DOI: 10.5203/pmuser.201841637

Research Article

Developing a Genotyping Scheme for *Mycobacterium abscessus* Complex Using Whole Genome Sequencing Data

Michelle Wuzinski[1], Meenu K. Sharma[2,3]

[1]Dept. of Microbiology, University of Manitoba, Winnipeg, MB, R3T 2N2
[2]Dept. of Medical Microbiology, University of Manitoba, Winnipeg, MB
[3]National Reference Centre for Mycobacteriology, National Microbiology Laboratory, Winnipeg, MB, R3E 3R4

Corresponding Author: M. Wuzinski (wuzinskm@myumanitoba.ca)

Abstract

Mycobacterium abscessus *complex is a rapid growing non-tuberculous mycobacteria (NTM) and a clinically significant pathogen capable of causing varying infections in humans. It is notoriously difficult to treat due to its inducible resistant state to clarithromycin and intrinsic resistant states to other drugs including rifampicin. Typing schemes for bacterial pathogens provide numerous applications including sourcing an outbreak, identifying cross contamination, chain of transmission and surveillance. However, they either lack or are limited for many NTMs including M. abscessus complex. The current publically available scheme on Pubmlst has not been updated in several years and was only able to apply a sequence type to less than half of tested isolates. This project was aimed at creating a workflow for the development of a multi locus sequence typing (*mlst*) scheme using whole genome data. A total of 104 genomes and 14 loci were used to build the scheme (including 3 type strains of each of the 3 subspecies). All 7 genes from the Pubmlst scheme were incorporated namely,* argH, cya, gnd, murC, pta, purH, *and* rpoB *and were expanded by 6, 9, 12, 8, 12, 10, and 8 alleles, respectively. Another 7 novel genes were added including* hsp65, erm(41), arr, rrs, rrl, gyrA, *and* gyrB *with 9, 14, 20, 7, 25, 24, 22 alleles, respectively, with 62 unique sequence types were identified among all isolates. This scheme can also differentiate* M. abscessus *complex to the subspecies level on the basis of 3 discriminatory genes and includes 6 genes related to antimicrobial resistance.*

Keywords: Whole Genome Sequencing, MLST, Mycobacterium abscessus Complex, NTM

1 Introduction

Mycobacterium abscessus is a rapidly-growing mycobacteria, commonly found in soil and water that is becoming a growing clinical concern[1]. It can cause a range of infections from pulmonary to soft tissue infections[2]. More concerning is its intrinsic resistant state to a large number of antimicrobials, making it one of the most resistant pathogenic rapidly growing mycobacteria (RGM)[3]. A particular challenge with M. abscessus is its controversial nomenclature. *M. abscessus* (also known as *M. abscessus* complex or *M. abscessus sensu lato*) has recently been divided into three subspecies, *M. abscessus* subspecies *abscessus*, *M. abscessus* subspecies *massiliense*, and *M. abscessus* subspecies *bolletii* (for simplicity each subspecies will be referred to as *M. abscessus*, *M. massiliense* and *M. bolletii*, respectively). Until 1992, *M. abscessus* was classified under the *M. chelonae* group and it was not until 2013 that it was split into the three subspecies that are now most commonly used[2, 4, 5].

Typically, mycobacteria are differentiated by their 16S rRNA sequences as it is well conserved within the genus[6]. However, this is not sufficient to differentiate the subspecies as they have been found to have identical 16S sequences[7]. While the subspecies are typically differentiated by PCR amplification and sequencing of the hsp65 or rpoB genes, with the growing trend of whole genome sequencing, it has become easier and less expensive to use whole genome data than ever before. A similar genotyping scheme reported that to obtain all the data for their analysis it both less expensive and less time consuming to use whole genome sequencing versus traditional Sanger sequencing. Bartels[8] Multi-locus sequence typing (MLST) has been a staple in molecular biology for the past 20 years. Its usefulness extends from an epidemiological tool to pathogenicity, evolution and surveillance[9].

Traditionally, MLST schemes are composed of seven house-keeping genes of 450–500 base pairs in length and each gene would be sequenced and analyzed (MLST-home, mlst.net) using an algorithm to assign allele numbers for each gene. However, with access to whole genome sequencing, this method is becoming outdated and whole genome data should be utilized in every way possible. It becomes much easier to use larger and a greater number of genes because individual gene sequencing is no longer required. With whole

Table 1: *Expansion of alleles from previously established loci as last updated by curator S. Kim on 2012/08/06, accessed 2018/04/10*

Alleles	argH	cya	gnd	murC	pta	purH	rpoB
Previous	6	5	7	8	11	7	4
New	6	9	12	8	12	10	8
Total	12	14	19	16	23	17	12

genome sequencing becoming more and more prevalent due to lowering costs, novel techniques are needed to be able to process and utilize more of the data and information that it brings. The publically available MLST database for *Mycobacterium abscessus* complex is split into two schemes, *M. abscessus* and *M. massiliense*, this paper and scheme is a modified and updated version of the *M. abscessus* scheme developed by Kim[10].

This project was aimed at creating a workflow for the development of a multi locus sequence typing (MLST) scheme using whole genome data.

2 Materials and Methods

2.1 Genome Collection

An original set of 16 genomes (subset 1; Appendix 1) were downloaded from the European Nucleotide Archive in fastq format. They were then uploaded to Galaxy[11] and run through the MentaLiST algorithm[12], an MLST allele caller that uses the publically available scheme on pubmlst.net as imported on the Galaxy and fastq formatted genomes. Two other MLST algorithms were tested for use, the first being the Centre for Genomic Epidemiology (CGE) and Stringmlst[13]. CGE was rejected due to taking over a day to type a single sample and often failed. Ultimately, MentaLiST was chosen over Stringmlst as it was faster and user friendly.

Reference genomes were obtained from ncbi for the type strains of the three subspecies, *M. abscessus* subspp. *abscessus*, *massiliense*, and *bolletii* under the accession numbers ATCC19977, CCUG48898 and CCUG50184, respectively. The genome size of each is roughly 5 Mb. An additional set of 97 genomes (subset 2; Appendix 2) (accessed from ncbi under the accession SRP127025) and 75 genomes (subset 3; Appendix 3) from the study ERP001039 were downloaded. All 198 genomes from the three subsets were subjected to quality filters on the following parameters: successfully be typed by MentaLiST, successfully assemble, have a coverage greater than 5 and at least 80% mapped to the reference strain. From the 198, 104 passed all parameters and were used to create the scheme.

2.1.1 Genome Assembly

The first 16 test genomes were assembled in irida (irida.ca) with default parameters. For simplicity and ease, the remaining genomes were assembled in Public Health Agency of Canada's Galaxy (developed by Bioinformatics core) using the SPAdes pipeline which provided the output of average coverages used as a parameter of greater than 5. While this coverage is generally considered low, due to the quality of available data it was decided to be sufficient in order to have a workable dataset. Sequencing reports were also run in Galaxy to ensure all were greater than 80% mapped to the reference of ATCC19977.

2.1.2 MLST Scheme Development

All 104 genomes were then run through a customized R script (known as "MasterBlaster") in R Studio (R Studio Inc., Boston, USA) initially developed in-house by Walter Demczuk (National Microbiology Lab, Winnipeg, MB) for genotyping of enteric pathogens. The script takes a user imported wild type gene and utilizes blast[14] to query a single genome (or several) against the wild type and identifies a match or calls it as not found and the user can input it as a novel allele. This was done for all seven genes of the previously established loci in the M. abscessus MLST scheme (pubmlst.com) as well *hsp65*, *erm(41)*, *rrs*, *gyrA*, *gyrB*, *rrl*, and *arr*. We developed an allele list specific for MAB incorporating essential genes for both identification genes and antimicrobial resistance genes. These genes are relevant in identification, clarithromycin resistance, aminoglycoside resistance, fluoroquinolone resistance (*gyrA* and *gyrB*), clarithromycin resistance and rifampicin resistance, respectively.

Wild type genes for the novel genes were arbitararily obtained from the genome of the ATCC19977 strain of *M. abscessus* (accession: NC_010397). The exception to this is the erm(41) where a partial sequence for the T28 sequevar was used as the wild type from accession HQ127365, the C28 sequevar was obtained from HQ127366, and the *M. massiliense* and *M. bolletii* alleles came from their respective type strains. All sequences from the M. abscessus database on Pubmlst were merged with the developing database with allele 1 from each respective loci becoming the 'wild type'.

The first part of the script (MasterBlaster) is known as the development stage. A wild type gene was selected (and became allele 1) and all genomes were blasted against this sequence. The script reports results for each genome on the basis of presence or absence of the gene (POS or NEG), if positive it proceeded to whether the allele matched one already in the database and it gave the allele number or reported it as not found (NF). If NF, the sequences were then opened in a sequence viewer, AliView v. 1.23[15], an alignment viewer and editor that can be downloaded for free from the internet.

Table 2: *Allelic characteristics of novel genes added to the scheme.*

	hsp65	erm(41)	arr	rrs	rrl	gyrA	gyrB
Alleles	9[1]	14	20	7	25	24	22
Length (bp)	424	360[2]	426	1504	3112	2520	2025

[1] An allele for *M. chelonae* was added in order to differentiate it from M. *abscessus* but not included in calculations in Table 3.

[2] The length for the allele correlating to *M. massiliense* is 283 base pairs.

Table 3: *Sequence diversity of each locus. Full similarity matrices of all alleles can be found in Appendix 4. erm(41) was excluded due to the large deletion in M. massiliense that skews results.*

Locus	Average Sequence Divergence	Average Percent Identity
argH	2.33	97.7
cya	1.82	98.2
gnd	2.58	97.5
murC	2.28	97.8
pta	1.47	98.6
purH	1.48	98.5
rpoB	2.67	97.4
hsp	3.33	96.8
arr	1.72	98.2
rrs	0.162	99.8
rrl	0.165	99.8
gyrA	1.25	98.8
gyrB	1.56	98.4

Table 4: *Allelic assignments for hsp65 with associated subspecies name and mutations. Mutation numbering based upon the M. abscessus type strain (ATCC19977)*

Allele #	Subspecies	Associated Mutations
1	*M.abscessus*	Wild type
2	*M.massiliense*	Wild type
3	*M. bolletii*	Wild type
4	*M. massiliense*	T293A
5	*M. bolletii*	T200C
6	*M. abscessus*	C200T
7	*M. abscesus*	C299A
8	*M. bolletii*	C173T and T200C
9	*M. chelonae*	Wild type

From this point, any duplicate sequences of an individual gene were identified in AliView and removed, leaving only unique alleles which were arbitrarily assigned a number and then used to build the allelic database for each individual loci. When the individual gene databases were completed, sequence types (STs) are defined by running all genomes with all loci to generate allele profiles through the MasterBlaster script. These allele profiles were unique to each of the 62 STs, in other words, each had different combinations of alleles at each locus. These were exported into Microsoft Excel and the "Remove Duplicates" function was used to filter out any of the 62 STs that may have been repeated among all isolates. After the definition of STs, the second and separate script, known as the MLST script, was used. This script reported allele numbers for a given isolate at each loci and reports the ST that corresponds with a given allele profile.

As it currently functions, the MLST script requires that alleles for a given gene must all be the same length in order for the algorithm to identify the sequence, due to this; multiple genes had to be trimmed. The last base pair of argH was deleted and the last codon from *gnd* and *gyrB* were removed. However, *erm(41)* posed a greater issue. Because *M. massiliense* has a truncated gene, it is significantly shorter than the others; when all alleles were trimmed to match the length of *M. massiliense*, multiple sequences were flagged as duplicates because their variation existed after the cutoff, thus introducing errors as these sequence were not actually duplicates. When samples are run through the MLST script, shorter sequences will be given an X (meaning no gene was found) and they can then be taken back to the MasterBlaster script and identified as containing the *M. massiliense* allele for *erm(41)*. When an allele was found that did not match one in the database, it was flagged with a question mark. This genome would then be taken back to the development stage (MasterBlaster script) and generate a new allele if applicable. This step is highly quality controlled.

Following completion of the scheme and establishment of sequence types, two clinical strains were extracted by a colleague using the InstaGene protocol (Bio-Rad, Hercules, California, USA) and assembled in IRIDA. The two samples, 1800282 and 1800298, were used to test the final scheme in the MLST phase and ensure it was functioning as expected.

3 Results

The original *M. abscessus* MLST scheme that is publically available on Pubmlst is comprised of 7 genes (*argH*, *cya*, *gnd*, *murC*, *pta*, *purH*, and *rpoB*) which were compiled into 26

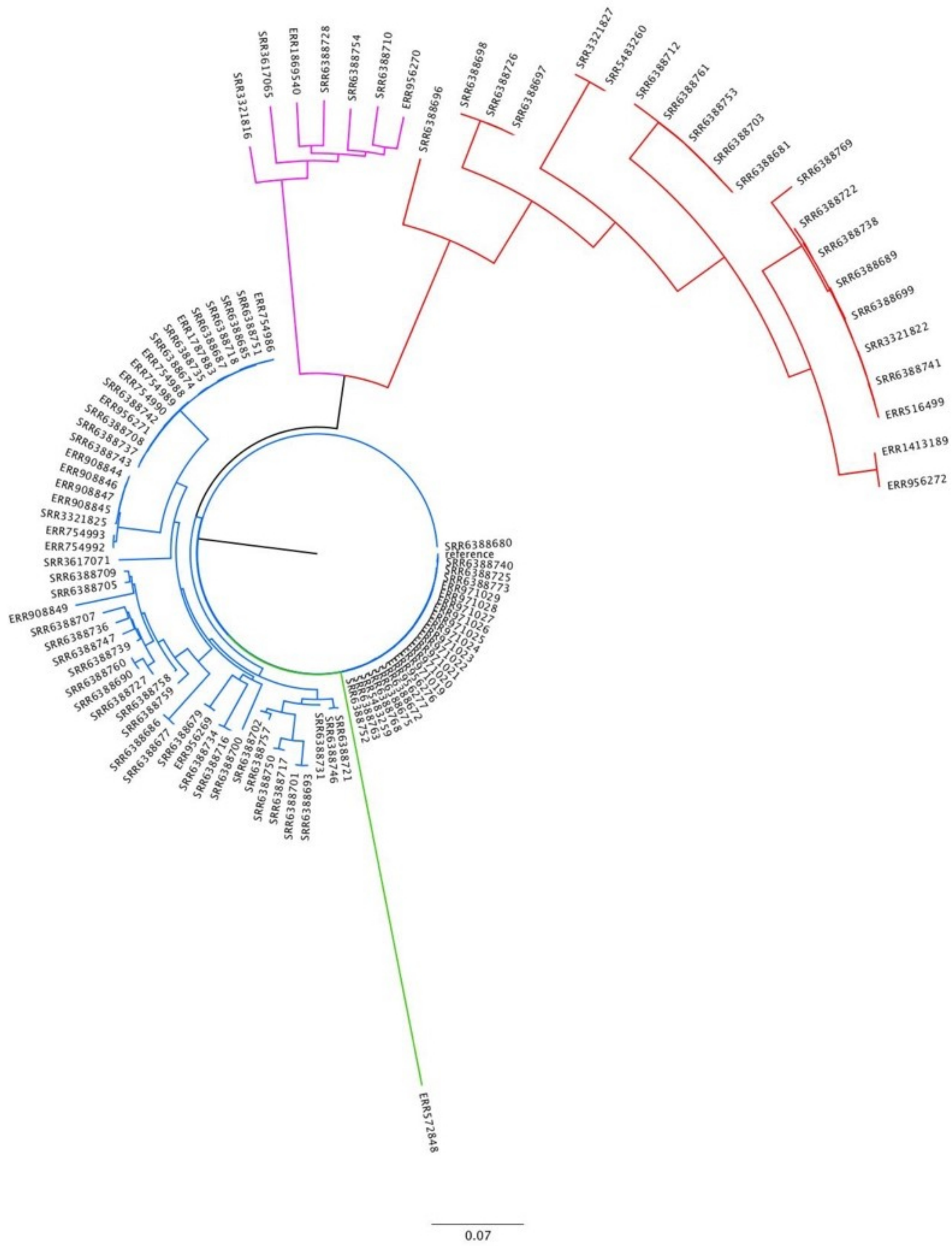

Figure 1: *Phylogenetic tree of all query genomes based on single nucleotide polymorphism analysis. Created as output of SNVPhyl pipeline (from Galaxy) analysis and visualized with FigTree v.1.3.4. Magenta branches represent M. bolletii, red is M. massiliense, blue is M. abscessus, and green is M. chelonae. Run with Galaxy's default parameters except the minimum coverage which was set to five.*

sequence types. A total of 7 new genes were added, all with capability to differentiate between the subspecies of *M. abscessus* complex or a gene involved in antimicrobial resistance. These novel genes are: *hsp65*, *erm(41)*, *gyrA*, *gyrB*, *rrs*, *rrl*, and *arr*. Similarity matrices for all alleles of all genes can be found in Appendix 4.

Table 5: *Characteristics of alleles used for erm(41).*

Allele #	Subspecies	Associated Mutations	Relevant Sequevar	Predicted Phenotype
1	*M. abscessus*	Wild Type	T28	Resistant
2	*M. abscessus*	Wild Type	C28	Susceptible
3	*M. massiliense*	Truncated/Wild Type	NA	Susceptible
4	*M. bolletii*	Wild Type	T28	Resistant
5	*M. abscessus*	G95T/T285C	T28	Resistant
6	*M. abscessus*	A246G/T285C	T28	Resistant
7	*M. bolletii*	G63A, C285T	T28	Resistant
8	*M. abscessus*	T285C	T28	Resistant
9	*M. bolletii*	A63G, C77T, T285C, A289G, T357C	T28	Resistant
10	*M. abscessus*	T198C, C285T	T28	Resistant
11	*M. bolletii*	A63G, T285C, T357C	T28	Resistant
12	*M. abscessus*	G294C	T28	Resistant
13	*M. bolletii*	A63G, A96T, C259T, T285C, A297G	T28	Resistant
14	*M. abscessus*	T154C, T285C	C28	Susceptible

Mutation numbering based on M. abscessus type strain (ATCC19977) as reference sequence. Due to a partial sequence being used, position 28 is 154 in this scheme.

3.1 MentaLiST Results

Using the MentaLiST algorithm, samples were first typed by the *M. abscessus* scheme from Pubmlst. Fastq files were not obtained for *M. massiliense* or *M. bolletii* type strains and thus did not undergo this typing. Of 102 samples, 45 were identified as a defined sequence type (ST), 23 belonging to ST5, 15 belonging to ST9 and 2 belonging to ST24. That leaves 56% of samples not matching a ST. While this can partially be explained by the fact that subspecies other than M. abscessus were included, only 26% of samples were identified by *hsp65* analysis as not belonging to the *M. abscessus* subspecies so many samples still fall outside of the defined sequence types.

Based on the correlation of sequences of their *rpoB*, *hsp*, and *erm* genes, 72 samples were found to be *M. abscessus*, 8 as *M. bolletii*, and 19 as *M. massiliense*. Three samples produced discrepant results (discussed below; Table 7) and one sample was identified as *M. chelonae* based on its *hsp65* and 16s RNA sequences (ERR572848). These identities align as expected with the clustering of the phylogenetic tree below (Fig. 1) where *M. abscessus* type strain was used as a reference.

3.2 Novel MLST Results

A total of 62 unique sequence types (STs) were identified among all samples with 41 duplicated STs. The STs were completely redefined in accordance with all loci and in no way relate to those on the published Pubmlst scheme. STs 1, 2 and 3 correspond to the type strains for *M. abscessus*, *M. bolletii* and *M. massiliense*, respectively. All remaining ST numbers were arbitrarily assigned based on unique allele profiles. However, at this point in time, due to the truncation in *erm(41)*, no *M. massiliense* STs can be automatically identified by the MLST script.

By increasing the amount of unique allele profiles from 26 to 62, the discriminatory power is increased by 4.2x. The amount by which each gene was added to in terms of alleles is displayed in Tables 1 and 2. Additionally, the allelic diversity of each gene is illustrated in Table 3. On the other hand, because there is double the number of genes involved in the scheme, it will become less likely for an isolate to have a 100% match to a given ST.

A total of 16 samples failed to be applied a specific ST. Of these, 24 individual errors within the 16 samples were identified. The reason for this is unclear because an allele was found for all genes and all genomes within the MasterBlaster script and a ST was assigned for all isolates in the scheme. Problems arose in alleles: *rrs24*, *rrs25*, *rrs1*, *rrs3*, *rrs7*, *purH12*, *pta16*, and *arr18*. All of these alleles with the exceptions of *arr18* and *purH12* were properly identified in other genomes. A possible explanation would be the assembly quality of the genomes as the parameters were set fairly

Table 6: *Allele information for rpoB.*

Allele #	Subspecies	Attributes
1	*M. abscessus*	Pubmlst original\Wild Type
2	*M. abscessus*	Pubmlst original
3	*M. abscessus*	Pubmlst original
4	*M. abscessus*	Pubmlst original
5	*M. bolletii*	Wild Type
6	*M. massiliense*	Wild Type
7	*M. massiliense*	
8	*M. massiliense*	
9	*M. massiliense*	
10	*M. massiliense*	
11	*M. massiliense*	

Table 7: *Gene Identities of Discrepant Genomes*

Accession #	rpoB Identity	hsp65 Identity	erm41 Identity
SRR5483260	*M. abscessus*	*M. massiliense*	*M. massiliense*
SRR3321827	*M. abscessus*	*M. massiliense*	*M. massiliense*
ERR908849	*M. abscessus*	*M. bolletii*	Non type strain

low due to the quality of available data. It is speculated that the whole genome sequencing (WGS) was missing data for some of these genes.

After competition of the scheme and establishment of STs, two clinical samples were analyzed as a test. While neither matched a previously defined ST, the first sample, 1800282, had 100% matches for 13/14 genes but lacked an exact match for *gyrB*. The second sample, 1800298 had 100% matches to 8/14 alleles including *hsp*, *erm*, *rrl*, *rrs*, *argH*, *gnd*, *murC*, and *purH*. Two new STs were added for these isolates.

3.3 Discrepant Genomes

Three samples presented with discordant results with differing identities at several identification loci, as seen in Table 7. According to the recommendations by Griffith[16] *M. massiliense* is that of the organism with a large deletion in the *erm(41)* gene making it non-functional, thus both SRR5483260 and SRR3321827 should be identified as *M. massiliense*. Furthermore, Macheras[17], isolated a strain that was identified as *M. abscessus* based on *rpoB* but *M. massiliense* on the basis of hsp65 which highlights the necessity of using multiple genes to identify the subspecies. Evidence of horizontal transfer of the rpoB gene has been reported and could explain this discrepancy[18]. Griffith[16], also recommends that *M. bolletii* be that which differs from *M. abscessus* and *M. massiliense* based on its *rpoB* sequence and has a functional *erm(41)*. On closer analysis of ERR908849's *erm(41)* sequence, it is 4 SNPs (single nucleotide polymorphism) away from the *M. bolletii* type strain but only 2 SNPs from the *M. abscessus* type strain (one being position 28 making it a C28 sequevar). When looking at the phylogenetic tree (Fig. 1), it clusters with *M. abscessus* but significantly further away than other samples. Based on this, a clear identity cannot be assigned to ERR908849.

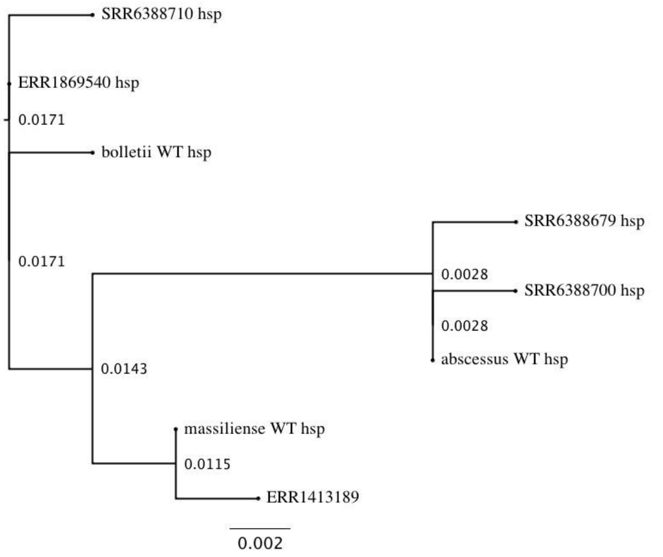

Figure 2: *Dendogram of all new alleles added to hsp65. Made with FastTree (Galaxy) and visualized with FigTree v.1.4.3.*

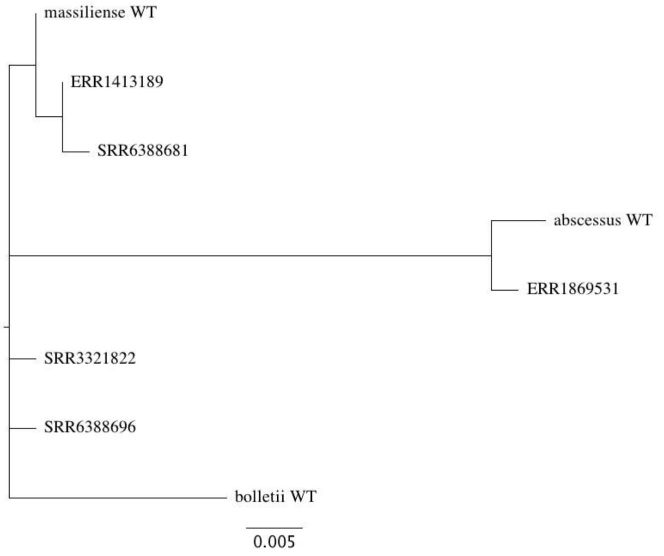

Figure 3: *Dendogram of sequences of new alleles added to rpoB. Made with fasttree, visualized with FigTree (v.1.4.3).*

4 Discussion

A major advantage of MLST is the fact that it does not account for how many nucleotide differences there are between alleles so point mutations and recombination are given the same weight in terms of diversity. Additionally, MLST has a high discriminatory power due to the rarity of an isolate exactly matching a sequence type[19].

This modified MLST that was developed takes advantage of some of the information that whole genome sequencing has been able to provide and is a simple and easy-to-use way to build an MLST scheme. A single locus could be added in a matter of minutes as all that is required is a wild type copy of the gene. However, it has flaws associated with it. First, any sequence with gaps or shortened for any reason, such as a single base pair deletion, cannot be properly recognized by the MLST script. While they can be identified with no issue in MasterBlaster, errors arise when they are brought into the MLST script as it looks for an exact user-specified length. Additionally, issue arose with the blast program downloaded from ncbi. Genes had to be trimmed because the blast output would be missing a base pair or two or have extra base pairs. The technical glitch identified was sample SRR6388768 originally identified as allele 18 for *rrl* but MLST was unable to provide an output.

For a select few genes, some extra information was encoded to be outputted when the script was run. Genes commonly used for identification (namely, *hsp65*, *erm(41)*, and *rpoB*) will display the associated subspecies name and in addition, *erm(41)* will output the relevant sequevar (elaborated on below). However, this output only exists in the MasterBlaster script as the MLST only reports the allele number of a given locus and any additional output is not shown. A solution to this however, is knowing what allele number corresponds with what genotype.

4.1 hsp65

Hsp65 sequencing is a common method for differentiating not only the subspecies of the *M. abscessus* complex but also from the closely related *Mycobacterium chelonae*[20]. This gene was not included in the original MLST scheme and thus was the first one added in order to give the scheme a higher discriminatory power. By using the hsp65 sequences for each reference genome of the respective subspecies, the scheme now has the power to subspeciate any *M. abscessus* genome it is given. 89% of samples matched the allele from one of the given type strains. The numerical allele assignments can be found in Table 4.

There were two samples, ERR1869540 and ERR1413189, that did not have a 100% match to one of the reference genome alleles and based on their phylogenetic clustering of the *hsp65* gene (Fig. 2), they were identified as *M. bolletii* and *M. massiliense*, respectively. Additionally, when these sequences were blasted online against the National Centre for Biotechnology Information (NCBI) database, they both obtained multiple 100% matches to other strains, implying that these alleles were not errors in sequencing but true differentiation. When subset two was applied, three new alleles were found; their subspecies identification was confirmed in two ways. First, the tree was analyzed and the identity found was confirmed with the *rpoB* identity. SRR6388679 and SRR6388700 clustered with the *M. abscessus* wild type and had the matching *rpoB* sequence, each with one mutation from the wild type (C200T and C299A, respectively). SRR6388710 clustered with the *M. bolletii* wild type *hsp65* and contained the *M. bolletii rpoB* sequence but had two mutations: C173T and T200C.

4.2 erm(41)

The *erm(41)* gene is important for *M. abscessus* for its ability to induce resistance to macrolides (namely clarithromycin) as well as differentiating the three subspecies[21]. They can be differentiated based on the presence or absence of a functional *erm(41)* gene. *M. massiliense* has a large deletion in this region which produces a truncated and non-functional erm protein. On the other hand, both *M. abscessus* and *M. bolletii* have functional erm genes but have unique sequences. Perhaps most importantly, a functional *erm(41)* confers inducible resistance meaning that after the standard 3 day incubation period in minimum inhibitory concentration testing, isolates may appear susceptible to clarithromycin and other macrolides. However, in an incubation period of up to 14 days, induction of the gene can occur and generate resistance to the drug[16]. Furthermore, some *M. abscessus* complex isolates have a T to C mutation at position 28 (position 154 in this scheme) which also demonstrates a susceptible macrolide phenotype. All of these factors are included and recognized by the scheme. While no 100% matches to the wild-type C28 sequevar were found, the allele is present in the scheme. However, one C28 sequevar was found but was a single base pair away from the wild type of C28 (2 base pairs from the T28 wild type) and was identified in 10 different samples. All alleles included in the scheme can be found in Table 5.

4.3 rpoB

Encoding for the beta subunit of RNA polymerase, *rpoB* was included in the previously established scheme and is involved with rifampicin resistance[22]. However, because of its use in subspecies identification[18], it requires special attention and curation in this new scheme. Firstly, alleles for *M. massiliense* and *M. bolletii* from their respective type strains were added, and then 5 other new alleles were found from data subsets 1 and 2. These 5 genomes that lacked a 100% match to an already established allele or an allele from a type strain were both blasted against the ncbi database and their phylogeny was compared against the type strains in order to assign a subspecies (Fig. 3). Furthermore, their identities were compared with the calls made by the other useful identification genes, *erm(41)* and *hsp65*. The alleles can be found in Table 6.

4.4 rrl

23s rRNA is encoded for by the *rrl* gene and, along with *erm(41)*, is involved in clarithromycin resistance[21]. While *erm(41)* is responsible for inducible resistance, both genes are important factors when looking at clarithromycin resistance. While 25 different alleles were identified among the samples, most differed by only a SNP or two from one another with only 25 variant positions found in 3112 total positions. Mutations known to be associated with resistance are 2058 (A to G, C or T) and 2059 (A to G, C, or T) (E. coli numbering)[23]. All alleles showed an A at position 2058 and a G at 2059 which implies that they are likely resistant to clarithromycin. As demonstrated by Bastian[21], *rrl* mutations are sufficient to produce a resistant phenotype in M. massiliense and C28 sequevars, genotypes that are typically associated with susceptibility to clarithromycin.

4.5 rrs

The *rrs* gene encodes the 16S ribosomal RNA. The least amount of variation was found within this gene with only 7 unique alleles identified in all samples and 6 variant positions. Resistance to 2-deoxystretamine aminoglycosides have been shown to be acquired through a single point mutation at position 1408 (*E. coli* numbering) from A to G[24]. This mutation was present in all 7 alleles in the scheme implying that they are all resistant to aminoglycosides. Nessar[2], showed that two other mutations (C1409T and G1491T) that are known to cause kanamycin resistance in *M. tuberculosis* also had the same effect in *M. abscessus*; thus, these mutations should be noted and added to the scheme.

4.6 gyrA and gyrB

Fluoroquinolone resistance is typically attributed to mutations in the *gyrA* and *gyrB* genes which encode for the A and B subunits of DNA gyrase, respectively. Resistance occurs through mutations within the quinolone resistance determining region (qrdr) which is a conserved region that interacts with the drugs[25]. As shown by Monego[26], the most prominent mutations in *gyrA* that confer resistance are in amino acid positions 90 and 94 (Mycobacterium tuberculosis numbering; 83 and 87 in *E. coli*). In the B subunit, positions 495, 516, and 533 (*M. tuberculosis* numbering, 426, 447, and 464 in *E. coli*)[25]. A susceptible phenotype codes for a serine at position 90 in *GyrA* while resistant organisms have an alanine to valine substitution.[26]. These particular mutations known to confer resistance were not added into the scheme simply due to time constraints. While they were not observed in the database, they would be useful for future analysis with additional genomes. However, the diversity among these genes was moderately low (Table 3), despite there being a large number of alleles.

4.7 arr

Rifampicin is a first line drug used to treat infection caused *M. tuberculosis*. The drug's activity is derived from its ability to binds to the beta subunit of the DNA-dependent RNA polymerase thereby inhibiting enzymatic activity. In most bacteria, rifampicin resistance is attributed to mutations in the *rpoB* gene (which encodes the beta subunit of RNA polymerase) that lower the drugs affinity for the enzyme. However, in addition to this mechanism, *M. abscessus*'s genome encodes a rifampicin ADP-ribosyltransferase (MAB_0591) which gives it intrinsic resistance to the drug. As shown by Rominski[27], when introduced into naturally susceptible organisms, the *arr* gene was enough to demonstrate rifampicin resistance. This gene was found in all strains that were analyzed and identified as belong to the M. abscessus complex. Interestingly, *arr* is not found in the closely related species *Mycobacterium chelonae*[27] and as expected, the program reported the gene as not found when the reference genome for *M. chelonae* was used (accession: CCUG47445). It was also lacking in the one test sample that was identified as *M. chelonae* (ERR572848).

5 Conclusion

In conclusion, the original goal of this project was to establish a whole genome multi locus sequence typing scheme. Once the desired genes are identified and genomes collected, the process is fairly straight forward, albeit labor intensive, from that point. The most significant advantage of this method is the ability to bypass the amplification and individual gene sequencing steps that come with traditional MLST scheme creation and use. All that is required is a collection of assembled genomes and a wild type allele which can easily be obtained from an online database such as ncbi. Additionally, commercial software such as Ridom SeqSphere's MLST+ Target Definer exist to identify genes to be used in a core genome

MLST (CGMLST) scheme and was used to create a CGMLST scheme for M. tuberculosis. An additional advantage over MentaLiST is that allele calls require a 100% match in order to be called that allele whereas MentaLiST (and Stringmlst) will simply assign the closest allele which could introduce inaccurate sequence typing but this issue is eliminated in the novel scheme.

It is a known problem in the science community that analytical tools for whole genome data produce a large block of results of which not everything is useful. We not only developed a novel scheme but an established a simple method for utilizing whole genome sequencing data. The script within R Studio is very user friendly and easy to use with little bioinformatics expertise required. Even if not used as a traditional MLST scheme, even looking at a single locus can provide numerous advantages from identifying a subspecies to detecting a resistant genotype.

Modifications can still be made in order to improve the scheme such as adjusting the script to allow for differing gene lengths and displaying subspecies name as part of the ST. Additionally, mutation information and predicted phenotypes could be applied to the other antimicrobial resistance genes as they were for *erm(41)* and for the identification genes.

References

1. Lee, M.-R., Sheng, W.-H., Hung, C.-C., *et al.* 2015. *Emerging infectious diseases*, 21: 1638–46, doi:10.3201/2109.141634.
2. Nessar, R., Cambau, E., Reyrat, J. M., *et al.* 2012. *Journal of Antimicrobial Chemotherapy*, 67: 810–818, doi:10.1093/jac/dkr578.
3. Brown-Elliott, B. A., Wallace, R. J., & Jr. 2002. *Clinical Microbiology Reviews*, 15: 716–46, doi:10.1128/CMR.15.4.716-746.2002.
4. Leao, S. C., Tortoli, E., Euzeby, J. P., *et al.* 2011. *International Journal of Systematic and Evolutionary Microbiology*, 61: 2311–2313, doi:10.1099/ijs.0.023770-0.
5. Tortoli, E., Kohl, T. A., Brown-Elliott, B. A., *et al.* 2016. *International Journal of Systematic and Evolutionary Microbiology*, 66: 4471–4479, doi:10.1099/ijsem.0.001376.
6. Han, X. Y., Dé, I., & Jacobson, K. L. 2007. *American Journal of Clinical Pathology*, 128: 612–621, doi:10.1309/1KB2GKYT1BUEYLB5.
7. Zelazny, A. M., Root, J. M., Shea, Y. R., *et al.* 2009. *Journal of Clinical Microbiology*, 47: 1985–95, doi:10.1128/JCM.01688-08.
8. Bartels, M. D., Petersen, A., Worning, P., *et al.* 2014. *Journal of Clinical Microbiology*, 52: 4305–4308.
9. Pérez-Losada, M., Arenas, M., Castro-Nallar, E., *et al.* 2017. *In: Genetics and Evolution of Infectious Diseases*, Elsevier, 383–404, doi:10.1016/B978-0-12-799942-5.00016-0.
10. Kim, S. Y., Kang, Y. A., Bae, I. K., *et al.* 2013. *Diagnostic Microbiology and Infectious Disease*, 77: 143–149, doi:10.1016/J.DIAGMICROBIO.2013.06.023.
11. Afgan, E., Baker, D., van den Beek, M., *et al.* 2016. *Nucleic Acids Research*, 44: W3–W10, doi:10.1093/nar/gkw343.
12. Feijao, P., Yao, H.-T., Fornika, D., *et al.* 2018. *Microbial Genomics*, 4, doi:10.1099/mgen.0.000146.
13. Gupta, A., Jordan, I. K., & Rishishwar, L. 2017. *Bioinformatics*, 33: 119–121, doi:10.1093/bioinformatics/btw586.
14. Altschul, S. F., Gish, W., Miller, W., *et al.* 1990. *Journal of molecular biology*, 215: 403–410.
15. Larsson, A. 2014. *Bioinformatics*, 30: 3276–3278, doi:10.1093/bioinformatics/btu531.
16. Griffith, D. E., Brown-Elliott, B. A., L. Benwill, J., *et al.* 2015. *Annals of the American Thoracic Society*, 12: 436–439, doi:10.1513/AnnalsATS.201501-015OI.
17. Macheras, E., Roux, A.-L., Ripoll, F., *et al.* 2009. *Journal of Clinical Microbiology*, 47: 2596–600, doi:10.1128/JCM.00037-09.
18. Macheras, E., Roux, A.-L., Bastian, S., *et al.* 2011. *Journal of Clinical Microbiology*, 49: 491–9, doi:10.1128/JCM.01274-10.
19. Enrigt, M. C. & Spratt, B. G. 1999. *Trends in Microbiology*, 7: 482–487, doi:10.1016/S0966-842X(99)01609-1.
20. Ringuet, H., Akoua-Koffi, C., Honore, S., *et al.* 1999. *Journal of Clinical Microbiology*, 37: 852–857.
21. Bastian, S., Veziris, N., Roux, A.-L., *et al.* 2011. *Antimicrobial Agents and Chemotherapy*, 55: 775–781, doi:10.1128/AAC.00861-10.
22. Andre, E., Goeminne, L., Cabibbe, A., *et al.* 2017. *Clinical Microbiology and Infection*, 23: 167–172, doi:10.1016/j.cmi.2016.09.006.
23. Mougari, F., Bouziane, F., Crockett, F., *et al.* 2017. *Antimicrobial Agents and Chemotherapy*, 61: e00,943–16, doi:10.1128/AAC.00943-16.
24. Prammananan, T., Sander, P., Brown, B. A., *et al.* 1998. *The Journal of Infectious Diseases*, 177: 1573–81.
25. de Moura, V. C. N., da Silva, M. G., Gomes, K. M., *et al.* 2012. *Journal of Medical Microbiology*, 61: 115–125, doi:10.1099/jmm.0.034942-0.
26. Monego, F., Duarte, R. S., & Biondo, A. W. 2012. *Microbial Drug Resistance*, 18: 1–6, doi:10.1089/mdr.2011.0047.
27. Rominski, A., Roditscheff, A., Selchow, P., *et al.* 2017. *Journal of Antimicrobial Chemotherapy*, 72: 376–384, doi:10.1093/jac/dkw466.