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

Thomas Clarke¹, Lauren Brinkac¹,², Joanna Manoranjan³, Alberto García-Basteiro⁴,⁵, Harleen Grewal⁶, Anthony Kiyimba⁷, Elisa Lopez⁴,⁵, Ragini Macaden⁶,⁸, Durval Respeito⁴,⁵, Willy Ssengooba⁷,⁹, Michele Tameris¹⁰, Granger Sutton¹

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

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 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.

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
Introduction

Non-tuberculous mycobacteria (NTM) are a ubiquitous group of diverse environmental mycobacteria related to *Mycobacterium tuberculosis*, the bacterium that causes tuberculosis (TB)\(^1\). 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\(^3\), established immuno-diagnostic methods, and the specificity of diagnostic and immune biomarkers\(^4\). 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 NTMs\(^5\). The NTM-NET published a survey of NTM patient isolates\(^6\): “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*, *Mycolicibacterien*, *Mycolicibacterium*, *Mycolicibacillus*, and *Mycobacteroides*\(^8\) 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)\(^9\) 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\(^10\) 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\(^11,12\). 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\(^13-20\). 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: clustering of *Mycobacteriaceae* family genomes using ANI, a pan-genome of representative genomes from each clade determined by ANI, an evaluation of our MLST scheme, and 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* family\(^7\). 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 S1\(^10\)) 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 approach. New clusters were then compared to our pan-genome to the work of Gupta in terms of genera assignments, species and subspecies definitions\(^21-30\). 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 table S2 and S3\(^31\)) using GGRAsP\(^32\). This has been shown to be a reasonable way to computationally computer species/subspecies clusters\(^32,33\). 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 S3\(^31\)). 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 S3). 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 S4). For each of the final 280 *Mycobacteriaceae* clusters, a representative genome was determined using GGHaSP 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. 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-genome 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 pan-genome 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 S4) 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.* 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.* which we also observed: some type strains have been genome sequenced more than once where the genomes are clearly different species from each other; 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 S4 we give all of the type strains in the cluster and in Extended data, Supplemental table S5 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 genomosp. 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 genomosp. 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). 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 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 S4, pending assignment of an additional genus name. NCBI has accepted the Gupta proposal in naming NTM species except for the misidentified
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.

Of the representative type strains, four have NCBI genus names which disagree with our assignment. These are labeled as *Mycobacterium* (*neumannii*, *syngathidarumi*, *dioxanotrophicus* and *aquaticum*), but are located in the *Mycolicibacterium* subtree. One of them (*nuemannii*) 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 S4. 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 S5). 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 *Mycobacteroides* (chelonae subsp. gwanake and subsp. chelonae), and two as *Mycolicibacterium* (kyoganese and lehmannii).

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 S5.

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, PRJDB4227, PRJNA299467) 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 S1) 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 *Mycolicibaculites* 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 past 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. 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 S6).

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 S7. Using our MLST schema an ST is never shared across species/subspecies clusters (Extended data, Supplemental table S5). 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 S5). Only for 142 genomes was the closest ST different than the ANI assigned cluster (Extended data, Supplemental table S5). 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 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 mlstverse 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 mlstverse 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 S8[45])3.
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 S9). Only a small number 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. Most 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 S8). 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 to our cluster representatives using ANI and ST distance (Extended data, Supplemental table S13). We also used our and Gupta’s CSP proteins to predict genera (Extended data, Supplemental table S13). 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 gallinaram, 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. Additionally, seven genomes are in previously untyped clusters, thus enabling the clades to be named with a species name (Extended data, Supplemental table S4). 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 – cluster 47D). This leaves 16 genomes with a best match to a cluster representative below 95% ANI. These 16 genomes would find 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/subspecies 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 S6a) 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 S6a). 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 S1c). 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 Manhiça (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. and Teleni et al. (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. for further details.) If any other bacteria were detected during smear microscopy examination for AFB, the cultures were subject to a second round of NACL-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. Supernant 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 3.1.1, which produced good quality assemblies as detailed in supplementary materials (Extended data, Supplemental Table S1c).

Genome annotation
The NTM genomes sequenced for this project were annotated by NCBI’s prokaryotic annotation pipeline, PGAP2.

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 Mycobacteriaeae strains excluded from RefSeq for either being untrustworthy as type or for no given reason and 11 Cornybacteriaeae 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 minHashing of kmer overlap\textsuperscript{16}. Pairwise comparisons of all the genomes was run with 10,000 21-mers per genome. We used the GGrasSP R package v1.0 which generated a complete linkage tree from the distance matrix calculated by subtracting 100 from the MASH ANI results\textsuperscript{31}. 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 GGrasSP 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 (\textit{N. nova} SAMN02900753, cluster 190) as an outgroup\textsuperscript{31,36}. 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\textsuperscript{31,36}. For four of the genera, the Least-Common Ancestor for all type strains designated by Gupta \textit{et al.} (found either within the text or in Tables 8–11) was identified\textsuperscript{31}. Since the genus \textit{Mycobacterium} only had a single designated 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 \textit{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\textsuperscript{35,36} 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 v3.0\textsuperscript{40}. 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 S4\textsuperscript{41}) with the command line typerpl --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/JCInstitute, including: JCVI pan-genome pipeline (PangenomePipeline), LOCUST, and GGrasSP.

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 EMBASS suite. Protein identifications for all of the Gupta \textit{et al.} genes were collected from tables 2, 5, and 7\textsuperscript{31} 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.

This BioProject contains all clinical NTM sample sequences; Individual biosample IDs of the genomes can be found in Extended data, Supplemental table 1\textsuperscript{41}. MLST allele sequences and corresponding schema is available at the PubMLST \textit{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 7\textsuperscript{43}. 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.v110.

Figshare: Supplemental Tables 2 & 3: NTM Genomes MASH ANI-based Clusters. https://doi.org/10.6084/m9.figshare.12024051.v112.

Figshare: Supplemental Table 4: NTM Clusters. https://doi.org/10.6084/m9.figshare.12024057.v114.

Figshare: Supplemental Table 5. NTM Genomes. https://doi.org/10.6084/m9.figshare.12024117.v117.

Figshare: Supplemental File 6. Pangenome Matchtable. https://doi.org/10.6084/m9.figshare.12024126.v119.

Figshare: Supplemental Table 7. NTM ST Definitions. https://doi.org/10.6084/m9.figshare.12024144.v121.

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.v221.

Figshare: Supplemental Table 10: NTM Pangenome-Based ANI All Versus All Table. https://doi.org/10.6084/m9.figshare.12024024.v234.

Figshare: Supplemental Table 11: NTM Genome MASH ANI Distances. https://doi.org/10.6084/m9.figshare.12024030.v236.

Figshare: Supplemental Table 12. NTM All versus All ST-based Distance Matrix. https://doi.org/10.6084/m9.figshare.12024303.v238.

Figshare: Supplemental Table 13. Analysis of Genomes published in Matsumoto et al. 2019. https://doi.org/10.6084/m9.figshare.12024324.v241.

Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (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 Beeri, 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): 1267–12.
   Published 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.
   Published 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): 319–34.
   Published Abstract | Publisher Full Text | Free Full Text

4. Tortoli E: The new mycobacteria: an update. FEMS Immunol Med Microbiol. 2006; 48(2): 159–78.
   Published Abstract | Publisher Full Text

5. Wolinsky E: Nontuberculous mycobacteria and associated diseases. Am Rev Respir Dis. 1979; 119(1): 107–59.
   Published Abstract

6. Wayne LG, Sneath HA: Agents of newly recognized or infrequently encountered mycobacterial diseases. Clin Microbiol Rev. 1992; 5(1): 1–25.
   Published 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.
   Published Abstract | Publisher Full Text | Free Full Text

8. Poyntz HC, Stiavano E, Griffiths KL, et al.: Non-tuberculous mycobacteria have diverse effects on BCG efficacy against Mycobacterium tuberculosis. Tuberculosis (Edinb). 2014; 94(3): 226–37.
   Published Abstract | Publisher Full Text | Free Full Text

9. Rowland R, McShane H: Tuberculosis vaccines in clinical trials. Expert Rev Vaccines. 2011; 10(5): 645–64.
   Published 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.
    Published 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): e515.
    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.
    Published Abstract | Publisher Full Text

13. Gupta RS, Lo B, Sinn 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.
    Published Abstract | Publisher Full Text | Free Full Text

14. Ezuzicky JP: List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. Int J Syst Bacteriol. 1997; 47(2): 590–2.
    Published 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.
    Published Abstract | Publisher Full Text | Free Full Text

16. Clarke Thomas: 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.
    Published 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.
    Published Abstract | Publisher Full Text | Free Full Text

19. Fedrizzi T, Meehan CJ, Gottora A, et al.: Genomic characterization of Nontuberculous Mycobacteria. Sci Rep. 2017; 7: 42528.
    Published 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. canariasense, M. brisbanense, and M. novocastrense. Genome Announc. 2016; 4(3): pii: e00322-16.
    Published Abstract | Publisher Full Text | Free Full Text

21. Sutton GG, Brinkac LM, Clarke TH, et al.: Enterobacter hormaechei subsp.
of the genus Acinetobacter. BMC Microbiol 2012; 12(1): 302.
Published Abstract | Publisher Full Text | Free 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.
Published Abstract | Publisher Full Text | Free Full Text

24. Chan JZ, Halachev MR, Loman NJ, et al.: Defining bacterial species in the genomic insights from the genus Acinetobacter. BMC Microbiol 2012; 12(1): 303.
Published Abstract | Publisher Full Text | Free Full Text

25. Gons J, Konstantinidis KT, Klapperbach 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.
Published Abstract | Publisher Full Text | Free 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(43): 19126–31.
Published 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): 75–88.
Published Abstract | Publisher Full Text | Free Full Text

28. Varghese NJ, Mukherjee S, Ivanova N, et al.: Microbial species delineation using whole genome sequence data. Nucleic Acids Res. 2015; 43(14): 6761–71.
Published Abstract | Publisher Full Text | Free Full Text

29. Colston SM, Fullmer MS, Bekal E, et al.: Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using Aeromonas as a test case. mBio. 2014; 5(6): e02136.
Published Abstract | Publisher Full Text | Free Full Text

30. Meier-Kolthoff JP, Aue AF, Klenk HP, et al.: Genome sequence-based species delineation with confidence intervals and improved distance functions. BMC Bioinformatics. 2013; 14(1): 60.
Published Abstract | Publisher Full Text | Free Full Text

31. Odnov BD, Treangen TJ, Meklet P, et al.: Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. 2016; 17(1): 132.
Published Abstract | Publisher Full Text | Free Full Text

32. Clarke Thomas: Supplemental Tables 2 & 3: NTM Genomes MASH ANI-based Clusters. figshare. Dataset. 2020.

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.
Published Abstract | Publisher Full Text | Free Full Text

34. Clarke Thomas: Supplemental Table 4: NTM Clusters. figshare. Dataset. 2020.

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.
Published 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. 2015; 31(5): 1049–50.
Published Abstract | Publisher Full Text | Free Full Text

37. Clarke Thomas: Supplemental Table 5: NTM Genomes. figshare. Dataset. 2020.

38. Torbø 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.
Published Abstract | Publisher Full Text

41. Jolley KA, Bray JE, Maiden MCJ: Open-access bacterial population genomics: BIGBis software, the PubMLST.org website and their applications. Welcome Open Res. 2018 [cited 2020 Jan 13]; 3: 124. PubMed Abstract | Publisher Full Text | Free Full Text

42. Clarke Thomas: Supplemental Table 6: PanGenome Matchtable. figshare. Dataset. 2020.

43. Clarke Thomas: Supplemental Table 7. NTM ST Definitions. figshare. Dataset. 2020.

44. Clarke Thomas: Supplemental Table 8 and 9. Analysis of Gupta et al 2018 and Pangenome-Derived Conserved Signature Proteins in NTM Genomes. figshare. Dataset. 2020.

45. Matsumoto Y, Kinjo T, Morotoka D, et al.: Comprehensive subspecies identification of 175 nontuberculous mycobacteria species based on 7547 genomic profiles. Emerg Microbes Infect. 2019; 8(1):1043–53.
Published Abstract | Publisher Full Text | Free Full Text

46. Clarke Thomas: Supplemental Tables 10 and 11: NTM Genome MASH ANI Distances. figshare. Dataset. 2020.

47. Clarke Thomas: Supplemental Table 12: Analysis of Genomes published in Matsumoto et al 2019. figshare. Dataset. 2020.

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.

49. Teleni 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(4): 175–8.
Published Abstract | Free Full Text

50. Hatherill M, Hawkebridge T, Whitelaw A, et al.: Isolation of non-tuberculosis mycobacteria in children investigated for pulmonary tuberculosis. PLoS One. 2006; 1(1): e21.
Published Abstract | Publisher Full Text | Free Full Text

51. Nurr S, Bankewich A, Antipov D, et al.: Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol. 2013; 20(10): 714–37.
Published 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 assembling single-cell genomes and mini-metagenomes from chimeric MDA products. J Comput Biol. 2013; 20(10): 714–37.
Published Abstract | Publisher Full Text | Free Full Text

54. 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

55. Clarke Thomas: Supplemental Table 10: NTM Pangenome-Based ANI All Versus All Table. figshare. Dataset. 2020.

56. Clarke Thomas: Supplemental Table 12. NTM All versus All ST-based Distance Matrix. figshare. Dataset. 2020.
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 then 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 Mycobactericeae 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: 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”.

- 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?
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 then 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
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 can 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, 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. 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 Mycobactericeae 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
"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.