more familiar with Mycobacteriaceae species to make those determinations” – while I can understand that it might be difficult to provide all possible checks for all possible samples, is there at least one collaborator that might be able to list some exceptions/limitations to purely using the ANI strategy for assignments (which could be an acknowledgment, if we are talking about a contribution of a few points without revising any analysis)?** We agree that this is a good idea, but all of the collaborators listed above are involved in health care, sequencing, or bioinformatics and thus not experts in microbiological techniques that we think would be required to analyze the traits of the different strains as described in the manuscript. We feel that this would be necessary to confirm the genera assignment similar to how Gupta et al. clarified their division as stated, and that this is currently outside the scope of this manuscript and the authors.
- **I think there should only be one period in a sentence in the introduction (either before or after the quote): different NTM species were encountered in this survey.”.**
- **I think there is a minor formatting issue where only part of the word is italicized in the Introduction: Mycolicibactergen (“gen” is not in italics)**
- **I think there should be a comma in the following sentence in the introduction: “While this was an important survey of NTM geographical distribution the low number of NTM species” à “While this was an important survey of NTM geographical distribution, the low number of NTM species”.**

We agree with all the above suggestions, and have implemented the changes suggested.

- **I think there is some unintended capitalization in the results: “of a given genus Using our” à “of a given genus using our” (this don’t look like the end of a sentence, so I don’t think the “U” should be capitalized)**

We agree that this is not correct. It should be the end of a sentence and we have added a period.

Thomas Clarke on behalf of all the authors

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research