Characterization of integrated prophages within diverse species of clinical nontuberculous mycobacteria

Cody Glickman (glickman.cody@gmail.com)
University of Colorado Anschutz Medical Campus https://orcid.org/0000-0002-6648-4832
Sara M. Kammlade
National Jewish Health
Nabeeh A. Hasan
National Jewish Health
L. Elaine Epperson
National Jewish Health
Rebecca M. Davidson
National Jewish Health
Michael Strong
University of Colorado Denver - Anschutz Medical Campus

Research

Keywords: Mycobacteriophage, NTM, Prophage, Virulence, Growth Rate

Posted Date: May 24th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-30072/v1

Version of Record: A version of this preprint was published on August 17th, 2020. See the published version at https://doi.org/10.1186/s12985-020-01394-y.

Abstract

Background Nontuberculous mycobacterial (NTM) infections are increasing in prevalence, with current estimates suggesting that over 100,000 people in the United States are affected each year. It is unclear how certain species of mycobacteria transition from environmental bacteria to clinical pathogens, or what genetic elements influence the differences in virulence among strains of the same species. A potential mechanism of genetic diversity within mycobacteria is the integration of viruses called prophages. Prophages may act as carriers of bacterial genes, with the potential of altering bacterial fitness through horizontal gene transfer. In this study, we quantify the frequency and composition of prophages within mycobacteria isolated from clinical samples and compare them against the composition of PhageDB, an environmental mycobacteriophage database.

Methods Prophages were predicted by agreement between two discovery tools, VirSorter and Phaster, and the frequencies of integrated prophages were compared by growth rate. Prophages were assigned to PhageDB lettered clusters using BLAST-p. Bacterial virulence gene frequency was calculated using a combination of the Virulence Factor Database (VFDB) and the Pathosystems Resource Integration Center virulence database (Patric-VF) within the gene annotation software Prokka. CRISPR elements were discovered using CRT. ARAGORN was used to quantify tRNAs.

Results Rapidly growing mycobacteria (RGM) were more likely to contain prophage than slowly growing mycobacteria (SGM). CRISPR elements were not associated with prophage abundance in mycobacteria. The abundance of tRNAs was enriched in SGM compared to RGM.

Introduction

NTM infections can cause serious pulmonary disease that may become chronic and affect quality of life (1). Many of the mycobacterial species that cause NTM infections are ubiquitous in the environment and are known to thrive in built environments, including premise plumbing (2–4). The mechanism by which these organisms, which have long been recognized as environmental, become clinical pathogens is an active area of research with prior studies exploring host susceptibility (5), geographic factors (6), and genomic phylogeny (7). Mycobacterial species are categorized into two broad groups based on differing growth rates in culture; rapidly growing (RGM) and slowly growing (SGM). Of the six species/subspecies examined in this study *M. avium, M. chimaera*, and *M. intracellulare* are described as having a slow growth rate, requiring more than 7 days to see colonies in culture. *M. abscessus subsp. massiliense, M. abscessus subsp. abscessus, and M. abscessus subsp. bolletii* have a rapid growth rate, requiring less than 7 days to be visible in culture (8).

Bacteriophages are double stranded DNA viruses that infect bacteria and are known to transfer genetic material between bacteria through a process called transduction (9). Mycobacteriophages are bacteriophages that target mycobacteria. Environmental bacteria are subject to external pressures to adapt their genomes through horizontal gene transfer, including bacteriophage transduction (9). Bacteriophages can exhibit both lysogenic and lytic phases during a life cycle. The lysogenic phase involves a

bacteriophage integrating genetic material into a bacterial genome and replicating in tandem until an external stimulus transitions the integrated bacteriophage, also known as a prophage, into a lytic life cycle. During the lytic phase, the bacteriophage utilizes the bacterial cellular machinery to create new phage particles that are then released during bacterial cell lysis. The newly created phage particles package bacterial genes at a low frequency, which are subsequently transduced during a new infection (10). Given the immense number of bacteriophage-bacterial interactions, transduction events are estimated to occur frequently in the environment (11).

Virulence is a general term that describes a pathogen's invasive power, ability to overcome host defenses, and the replication efficiency of a pathogen within a host (12). Bacterial susceptibility to bacteriophages is positively correlated with the overall virulence of the bacteria (13). There is a selective advantage for bacteria that contain prophages with genetic elements capable of increasing fitness and propagative success. It is possible that genes carried by mycobacteriophages during transduction events could impact virulence as seen in other bacteria such as *Vibro cholerae, Corynebacterium diphtheriae, and Streptococcus pyogenes* (14). Prophages in these serious bacterial pathogens contain elements that contribute to quorum sensing, enzymatic functions, and even extracellular toxicity (14, 15).

Here, we explore the frequency of integrated prophages in various NTM genomes and characterize the composition of bacterial genes within predicted prophage elements. Genetic elements including tRNAs act as a potential insertion site for mycobacteriophages (16). Our hypothesis is that more tRNAs are associated with an increase in the abundance of integrated prophages due to there being more targets for integration. CRISPR elements are a bacterial defense mechanism against viral integration, and CRISPR elements with fewer spacer elements are more susceptible to prophage integration (17). We hypothesize that the presence of CRISPR elements will reduce the number of integrated prophage elements.

To explore differences between clinical prophages and environmental mycobacteriophages, we utilized PhageDB, which is a data repository of mycobacteriophages isolated from the environment (18). Most of these mycobacteriophages were identified, using a phage plaque screening assay to identify lytic mycobacteriophage capable of lysing the non-pathogenic species *M. smegmatis*. Mycobacteriophages from PhageDB are organized into sequence clusters, indicated by letters (A-Z), based on sequence similarity and shared functional protein families (19). Lettered clusters typically exhibit similar lifestyle and functional behavior. Prior works have suggested PhageDB mycobacteriophages clusters N, K, G, and A are capable of infecting clinical NTM (20, 21). Our hypothesis was that prophages from clinical genomes may be enriched for bacterial virulence genes compared to environmental mycobacteriophages from PhageDB. Exploring this hypothesis will elucidate the ability of prophages to act as a genetic repository for bacterial virulence genes within clinical NTM genome.

Methods

Bacterial Genome Assembly and Isolate Datasets

In this study, we utilized two publically available datasets. First, we downloaded complete genomes from two species in the NCBI assembly database (**Supplementary File 1**), *M. abscessus* and *M. avium.* These species were selected because they represent the two most common species of clinical significance (22). All complete genomes were isolated from clinical sources with the exception of an *M. avium* isolate from a hospital water source (23). Our decision to look at complete genomes was to examine bacteriophage trends between species regardless of assembly status (complete vs. draft genome).

The second dataset includes a collection of 318 NTM draft genomes from 168 individuals diagnosed with cystic fibrosis (CF) (24). Isolates were collected in a longitudinal manner, however, only one isolate genome per patient per species was retained for prophage analysis. Fourteen patients had multiple NTM species, and in these cases, one isolate from each species was retained (n = 182 draft genomes). This dataset includes six different mycobacterial species and subspecies: *M. abscessus subsp. massiliense, M. abscessus subsp. abscessus, M. abscessus subsp. bolletii, M. avium, M. chimaera*, and *M. intracellulare* (Table 1).

Table 1
Summary statistics of prophages predicted in the NTM draft genomes. Median counts of tRNA, N50 Lengths, and counts of contigs are shown with the ranges in parenthesis. The edge case average is the total number of edge cases divided by the isolate count.

Species	Genome Count	Genomes with Phage	Predicted Prophages	Median tRNA Count	Median N50 Length	Median Contig Count	Edge Cases Average
M. avium	43	2	2	57 (54–108)	81670 (56262–116826)	206 (126–336)	1.07
M. chimera	11	1	1	55 (49–81)	86644 (78123–101816)	193 (116–251)	1.00
M. intracellulare	23	3	5	52 (50–54)	101938 (82877–131830)	104 (83–145)	0.13
M. abscessus	76	63	110	48 (45–116)	168941 (87141–327102)	67 (41–191)	2.60
M. bolletti	4	4	7	48.5 (48–49)	208710 (140476–226296)	45 (37–59)	2.50
M. massiliense	25	18	25	51 (46–84)	215865 (97542–413240)	53 (39–143)	2.84

Paired-end reads from draft genomes in the CF dataset were assembled using Unicycler into contiguous sequences (contigs), known as draft genomes, with a median N50 ranging from ~ 82 kilobases to ~ 216

kilobases (Table 1) (25). Numbers of contigs in the draft genome assemblies ranged from 37 to 336. Analysis of assembly completeness based on N50 length and the number of contigs metrics revealed three outliers, which were removed from downstream analysis. To understand if assembly fragmentation affected prophage prediction in our draft genomes, we explored the edge cases, which are defined as predicted prophages within 100 bases of either end of a contig.

Species identifications of draft genomes were determined using a method of average nucleotide identity (ANI) as described previously (24) and sequence reads mapped to active reference genomes to generate phylogenies (26, 27). Reads were mapped using Bowtie2 software (28) and single nucleotide polymorphisms (SNPs) were determined using Samtools mpileup (29). Mycobacterial genotypes were concatenated and used to make phylogenetic trees using maximum likelihood with 1000 bootstrap replicates in Randomized Axelerated Maximum Likelihood-Next Generation (RAxML) (30). SNP distances between groups in the tree file were used to perform PERMANOVA analysis in R. Phylogenetic trees were annotated and visualized with ggtree (31). Additional tree file manipulations and visualizations were performed with ETE3 (32).

ARAGORN v1.2.38 was used to quantify the tRNAs in the mycobacterium (33). CRISPR identification was performed using CRT with default parameters (minimum 3 repeats, minimum length 19 base pairs, maximum length 38 base pairs) (34).

Integrated Prophage Discovery

Prophage discovery was performed using the agreement of two prophage discovery tools, Phaster and VirSorter (35, 36). Phaster and VirSorter use sequence similarity methods against known viruses to find prophage elements within bacterial genomes. A custom application programming interface (API) script was used to identify prophages from the Phaster web server, whereas VirSorter prophage identification was performed locally. The output of Phaster contains three confidence levels for a prophage prediction with “intact” possessing the strongest support, “questionable” having some support, and “incomplete” being the least confident. VirSorter also ranks prophage predictions into three numeric levels (4 strongest, 5 middle, and 6 low support) in addition to predicting individual contigs from the draft genomes as stand-alone viruses (1 strongest, 2 middle, and 3 low support). Confidence levels of Phaster predictions were manually set to the scale of VirSorter prophages. Predicted prophages with overlapping ranges between the two tools were retained, prioritizing the predicted range with higher confidence level followed by longer length. Predicted stand-alone viruses identified in VirSorter were used to confirm overlap between tools then discarded. Edge cases incorporate all levels of predicted prophages not only those selected by both prophage prediction tools.

Dereplication of predicted prophage elements was performed using VSEARCH (37) to identify genetically identical prophages between different genomes, which may be evidence of transmission or contamination.

Prophage Identification and Genome Annotation

PhageDB uses MMSEQ2 to cluster mycobacteriophage gene product into gene “phamilies”, then uses Splitstree to create functional clusters based on presence or absence of gene “phamilies” (38). The version

of PhageDB used was filtered to only contain mycobacteriophages. Our approach to assign predicted prophages to clusters began by filtering BLASTp hits. We selected BLASTp because the gene “phamilies” are not included in the data API and the parameters of MMSEQ2 cluster are not clearly defined in prior works (19). We retained prophages that matched to at least 10 unique genes in the PhageDB gene database. A match was based on DIAMOND Blastp homology with default parameters (e-value < 1e-12, query coverage > = 70, identity cut-off > = 70) against proteins from PhageDB downloaded using the PhageDB API (39). The database containing all mycobacteriophage genes was compiled into a list of 185,629 proteins (accessed on 9/26/2019). Gene products of predicted prophages were identified using Prodigal (40). The aggregation of cluster identifications from gene hits was used to determine the projected cluster of the predicted prophage. The clustering of predicted prophages was visualized using RAWGraphs.io (41).

Bacterial genes within predicted prophages were annotated using Prokka (42) with an additional virulence factor and phage protein database combining VFDB, Patric-VF, and PhageDB proteins added as a parameter (18, 42–44). Proteins were predicted using Prodigal and subsequently annotated with a combination of DIAMOND BLASTp and HMMscan using default parameters (39, 40, 45). Pairwise significance testing comparing the abundance of virulence genes in the PhageDB and the predicted prophages was performed using a z-score test for two population proportions with binary success defined as having any bacterial virulence gene in the genome.

Prophage counts by mycobacterial species were visualized using pandas (version 0.20.3) and matplotlib (version 2.1.0) (46, 47). Pairwise significance testing comparing the abundance of prophages between species growth rate was performed using a z-score test for two population proportions with binary success defined as having a prophage with more than 10 gene matches against PhageDB in the genome.

Pangenome analyses were performed using Roary (48) to identify core genes (present in > 95% of prophages) and shell genes (present in between 15–95% of prophages) amongst all predicted prophages (n = 188). In addition, assigned PhageDB clusters of predicted prophages with more than five prophages per cluster were subjected to another pangenome analysis (n = 173).

Results

Abundance of Prophages in NTM Species

Integrated prophages were present in 91 of the 182 (50.0%) draft genomes and present in 26 of the 53 (49.1%) complete genomes. A total of 150 unique prophages were found in the 91 draft genomes and 38 unique prophages were predicted amongst the 26 complete genomes. The interquartile range of prophage lengths across all species was ~ 26 kilobases to ~ 47 kilobases (**Supp.** Figure 1**)**. We found that integrated prophage elements are more likely to be found within RGM than in SGM in both draft genomes and complete genomes (F = 9.80, p = 1.10e-22 and F = 4.99, p = 5.89e-07, proportions z-test, Fig. 1 **| Supp.** Figure 2). In the RGM, *M. abscessus*. 88 out of 109 (80.7%) draft genomes have intact prophage elements,

while of the SGM, only 3 out of 63 *M. avium* (4.8%), 3 out of 23 *M. intracellulare* (13%) and 1 out of 11 *M. chimaera* (9.1%) draft genomes have intact prophage elements

To test if the prophage discovery tools are biasing prophage predictions in longer contiguous sequences (i.e. complete genomes), we quantified the number of edge cases by species in the 182 draft genomes. The number of edge cases are normalized by genome count per species to generate an average number of edge cases per species (Table 1). Edge cases are only quantified in the draft genome assemblies because the complete genomes are assumed to have only one contiguous sequence (Table 2**)**. The number of contigs, NG50 length (base pairs), and the number of contigs greater than 1500 base pairs showed no significant correlation with number of prophages across all species (**Supp.** Figure 3**)** indicating that assembly fragmentation likely has little effect on the number of predicted prophages between species. Thus, in future figures, draft and complete genomes are combined for downstream analysis.

Table 2
Summary statistics of predicted prophages within complete NTM genomes selected from the NCBI assembly database. Complete genomes are not fragmented into contigs, thus N50 statistics, contig count, and edge cases are not applicable. Additional information about these genomes are available in **Supplementary File 1**.

Species	Genome Count	Genomes with Phage	Predicted Prophages	Median tRNA Count
M. avium	20	1	1	59 (54–59)
M. abscessus	33	25	37	49 (46–104)

[Table 1 **and** Table 2**]**

Predicted Prophages Annotated Using Existing Mycobacteriophage Clusters

PhageDB is a database containing mycobacteriophage genomes categorized into clusters and sub clusters based on sequence similarity and shared functional genes. Predicted prophages in our study were annotated across 12 lettered clusters and 15 sub clusters (Fig. 2) using protein sequence similarity. The RGM had prophages in all 13 lettered groups, and SGM had 5 lettered groups. Most prophages (46.3%) fell into the “no cluster” category, which represents prophages having less than 10 gene hits against PhageDB or clustering with viruses without a defined cluster.

Across the entire dataset (draft and complete genomes), the 188 identified prophages consisted of 186 genotypically unique sequences. Two instances of identical prophages occurred between a prophage from a draft *M. abscessus* genome and a complete *M. abscessus* genome and could represent an evolutionary artifact.

The pangenome analysis using all predicted prophages resulted in no core genes shared among all 186 unique prophages. The largest number of prophages that shared a gene was 29 prophages and the gene function was undefined. **Supplementary Table 1** details the pangenome analyses, including shell gene

counts between prophages discretized by the PhageDB cluster. No core genes are found in any lettered cluster pangenome analysis. Hypothetical proteins of unknown function are most commonly shared shell genes within the predicted prophages.

Prophage Annotation Results and Virulence Factors by Species

There was an average of 58 open reading frames (ORFs) predicted within each of the 188 prophages from the draft and complete genomes. The average amino acid length of ORFs was 221 amino acids. The total number of ORFs among all prophages was 10,870. Of these predicted ORFs, 66.5% were labeled hypothetical proteins without a known function. More than half of the ORFs (50.5%) were annotated using sequence similarity to a protein within the PhageDB mycobacteriophage database. ARAGORN identified 161 tRNA sites with threonine carriers as the most abundant, comprising 16.8% of tRNAs. The number of ORFs annotated as bacterial genes in the dataset, not including tRNAs, was 337 (3.1%). Of this small group, 103 ORFs (0.95%) were predicted to derive from bacterial virulence factors.

Virulence factor annotations within the predicted prophages comprised less than 1% of the ORFs. The 103 predicted ORFs annotated as bacterial virulence genes were present across 53 prophages from 48 different NTM genomes. Figure 3 highlights the presence/absence of bacterial virulence factors within the predicted prophages across mycobacterial species. Of the genes annotated as virulence factors by VFDB or Patric-VF, 63.1% were originally identified in *Mycobacterium tuberculosis.*

PhageDB contains 1,795 mycobacteriophages. Within these mycobacteriophages, 187 mycobacteriophages contained 249 virulence genes or 0.13% of all predicted ORFs. To more fairly compare virulence gene abundance between our predicted prophages and PhageDB, mycobacteriophages from the clusters assigned to the predicted prophages were subset from PhageDB leaving 693 mycobacteriophages and 64,515 ORFs. Within these 693 mycobacteriophages, 124 mycobacteriophages contain 179 annotated virulence genes (0.28%). We found that bacterial virulence genes are more likely to be present within prophage elements derived from clinical sources than from all mycobacteriophage genomes in PhageDB isolated from the environmental M. *smegmatis* and the subset selected from matching clusters (F = 7.11, p = 1.17e-12 and F = 3.12, p = 1.77e-03, proportions z-test, Fig. 4).

Mycobacterium Influences on Prophage Frequency

To test if NTM isolates with prophages are more evolutionarily similar to each other than to NTM isolates without prophage, we performed a PERMANOVA test of phylogenetic distance metrics. The genome wide genetic distances of *M. abscessus subsp. abscessus* isolates are more similar within the groups: with and without prophages, than between groups (F = 2.44, p = 0.012, PERMANOVA). Figure 5 displays the distribution of prophages in M. *abscessus subsp. abscessus* and M. *abscessus subsp. bolletti* in the context of the bacterial phylogeny.

CRISPR elements have the capability to protect a bacterium against prophage integration. CRISPR elements were present in only 4 of the 53 (7.5%) complete genomes and 12 of the 182 (6.6%) draft

genomes. Presence of CRISPR elements was not associated with the number of prophages (H = 0.617, *p* = 0.43, Kruskal-Wallis). The abundance of tRNAs in the mycobacterial genomes was significantly different by growth rate with slowly growing mycobacteria species having a greater number of tRNAs (H = 89.43, *p* = 3.18e-21, Kruskal-Wallis). In addition, the relationship of tRNAs and prophage frequency corrected by species reveals a positive linear correlation only in *M. abscessus* (R^2 = 0.260, *p* = 6.4e-3).

Discussion

In this study, we detail the frequency of prophages within six different species of NTM from 182 draft genomes and 53 complete genomes. Confidence in the predicted prophages is supported by identification using two different prophage prediction tools. The number of prophages within mycobacteria with a rapid growth rate is greater relative to the number found in slowly growing mycobacteria (Fig. 1). Assemblies of slowly growing mycobacteria are more GC rich than rapid growers, which is correlated with higher contig numbers (**Supp.** Figure 2). Prophage identification does not appear to be driven by assembly fragmentation as evident by the edge case ratio. A higher edge case ratio in RGM compared to SGM means that if prophages are missing from the assembly they are more likely to be missing in RGM. In addition, there was no correlation between the number of contigs, median contig length, number of sequences greater than 1,500 base pairs, and the number of predicted prophages (**Supp.** Figure 3). This further supports the notion that assembly fragmentation is not affecting the identification of prophages.

Our analyses revealed that functioning CRISPR elements are rare in NTM, as only 16 of the total 235 samples (6.8%) contained CRISPR spacers (49). Prophage frequency did not correlate with the presence of CRISPR spacers, although the sample size of mycobacteria with CRISPR spacers could be a limiting factor within our study.

tRNAs can act as an insertion site target for prophage integration (16). Our hypothesis was that an increased abundance of tRNA would result in more prophages due to the increase in potential target integration sites. Interestingly, the abundance of tRNAs in SGM is enriched relative to RGM. This is counterintuitive to our hypothesis considering the frequency of prophages by growth rate. We did observe a positive correlation of tRNA counts and prophages in *M. abscessus* (**Supp.** Figure 4) though our sample size is limited for isolates with higher numbers of prophages. Interestingly, the increased abundance of tRNAs in SGMs did not translate to increased frequency of prophages compared to RGMs.

The evolutionary advantage for rapid-growing NTM to have more prophages is unclear. We hypothesize that the process of cell division may increase bacterial susceptibility to prophage integration. Evidence of this phenomenon is supported by prior work in E. *coli* (50) and additionally by the growth rate of *M. smegmatis*, the nonpathogenic rapid growing model organism that is commonly used to isolate mycobacteriophages (20).

Most prophages within our draft genomes did not share significant similarity to other known mycobacteriophages. The prophages predicted in this study do not share a core genome, and may reflect the wide variability of viruses. The gene shared amongst the most prophages was present in only 29

prophages. This gene and many others that are highly shared have no defined function and are labeled hypothetical proteins.

Annotating the prophage elements showcased the rarity of transduction events, transferring bacterial derived genes via integrated prophage, with only 3.1% of ORFs predicted annotated as bacterial genes. Also, virulence genes were more abundant within prophages from clinical NTM genomes than environmental mycobacteriophages cataloged in PhageDB (0.95% Clinical NTM vs 0.28\% PhageDB). Our results support our hypothesis suggesting prophages could act as a reservoir of bacterial genes important for virulence (Fig. 4).

Though clinical NTM are known to contain different levels of virulence, even within a species, it is unclear if virulence genes within prophage elements affect patient outcomes (51). Presence of virulence genes alone does not mean these genes are actively expressed, and the presence of a prophage in a genome does not guarantee a functional or excisable virus. Further studies exploring RNA transcription of mycobacteria with prophages would be helpful in characterizing the expression of phage genes. In addition, our study relied on PhageDB as an environmental proxy of mycobacteriophage. Future studies exploring prophage frequency within environmental isolates of mycobacteria are needed to directly compare prophage susceptibility of clinical and environmental NTM genomes.

This study demonstrates the presence of prophages in clinical species of mycobacteria. Prophages offer a mechanism for the genetic mosaicism of mycobacteria observed in patients with cystic fibrosis which have been observed to lack a distributed conjugal transfer (DCT) protein (52). An increase in the genetic fluidity of a bacterial infection by prophage elements can impact patient outcomes as seen in other pathogens (53). Mycobacteriophages may contribute to the pathogenicity of environmental mycobacteria and impact disease progression. Additional work is needed to understand the role of mycobacteriophages in shaping the dynamics of mycobacterial infections.

Conclusions

In summary, our results indicate that prophages are present in the genomes of clinical mycobacteria. Prophages are more likely to be present in mycobacteria with a rapid growth rate compared to slowly growing species. The mechanism and selective advantage of this enrichment by growth rate remains unclear. Prophages within mycobacteria do not share a core genome and are genetically distinct. Comparisons to other mycobacteriophages from PhageDB revealed some similarities, including shared members of lettered clusters, however the largest group of integrated prophages were not assigned to a previously defined cluster. In addition, bacterial virulence genes were enriched in predicted prophages from clinical genomes relative to environmental mycobacteriophages from PhageDB. Our comprehensive analysis of prophage frequency and their genetic composition provides insight into the capability of mycobacteriophages to transduce bacterial genes relevant to bacterial virulence, potentially influencing the progression of disease.

Abbreviations

NTM
nontuberculous mycobacteria
RGM
rapidly growing mycobacteria
SGM
slowly growing mycobacteria
MAC
M. avium complex
PATRIC-VF
Pathosystems Resource Integration Center virulence factor database
VFDB
Virulence Factor Database
CRISPR
clustered regularly interspaced short palindromic repeats
RAxML
Randomized Axelerated Maximum Likelihood-Next Generation
DCT
distributed conjugal transfer

Declarations

Ethics approval and consent to participate

Not applicable

Availability of data and material

All scripts used to derive the figures and additional preprocessing including tool overlap identification are available on the Strong Lab GitHub at https://github.com/Strong-Lab/Prophage_In_NTM. The raw data of the clinical NTM draft genomes are available at BioProject 319839 and the sources of the complete genomes are listed in **Supplemental File 1.**

Competing interests

The authors declare no competing interests financial or otherwise related to this project.

Supplementary Table 1

Abundance of shell genes in prophages calculated from pangenome analysis using Roary. Shell genes are defined as genes present in 15–95% of genomes in a cluster. This analysis of prophages assigned to

lettered clusters was only applied to clusters with 5 or more prophages.

Supplementary Fig. 1

Boxplots detailing prophage length distributions colored by growth rate. The interquartile range of prophage length from all species is 27,815 to 50,923 base pairs. A Kruskal-Wallis test of prophage lengths by growth rate failed to reject the null hypothesis suggesting no difference in the size of integrated prophages by growth rate ($H = 0.692$, $p = 0.406$).

Supplementary Fig. 2

Bar plots showing the relative abundance of prophage presence

Supplementary Fig. 3

Details how assembly statistics compare to prophage frequency in clinical NTM isolates. **A)** Correlation plot detailing the Pearson correlation values between prophage frequency and assembly statistics. **B)** Scatterplot matrix showing the individual samples colored by mycobacteria growth rate. Given the clear distinctions in assembly statistics by growth rate, linear mixed modeling was performed to calculate statistical significance of assembly features against prophage frequency. Linear mixed models of prophage frequency by number of contigs ($Z=-1.513$, $p = 0.130$), NG50 Length ($Z = 1.575$, $p = 0.115$), and number of contigs > 1500 base pairs ($Z=-0.914$, $p = 0.361$) all failed to reject the null hypothesis when modeling by species.

Supplementary Fig. 4

Phylogeny of host *M. absessus subsp. massiliense* genomes. The shaded boxes are located along an x axis, which lists the lettered PhageDB clusters. The presence of prophage in a sample is noted by a shaded box in any column except NA/Control. No shading means the sample does not have a prophage in that lettered cluster. Statistical significance was not achieved using PERMANOVA on this tree ($F = 1.398$, $p = 0.225$).

Supplementary Fig. 5

Phylogeny of host M. *chimaera* and M. *intracellulare* genomes. The shaded boxes are located along an x axis, which lists the lettered PhageDB clusters. The presence of prophage in a sample is noted by a shaded box in any column except NA/Control. No shading means the sample does not have a prophage in that lettered cluster. Statistical significance was not achieved using PERMANOVA on this tree ($F = 1.12$, $p = 0.321$).

Funding

CG is supported by NLM 5 T15 LM009451-12. The authors would like to thank the Cystic Fibrosis Foundation for funding.

Acknowledgements

The authors would like to thank the Colorado Cystic Fibrosis Research Development Program for making the clinical NTM genomes publicly available.

Tables and Figure Captions

References

1. Jhun BW. Prognostic Factors Associated With Long-Term Mortality in 1445 Patients With Nontuberculous Mycobacterial Pulmonary Disease: A 15-year Follow-Up Study.
2. Honda JR, NA Hasan RD. Environmental Nontuberculous Mycobacteria in the Hawaiian Islands. 2014 [cited 2014]; Available from: http://journals.plos.org/plosntds/article?id=10.1371/journal.pntd.0005068
3. Covert TC, Rodgers MR, Reyes AL, Stelma GN. Occurrence of nontuberculous mycobacteria in environmental samples. Appl Environ Microbiol [Internet]. 1999 Jun 1 [cited 1999 Jun 1];65(6):2492–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10347032
4. Gebert MJ, Delgado-Baquerizo M, Oliverio AM, Webster TM, Nichols LM, Honda JR, et al. Ecological Analyses of Mycobacteria in Showerhead Biofilms and Their Relevance to Human Health. mBio [Internet]. 2018 Oct 30 [cited 2018 Oct 30];9(5). Available from: http://www.ncbi.nlm.nih.gov/pubmed/30377276
5. Honda JR, Alper S, Bai X, Chan ED. Acquired and genetic host susceptibility factors and microbial pathogenic factors that predispose to nontuberculous mycobacterial infections. Curr Opin Immunol [Internet]. 2018 Oct 21 [cited 2018 Oct 21];54:66–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29936307
6. Spaulding AB, Lai YL, Zelazny AM, Olivier KN, Kadri SS, Prevots DR, et al. Geographic Distribution of Nontuberculous Mycobacterial Species Identified among Clinical Isolates in the United States, 2009-2013. Ann Am Thorac Soc [Internet]. 2017 Nov 1 [cited 2017 Nov 1];14(11):1655–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28817307
7. Lipner EM, Garcia BJ, Strong M. Network Analysis of Human Genes Influencing Susceptibility to Mycobacterial Infections. PloS one [Internet]. 2016 Jan 11 [cited 2016 Jan 11];11(1):e0146585. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26751573
8. Slowly Growing Nontuberculous Mycobacteria (NTM) - Infectious Disease Advisor [Internet]. 2020 [cited 2020 Apr 27]. Available from: https://www.infectiousdiseaseadvisor.com/home/decision-support-in-medicine/infectious-diseases/slowly-growing-nontuberculous-mycobacteria-ntm/

9. Penadés JR, Chen J, Quiles-Puchalt N, Carpena N, Novick RP. Bacteriophage-mediated spread of bacterial virulence genes. Curr Opin Microbiol [Internet]. 2015 Feb 19 [cited 2015 Feb 19];23:171–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25528295
10. Olson ME, Horswill AR. Bacteriophage Transduction in Staphylococcus epidermidis. Methods Mol Biol [Internet]. 2011 [cited 2011];1106:167–72. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24222465
11. Jiang SC, Paul JH. Gene transfer by transduction in the marine environment. Appl Environ Microbiol [Internet]. 1998 Aug 1 [cited 1998 Aug 1];64(8):2780–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9687430
12. Casadevall A, Pirofski LA. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. Infect Immun [Internet]. 1999 Aug 1 [cited 1999 Aug 1];67(8):3703–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10417127
13. León M, Bastías R. Virulence reduction in bacteriophage resistant bacteria. Front Microbiol [Internet]. 2015 Apr 23 [cited 2015 Apr 23];6:343. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25954266
14. Brüssow H, Canchaya C, Hardt W-D. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. Microbiol Mol Biol Rev MMBR [Internet]. 2004 Sep 1 [cited 2004 Sep 1];68(3):560–602, table of contents. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15353570
15. Hargreaves KR, Kropinski AM, Clokie MRJ. What Does the Talking?: Quorum Sensing Signalling Genes Discovered in a Bacteriophage Genome.
16. Williams KP. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res [Internet]. 2002 Feb 15 [cited 2002 Feb 15];30(4). Available from: http://dx.doi.org/10.1093/nar/30.4.866
17. Zeng H, Zhang J, Li C, Xie T, Ling N, Wu Q, et al. The driving force of prophages and CRISPR-Cas system in the evolution of Cronobacter sakazakii.
18. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. Bioinforma [Internet]. 2017 Mar 1 [cited 2017 Mar 1];33(5):784–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28365761
19. Hatfull GF, Cresawn SG, Hendrix RW. Comparative genomics of the mycobacteriophages: Insights into bacteriophage evolution.
20. Rybniker J, Kramme S, Small PL. Host range of 14 mycobacteriophages in Mycobacterium ulcerans and seven other mycobacteria including Mycobacterium tuberculosis--application for identification and susceptibility testing. J Med Microbiol [Internet]. 2006 Jan 1 [cited 2006 Jan 1];55(Pt 1):37–42. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16388028
21. Jacobs-Sera D, Marinelli LJ, Bowman C, Broussard GW, Guerrero Bustamante C, Boyle MM, et al. On the nature of mycobacteriophage diversity and host preference. Virology [Internet]. 2012 Dec 20 [cited 2012 Dec 20];434(2):187–201. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23084079

22. Seddon P, Fidler K, Raman S, Wyatt H, Ruiz G, Elston C, et al. Prevalence of Nontuberculous Mycobacteria in Cystic Fibrosis Clinics, United Kingdom, 2009. Emerg Infect Dis [Internet]. 2013 Jul 1 [cited 2013 Jul 1];19(7). Available from: http://wwwnc.cdc.gov/eid/article/19/7/12-0615_article.htm
23. Zhao X, Epperson LE, Hasan NA, Honda JR, Chan ED, Strong M, et al. Complete Genome Sequence of subsp. Strain H87 Isolated from an Indoor Water Sample. Genome Announc [Internet]. 2017 Apr 20 [cited 2017 Apr 20];5(16). Available from: http://www.ncbi.nlm.nih.gov/pubmed/28428297
24. Hasan NA, Davidson RM, Epperson LE, Kammlade SM, Rodger RR, Levin AR, et al. Population Genomics of Nontuberculous Mycobacteria Recovered from United States Cystic Fibrosis Patients. bioRxiv [Internet]. 2019 Jan 1 [cited 2019 Jan 1]; Available from: https://www.biorxiv.org/content/10.1101/663559v1
25. Wick RR, Judd LM, Gorrie CL, Holt KE. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol [Internet]. 2017 Jun 8 [cited 2017 Jun 8];13(6):e1005595. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28594827
26. Davidson RM, Hasan NA, de Moura VCN, Duarte RS, Jackson M, Strong M. Phylogenomics of Brazilian epidemic isolates of Mycobacterium abscessus subsp. bolletii reveals relationships of global outbreak strains. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis [Internet]. 2013 Dec 18 [cited 2013 Dec 18];20:292–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24055961
27. Datta G, Nieto LM, Davidson RM, Mehaffy C, Pederson C, Dobos KM, et al. Longitudinal whole genome analysis of pre and post drug treatment Mycobacterium tuberculosis isolates reveals progressive steps to drug resistance. Tuberc [Internet]. 2016 May 26 [cited 2016 May 26];98:50–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27156618
28. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat methods [Internet]. 2012 Mar 4 [cited 2012 Mar 4];9(4):357–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22388286
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinforma [Internet]. 2009 Aug 15 [cited 2009 Aug 15];25(16):2078–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19505943
30. Kozlov A, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. bioRxiv [Internet]. 2018 Jan 1 [cited 2018 Jan 1]; Available from: https://www.biorxiv.org/content/early/2018/10/18/447110
31. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: anrpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. McInerny G, editor. Methods Ecol Evol [Internet]. 2017 Jan 1 [cited 2017 Jan 1];8(1). Available from: http://doi.wiley.com/10.1111/2041-210X.12628
32. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinforma [Internet]. 2010 Jan 13 [cited 2010 Jan 13];11:24. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20070885
33. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic acids Res [Internet]. 2004 Jan 2 [cited 2004 Jan 2];32(1):11–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/14704338

34. Bland C, Ramsey TL, Sabree F, Lowe M, Brown K, Kyrpides NC, et al. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinforma [Internet]. 2005 [cited 2005];8:209. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17577412http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1924867&tool=pmcentrez&rendertype=abstract
35. Arndt D, Grant JR, Marcu A, Sajed T, Pon A, Liang Y, et al. PHASTER: a better, faster version of the PHAST phage search tool. Nucleic acids Res [Internet]. 2016 Jul 8 [cited 2016 Jul 8];44(W1):W16–21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27141966
36. Sullivan MB, Hurwitz BL, Roux S, Enault F. VirSorter: mining viral signal from microbial genomic data. PeerJ [Internet]. 2015 May 28 [cited 2015 May 28];3. Available from: https://peerj.com/articles/985/
37. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ [Internet]. 2016 Oct 18 [cited 2016 Oct 18];4:e2584. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27781170
38. Kloepper TH, Huson DH. Drawing explicit phylogenetic networks and their integration into SplitsTree. BMC Evol Biol [Internet]. 2008 Jan 24 [cited 2008 Jan 24];8:22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/18218099
39. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat methods [Internet]. 2015 Jan 17 [cited 2015 Jan 17];12(1):59–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25402007
40. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinforma [Internet]. 2010 Mar 8 [cited 2010 Mar 8];11:119. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20211023
41. Mauri M, Elli T, Caviglia G, Uboldi G, Azzi M. RAWGraphs: A Visualisation Platform to Create Open Outputs. In: Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter - CHItaly '17 [Internet]. ACM Press; 2015 [cited 2015]. p. 1–5. Available from: http://dl.acm.org/citation.cfm?doid=3125571.3125585
42. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinforma [Internet]. 2014 Jul 15 [cited 2014 Jul 15];30(14):2068–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24642063
43. Chen L. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res [Internet]. 2004 Dec 17 [cited 2004 Dec 17];33(Database issue). Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki008
44. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic acids Res [Internet]. 2014 Jan 12 [cited 2014 Jan 12];42(Database issue):D581–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24225323
45. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic acids Res [Internet]. 1998 Jan 1 [cited 1998 Jan 1];26(1):320–2. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9399864

46. McKinney W. Data Structures for Statistical Computing in Python. Proc 9th Python Sci Conf. 2007;1697900:51–6.
47. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci & Eng. 9(3).
48. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinforma [Internet]. 2015 Nov 15 [cited 2015 Nov 15];31(22):3691–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26198102
49. He L, Fan X, Xie J. Comparative genomic structures of Mycobacterium CRISPR-Cas. J Cell Biochem [Internet]. 2012 Jul 1 [cited 2012 Jul 1];113(7):2464–73. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22396173
50. Nabergoj D, Modic P, Podgornik A. Effect of bacterial growth rate on bacteriophage population growth rate. MicrobiologyOpen [Internet]. 2018 Apr 1 [cited 2018 Apr 1];7(2):e00558. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29195013
51. González-Pérez M, Mariño-Ramírez L, Parra-López CA, Murcia MI, Marquina B, Mata-Espinoza D, et al. Virulence and immune response induced by Mycobacterium avium complex strains in a model of progressive pulmonary tuberculosis and subcutaneous infection in BALB/c mice. Infect Immun [Internet]. 2013 Nov 19 [cited 2013 Nov 19];81(11):4001–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23959717
52. Sapriel G, Konjek J, Orgeur M, Bouri L, Frézal L, Roux A-L, et al. Genome-wide mosaicism within Mycobacterium abscessus: evolutionary and epidemiological implications. BMC Genomic- [Internet]. 2016 Feb 17 [cited 2016 Feb 17];17:118. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26884275
53. Malachowa N, DeLeo FR. Mobile genetic elements of Staphylococcus aureus. Cell Mol life Sci CMLS [Internet]. 2010 Sep 29 [cited 2010 Sep 29];67(18):3057–71. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20668911

Figures

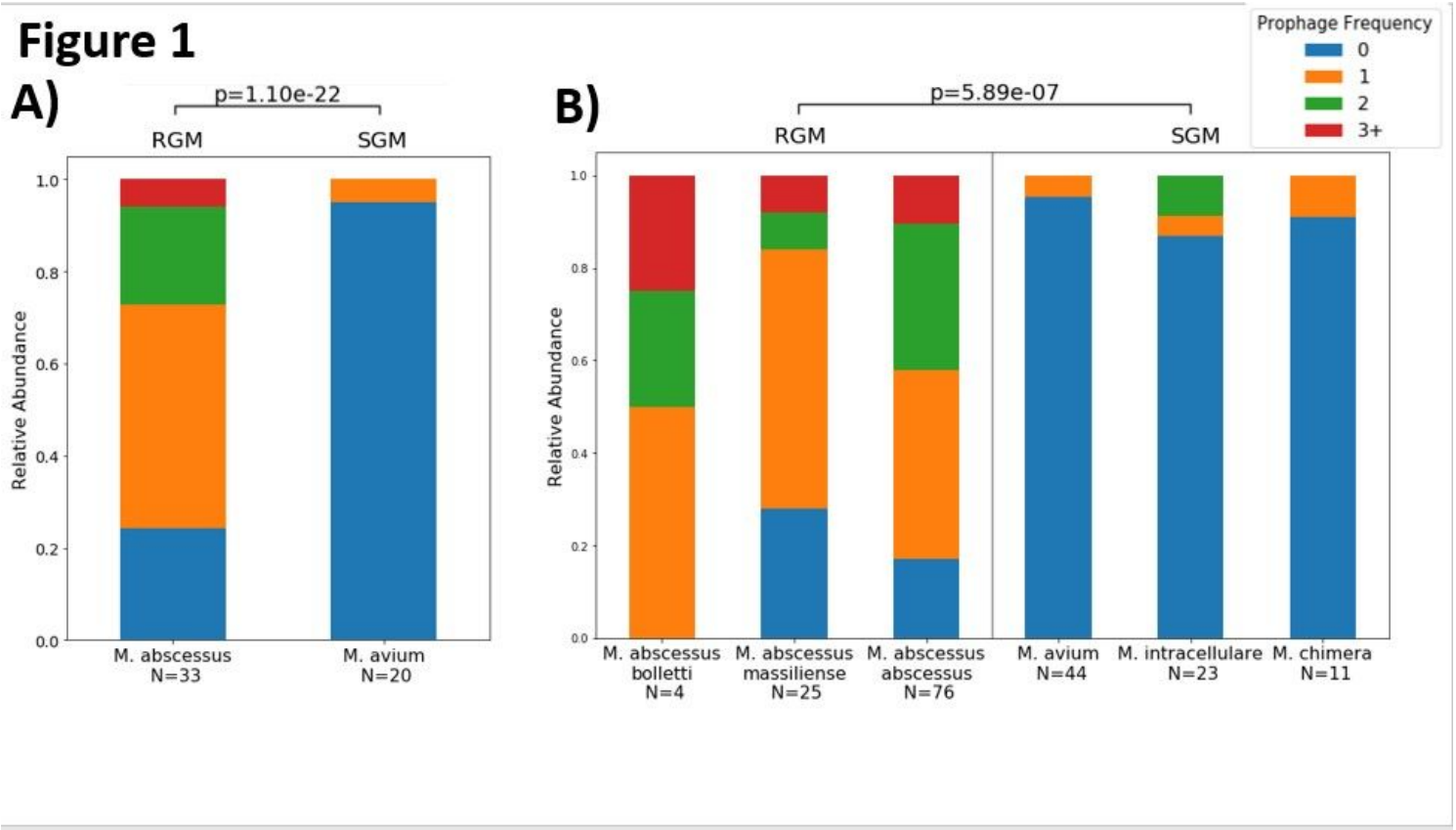

Figure 1

Prophage Frequency by NTM Species: Bar plots show relative abundance of prophage frequency in samples. Rapidly growing mycobacteria species are on the left, and slowly growing mycobacteria species are on the right. A) The frequency of prophages by genome in complete NTM genomes. The presence of prophages is statistically significant by growth rate (p=1.10e-22). B) The frequency of prophages per isolate from draft NTM genomes. The presence of prophages is statistically significant by growth rate (p=5.89e-07).

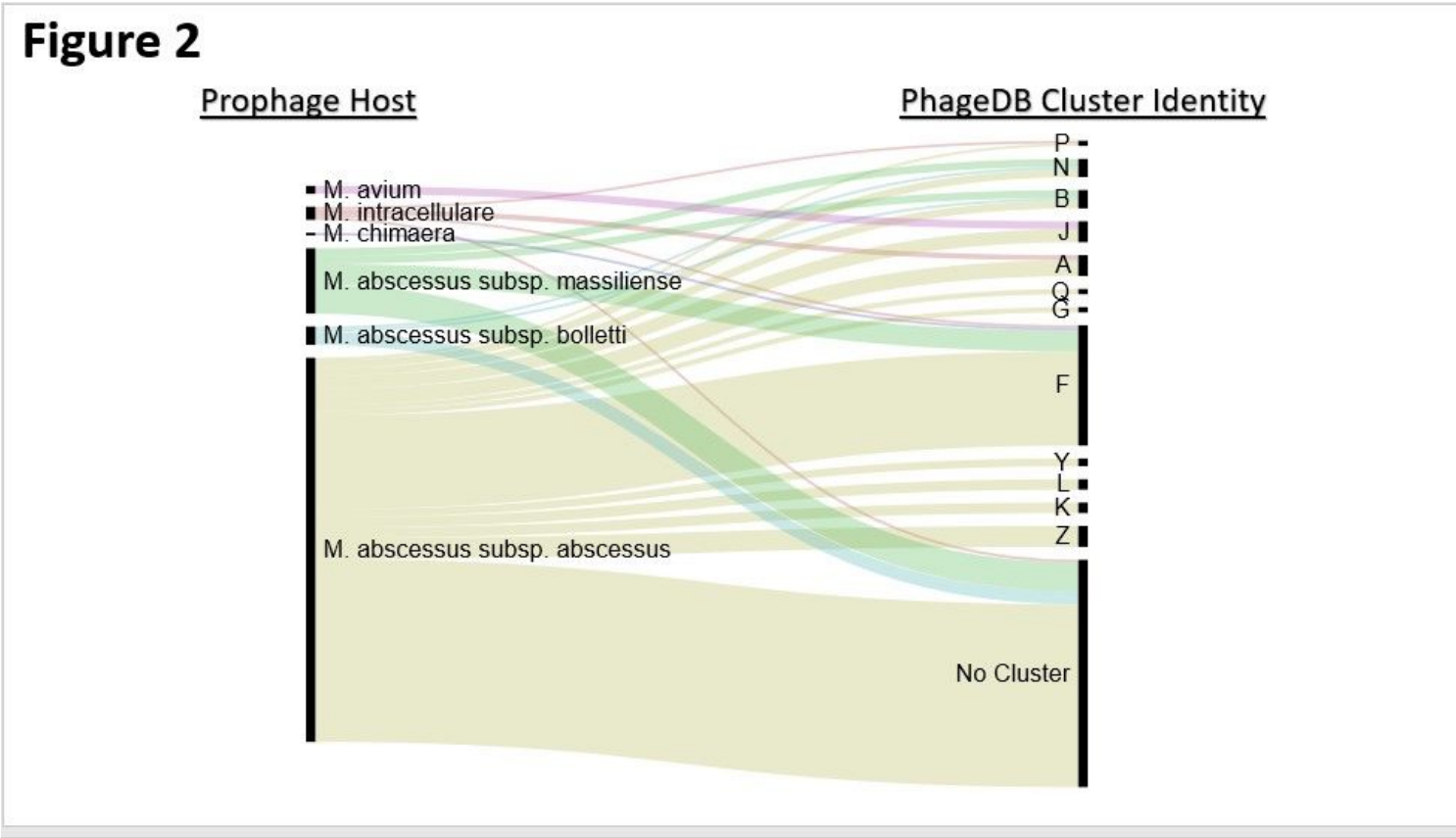

Figure 2

Prophage Assignments to PhageDB Clusters: Alluvial graph depicting assignment of predicted prophages by NTM species to a PhageDB lettered cluster (on the right) and NTM species (on the left). Line width corresponds to the number of predicted prophages from a genome that are assigned to a specific PhageDB cluster.

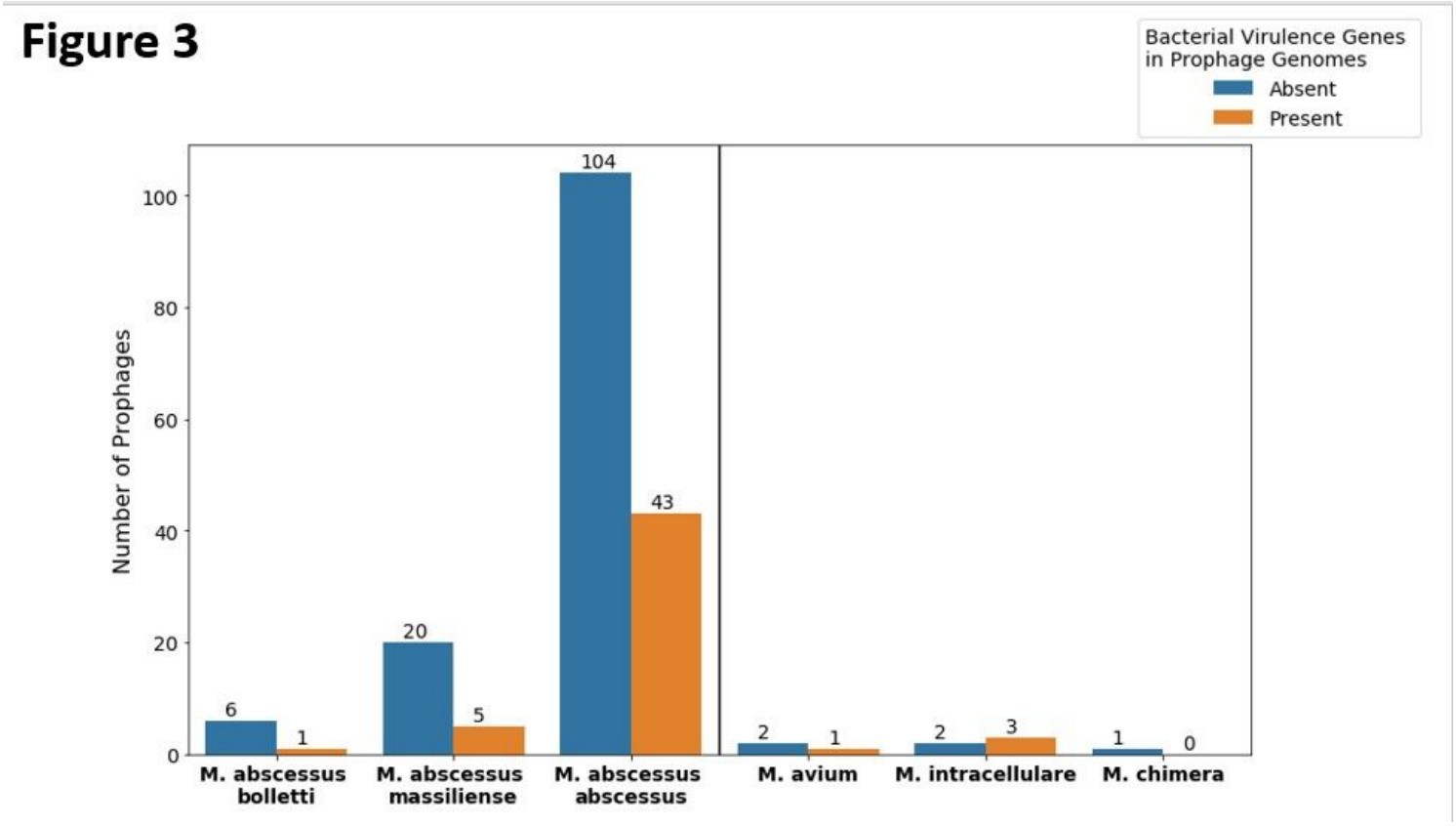

Figure 3

Bacterial Virulence Frequency in Prophages by Species: Bar plots showing the frequency of a bacterial virulence factors within predicted prophages by mycobacterial species. The presence of bacterial virulence genes in prophage genomes is not statistically significant by growth rate (p=0.267).

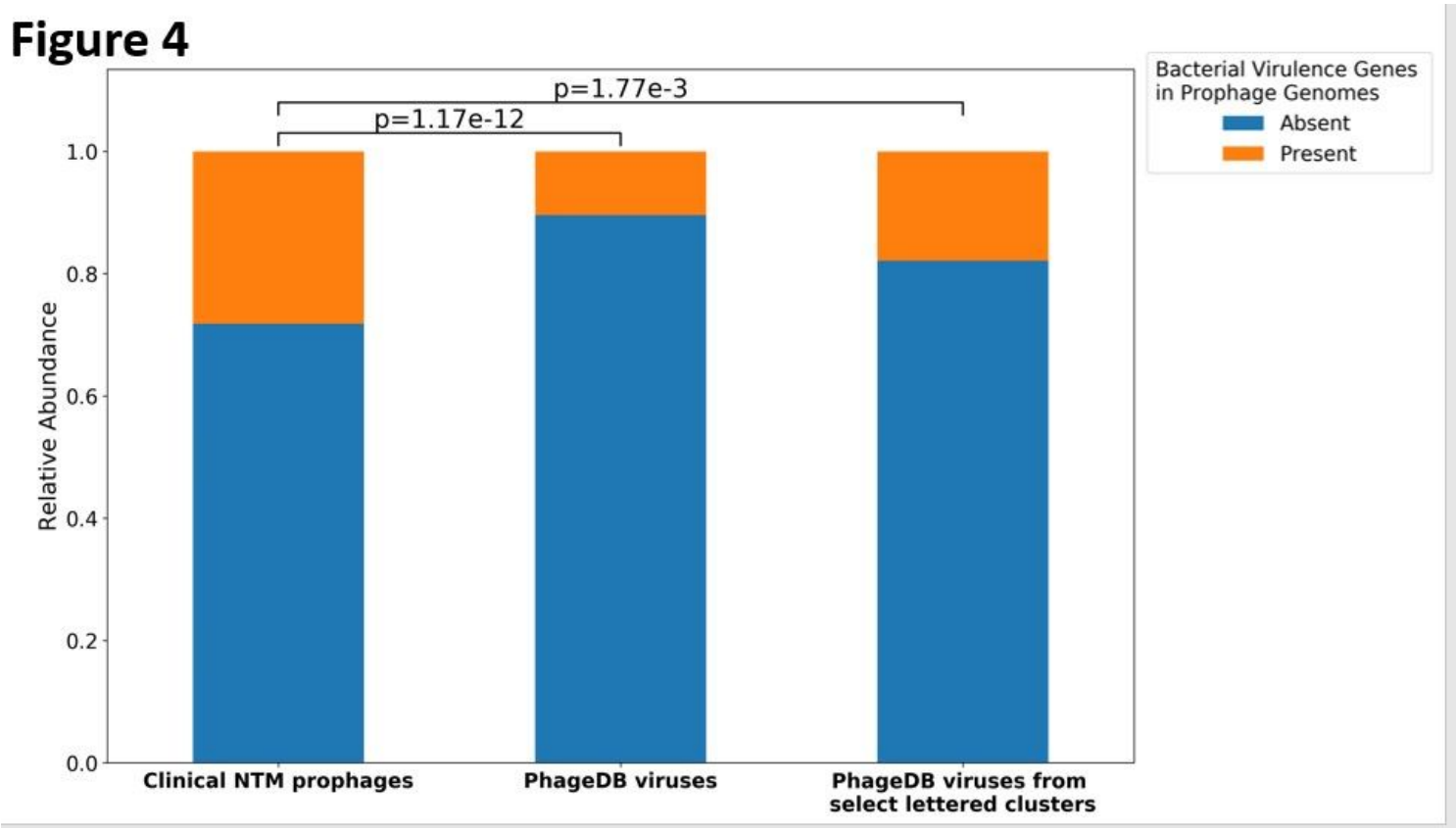

Figure 4

Bacterial Virulence Frequency in Prophages by Data Source: Bar plots showing relative abundance of bacterial virulence genes within viral genomes from our predicted prophages, mycobacteriophages from PhageDB, and a subset of PhageDB mycobacteriophages from lettered clusters our predicted prophages matched against. The presence of bacterial virulence genes in the genomes of our predicted prophages is statistically significant against the presence of bacterial virulence genes in both the full PhageDB mycobacteriophages and the select PhageDB mycobacteriophages from matched lettered clusters ($F = 7.11$, $p = 1.17e\text{-}12$ and $F = 3.12$, $p = 1.77e\text{-}03$, proportions z-test).

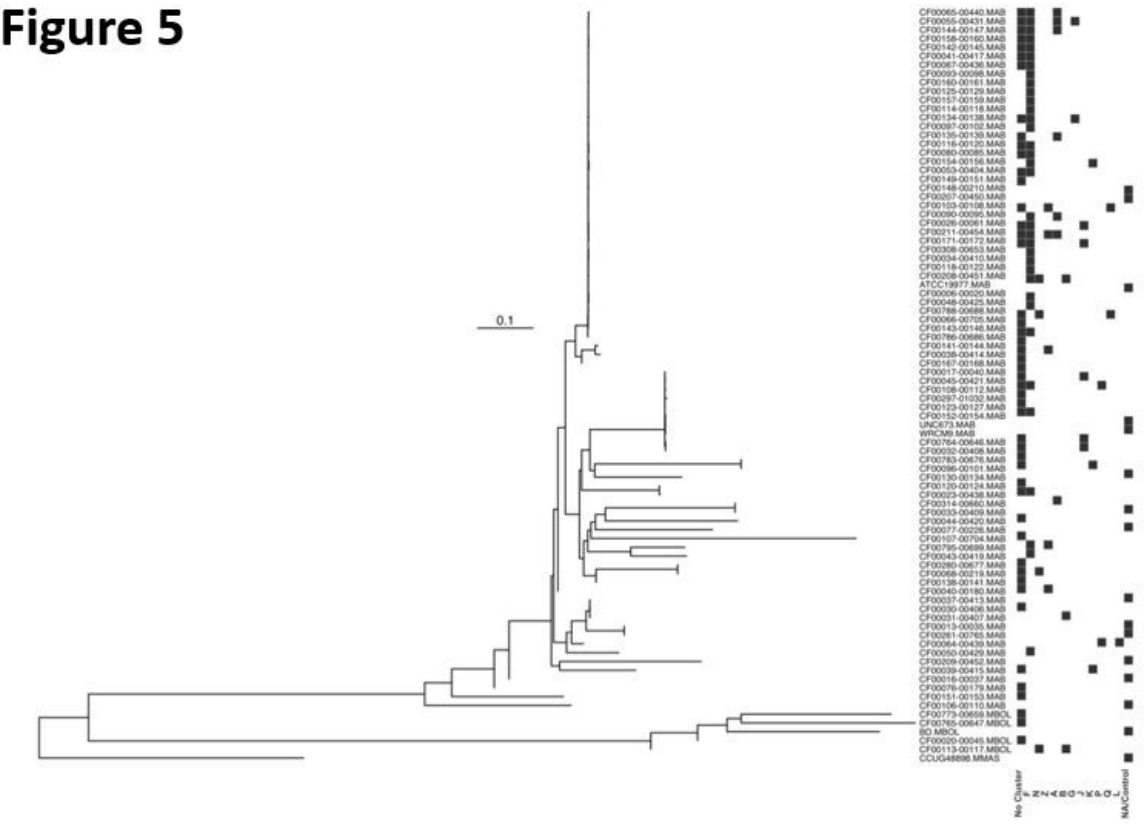

Figure 5

Phylogeny and Prophage Frequency: Genome wide phylogeny of 81 M. abscessus subsp. abscessus and M. abscessus subsp boletti genomes and 5 control genomes. The heatmap on the right shows the presence (black) or absence (white) of the lettered PhageDB clusters (x-axis). The presence of prophage in a sample is noted by a shaded box in any column except NA/Control. NA/Control shading signifies genomes without a prophage and added controls not analyzed for prophages in this study. No shading means the sample does not have a prophage in that lettered cluster.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- S4.jpg
- S3.jpg
- S5.jpg
- S2.jpg
- S1.jpg
- SupplementaryTablesandSupplementaryFigureCaptions.docx
- SupplementaryFile1Glickmanetal.csv