The protein interactome of mycobacteriophage Giles predicts functions for unknown proteins

Running Title: Protein interactome of Mycobacteriophage Giles

Jitender Mehla¹, Rebekah M. Dedrick², J. Harry Caufield¹, Rachel Siefing¹, Megan Mair¹, Allison Johnson¹, Graham F. Hatfull², and Peter Uetz¹*

¹ Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284
² Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260

* corresponding author: peter@uetz.us.

Word count: ~5,300 (main text, excluding references and figure legends)

Word count: 154 (abstract)
**Abstract**

Mycobacteriophages are viruses that infect mycobacterial hosts and are prevalent in the environment. Nearly 700 mycobacteriophage genomes have been completely sequenced, revealing considerable diversity and genetic novelty. Here, we have determined the protein complement of mycobacteriophage Giles by mass spectrometry and mapped its genome-wide protein interactome to help elucidate the roles of its 77 predicted proteins, 50% of which have no known function. About 22,000 individual Yeast two-hybrid (Y2H) tests with four different Y2H vectors, followed by filtering and re-test screens, resulted in 324 reproducible protein-protein interactions, including 171 (137 non-redundant) high confidence interactions. The complete set of high-confidence interactions among Giles proteins reveals new mechanistic details and predicts functions for unknown proteins. The Giles interactome is the first for any mycobacteriophage and one of just five known phage interactomes so far. Our results will help in understanding mycobacteriophage biology and aid in development of new genetic and therapeutic tools to understand *Mycobacterium tuberculosis*.

**Importance:**

*Mycobacterium tuberculosis* causes over 9 million new cases of tuberculosis each year. Mycobacteriophages, viruses of mycobacterial hosts, hold considerable potential to understand phage diversity, evolution, and mycobacterial biology, aiding in development of therapeutic tools to control mycobacterial infections. The mycobacteriophage Giles protein-protein interactions and interactions network allow us to predict functions for unknown proteins and shed light on major biological processes in phage biology. For example, Giles gp76, a protein of unknown function, is found to associate with phage packaging and maturation. The functions of...
mycobacteriophage-derived proteins may suggest novel therapeutic approaches for tuberculosis.

Our ORFeome clone set of Giles proteins and the interactome data will be useful resources for phage interactomics.
Introduction

The continuous emergence of bacterial pathogens resistant to antibiotics is an increasing medical problem (1). *Mycobacterium tuberculosis* is prominent among these pathogens, with over 9 million new cases of tuberculosis reported each year. There is an urgent need for alternate ways to control *M. tuberculosis* infections, and one potential strategy involves using mycobacteriophages for prophylaxis or therapy (2). The emergence of extensively drug-resistant (XDR) and totally drug-resistant (TDR) strains of *M. tuberculosis*, both of which are especially difficult to control (3), has spurred renewed interest in the therapeutic use of bacteriophages.

Mycobacteriophages are known to infect many different species of both fast- and slow-growing mycobacteria, including *M. tuberculosis* and *Mycobacterium smegmatis* (4, 5). Over the past decade, thousands of mycobacteriophages have been isolated and hundreds have been completely sequenced (http://phagesdb.org/). Mycobacteriophage genomes are highly mosaic due to horizontal genetic exchange (6-8) but can be grouped into 20 clusters and eight singletons, i.e. phages for which no close relatives have yet been identified (9, 10). Genomic characterization of nearly 700 of these phages (http://phagesdb.org/) reveals a staggeringly large number of genes (>40,000) coding for products of unknown function.

Giles, a temperate mycobacteriophage, has a 53,746 bp genome coding for 77 proteins (11, 12). Giles belongs to the Q cluster of mycobacteriophages, which includes four other phages with very similar genome sequences. More than half of the proteins encoded by Giles are functionally uncharacterized and most of its unknown proteins do not have close homologs or orthologs in other mycobacteriophages outside of the Q cluster (12). Thus, no clues are available for these
proteins of unknown function. One key to understanding Giles biology will rely upon elucidating uncharacterized proteins through their protein-protein interactions (PPIs). The full set of PPIs for this virus (that is, its interactome) should provide functional clues unattainable by studying proteins individually (13).

PPIs are essential to understanding how different proteins in phages perform their functions in vivo, either alone or in a well-coordinated, cross-regulated interaction network. Understanding the interactome of a given organism provides new insights to the key steps of major biological pathways. A better comprehension of these pathways, especially in the context of bacterial host pathogenicity, will assist us in treating diseases. Mapping PPIs is specifically essential to unraveling the biology of uncharacterized proteins in Giles. An established strategy to investigate protein function, especially among bacteriophage proteins, is to identify PPIs that may link uncharacterized proteins to well-studied proteins (14). A protein interaction network for any mycobacteriophage may also shed light on the function of many proteins in hundreds of related phages. Exploring the activities of mycobacteriophage-derived proteins provides an arsenal of potential therapies for *M. tuberculosis* and other mycobacterial infections.

In the present study, we have comprehensively and systematically analyzed the proteome of mycobacteriophage Giles for binary protein-protein interactions using a combination of four Y2H vectors to maximize coverage and reliability of the protein-protein interaction data set. This study provides the first interactome of any mycobacteriophage and one of only 5 published phage interactomes to date (15,16,17,18).
Results

Mapping the Giles interactome with a high throughput Multi Vector-Yeast Two Hybrid (MV-Y2H) approach

We successfully screened a total of 2 x 74 and 2x73 bait strains (using all vectors) corresponding to 100% of all the available ORFs or ~95% of ORFs in the mycobacteriophage Giles genome (Fig. 1A, see Suppl. Table S1A for the full list). Our Y2H screens covered 91.12% of all 2.96·10^3 possible bait-prey combinations corresponding to 2.7·10^3 possible interactions. The first set of Y2H screens yielded 2662 raw binary interactions. This set was filtered to produce a set of retesting candidates (~ 1000); retesting these baits and preys yielded 324 reproducible interactions of which 171 PPIs (~53% of the reproducible results, ~17% of filtered set and ~6% of the raw data) were deemed “high confidence”. Representative screens are shown in Suppl. Figures S1-3.

In this study, we used four different Gateway-compatible vectors (pGBGT7g, pGADT7g, pGBKCg, and pGADCg) to test for interactions among nearly all protein-protein pairs in the mycobacteriophage Giles proteome. We constructed bait and prey arrays by transferring Gateway-compatible Giles entry clones into the Y2H expression vectors followed by transformation into mating-competent yeast strains. Our final arrays contained 74 and 73 baits (in pGBGT7g and pGBKCg as fusions to the Gal4 DNA-binding domain, respectively) and 73 preys (in both pGADT7g and pGADCg as fusions to the Gal4 activation domain) (Fig. 1A). Clones corresponding to Giles proteins gp20 and gp22 were not available and were not included in the final arrays (Suppl. Table S1A).
After constructing the arrays, the self-activating baits were identified within each vector combination (pGBK Cg vs pGAD Cg; pGBK Cg vs pGAD T7 g; pGBGT7 g vs pGAD T7 g and pGBGT7 g vs pGAD Cg). The self-activating baits (a full list is available in Suppl. Table S1B) allowed for background growth signal when mated with an empty prey vector (that is, in the absence of any interacting protein partner). Hence self-activating baits are not ignored but screened and included in the dataset.

Special precautions were taken to ensure that bait self-activation did not lead to false-positive results. A series of 3-AT concentrations between 0 and 100 mM was used to suppress the background from self-activating baits (see Materials and Methods for details regarding use of 3-AT). A few baits (gp11 and gp25), when expressed from pGBK Cg, show background growth even at 100 mM 3-AT. The bait gp25 was an auto-activator in pGBK Cg but not in other vector combinations.

In an effort to maximize assay sensitivity, we screened all the baits in binary pairs with preys. All interactions discussed in this study were collected from binary screens followed by re-testing. Images of the each screen plate were produced (raw result images can be provided on request). All screens were performed in 384-colony format, with each plate including ~73 preys screened against each bait, performed in quadruplicate. An overview of the process is presented in Fig. 1B.

A series of 3-AT concentrations (0, 1, 3, 10, 25 and 50 mM) was used in Y2H re-test screens to quantitatively assess PPIs. The re-testing of filtered PPIs results in a data set of 324 reproducible PPIs. This dataset includes 75 out of the 77 proteins (~97%) in the mycobacteriophage Giles proteome (Fig. 1B).

Coverage and completeness of the Giles interactome
Each protein was tested in 8 different configurations, either as bait or prey, using four sets of Y2H expression vectors (pGBKCG/pGADCg, pGBKCG/pGADT7g, pGBGT7g/pGADCg and pGBGT7g/pGADT7g) (Fig. 1A). Out of the 77 predicted Giles proteins (or 75 proteins tested either as bait or prey) available in the Giles proteome, 75 proteins were found to be involved in protein-protein interactions (Fig. 1B). Out of 74 baits, 71 (95.9%) were observed in the interactome. In contrast, out of 73 tested preys, just 58 (79.4%) contributed to the interactome, though all 75 proteins contributed interactions across all bait vs. prey combinations. Thus, the Giles interactome covers between 95% and 100% of its available, predicted proteome. The interactome size was calculated only for reproducible PPIs in our retest data set (Table S2B). For the set of high confidence PPIs (IScore> 0; 171 PPIs), 70 Giles proteins were found to have at least one interacting partner, representing about 90% of the Giles predicted proteome. Ten Giles proteins (including scaffold protein gp8 and tail protein gp16, among others) were involved in only one interacting pair each (excluding self-interactions) and 6 proteins (gp24, gp31, gp39, gp60, gp62 and gp65) were found to have more than 10 interacting partners. The completeness of the Giles interactome space (that is, the percentage of tested pairs out of all possible protein pairs) was 91.12%. Full lists of protein-protein interactions are available in Suppl. Tables S2A, S2B, and S2C.

Each vector combination (2 N-terminal and 2 C-terminal fusions, Fig 2A) produced different, non-overlapping, PPI sets. For example, the vector combination “CC” produced a set of total 79 PPIs including 14 overlapping interactions with other vector combinations (e.g. 1 with NN; and 13 with NC, Fig. 2B). None of the high-confidence interactions were visible in more than two expression vector combinations. For example, the bait gp62, a putative DNA methylase protein, produced a total of 9 PPIs with CC and CN combinations yielding 1 and 3 interactions,
respectively. The N-terminal fusions of gp62 (NC and NN) yielded 3 and 2 interactions, respectively. No gp62 PPI was observed in more than one vector pair, confirming that different protein fusion result in different sets of PPIs. Numerous other examples are present across the interactome dataset. Only 20 redundant PPIs (that is, those detected by more than one expression vector combination) were observed, comprising 6% of the dataset. No interactions were detected in three or four vector combinations. The results clearly indicate that different vector combinations have different potency to dissect the Giles interactome by exploring interactions resulting from different regions of each protein. Furthermore, the vector combinations produced a varied number of total PPIs at different 3-AT concentrations (Fig. 2A). The data shown here was reproducible in re-tests and was obtained at 3-AT concentrations which clearly differentiate background signal from true positive interactions.

Identification of high confidence protein-protein interactions

The filtering and re-testing of the raw Y2H data resulted in a reproducible data set of 324 PPIs. To further select the highest confidence PPIs, an IScore was calculated and assigned to each reproducible interacting pair (Suppl. Fig. S4, S8, and Methods). The IScore for all PPIs ranges from -20 to 100. A PPI with an IScore of 100 was classified as the most reliable and -20 was the least reliable. All PPIs with an IScore > 0 were classified as high confidence PPIs. Further scoring of PPIs resulted in a final set of 137 high confidence (excluding redundant) PPIs (see Suppl. Table S2C).

The structure of the Giles interactome

The Giles protein–protein interactome appears to be tightly connected and intricately cross-regulated. Few portions of the phage proteome appear to be enriched for interactions with
neighboring proteins or with proteins in distinct regions (Fig. 3A). Rather, some segments of the proteome do not appear to interact with other segments: gp3 through gp12 produced very few interactions with each other, while proteins gp42 through gp57 produced few interactions among themselves or with gp1 through gp12. Most other regions of the interactome space yielded a pattern of interactions scattered across the other regions. The Giles lytic phase transcripts (described in further detail in (12) place the interactome in the context of gene expression (Fig. 3A). This context reveals that protein interactions appear to cross transcript boundaries, a finding suggesting intricate cross-regulation.

The predicted proteome of Giles can be organized into functional groups, including structural (head and tail assembly), recombination (integration and excision), etc. We found many interactions within groups such as those between phage structural proteins (~11% of the total PPIs) (Fig. 3B, Suppl. Table S2F). For example, gp6 (portal) was found to interact with gp24 and gp25, two tail assembly proteins.

There are 131 cross-functional interactions (~40.5%) between structural proteins and other functional/regulatory proteins, including proteins of unknown function. For instance, gp26, (a tail assembly protein) and gp15 (a minor tail protein) interact with gp31 (LysA), thus connecting two different biological processes.

Another 157 PPIs (~48.6%) were detected between non-structural proteins, which include all 45 proteins of unknown function. For example, gp62, a DNA methylase, interacts with gp56 and gp57, two uncharacterized, potential regulatory proteins. Out of all PPIs, 136 are detected between two proteins of unknown function, accounting for about 86% of non-structural PPIs or 42% of total PPIs. Thus, a major part of the Giles interactome is between functionally uncharacterized proteins (Fig. 3B), emphasizing the need for further protein characterization.
Some proteins, including the RecE and RecT-like gp52 and gp53 respectively, have predicted functions based on their homology with other protein sequences but are considered uncharacterized in this context due to their low sequence similarity with their closest potential homologues.

The topology of the Giles protein interaction network

Fig. 4A presents the Giles protein interaction network: each edge indicates a pair of nodes that interact at least once in the high-confidence data. The full high-confidence interaction set contains 171 pairs, though these include reciprocal interactions and make distinctions between different expression vector pairs. The compressed network of high-confidence interactions among Giles proteins, as seen in Fig. 4A, is a set of 70 nodes and 136 edges, where each node is a unique protein and each edge is an interaction between two proteins. The average number of neighbors within this network is 3.69; all nodes except gp1 are connected in at least one path. The network demonstrates a loose fit to the generally expected power-law distribution ($R^2 = 0.615$) though this may be the result of a small data set and the filters used to render it more biologically relevant. Despite its small size, this network has a clustering coefficient of 0.068, a value similar to those observed for yeast and C. elegans PPI networks (19). Even after filtering, the Giles interactome demonstrates high interconnectivity, with the majority of shortest paths between nodes including 4 edges at most.

The Giles interactome network can be separated into subgraphs on the basis of gene essentiality, such that each subgraph only contains interactions between proteins essential or non-essential to the lytic cycle. These subgraphs contain all nodes with a specified property (essential or non-essential) with the exception of those which do not interact with any proteins with the same property. Examining interactions between essential, or likely essential (12), proteins (Fig. 4A),
we see that this set of 32 nodes contains predominantly structural and assembly proteins (18 in total). Out of these 32 proteins, 9 have no known or predicted function. Conversely, the set of interactions between nonessential proteins (Fig. 4A) is enriched for uncharacterized proteins. This 25 node set contains just 5 nodes with any predicted function.

Benchmarking and validating the Giles interactome using mass spectrometry

Neither the genome nor the interactome provides any information about the abundance of proteins or whether they are expressed. One method of benchmarking detected protein-protein interactions is to measure co-expression of proteins. If two interacting proteins are co-expressed at the same time interval during the bacteriophage life cycle, they are more likely to interact than two proteins present during different times of the cycle. Thus, to provide further support for our data, we did mass spectrometry analysis of phage particles and infected host cells (see Methods and Table S1A). Interestingly, we were able to detect 46 of the 77 Giles proteins by MS (Fig. 4A, B, and Suppl. Table S1A) and this subset was enriched for essential proteins (29 in total, vs. 12 amongst undetected proteins) and structural/assembly proteins (19 in total, vs. 2 among undetected proteins). A set of 42 proteins was detected in the Giles particle alone, while early (30 minutes post-infection) and late (2.5 hours post-infection) lysate samples contained 16 and 42 unique proteins, respectively. When arranged with the interaction network, these results also provide evidence for interactions among more than two proteins at a time. For instance, the minor tail protein gp21 may participate in interactions with both gp36 (a predicted virion protein without a known function) and the repressor protein gp47.
About one quarter of interacting pairs are co-expressed and detected in the same sample, validating the PPI data set. Thirty-one Giles proteins, however, including both essential and structural proteins, were not captured in any MS sample (Fig. 4A, B).

**Predicting functions for unknown Giles proteins**

Connections between proteins of known and unknown function can provide evidence to aid in function prediction. For example, gp59, a protein of unknown function that is nonessential for lytic growth, interacts with almost all of the tail assembly proteins (including gp24, gp25, and gp26). The gp59 deletion mutant has an increased rate of lysogeny and reduced fecundity (12). Due to its interactions with tail assembly proteins, we hypothesize that gp59 is also required for tail assembly and thus, deletion of the protein may lead to structural defects that cause failure in assembly and prevent completion of the lytic cycle.

gp50 (a protein of unknown function) appears to interact with gp61 (DnaQ), a replication associated protein. Thus, gp50 may have a shared functional role in DNA replication. Additionally, the results of fecundity and lysogeny assays with gp50 deletion mutants (12) suggest that gp50 could be a protein that affects a major biological process common to both lytic and lysogenic phases. Dedrick et al. (12) also predicted that gp50 is a DNA replication associated protein further supporting our data.

**Protein-protein interactions involving structural proteins**

Proteins in the virion are expected to be among the most abundant phage proteins, hence they are detected relatively easily by MS (Table S1A). For instance, the head-tail connector structure is well-represented: gp12 and 14 both interact with tail assembly proteins. The structural role of head-tail connector proteins gp10 and 11 is less clear, as they interact with each other but not...
with any other head-tail connector components. Out of 9 tail proteins confirmed to be present in
the virion (12), 6 yielded direct interactions with other structural components (Table S2D), and 7
were found in interactions with other non-structural proteins (Table S2E), though just 5
interactions are between proteins found co-occurring in MS results (Fig. 4A, B). This suggests
that many more proteins are part of the virion (or involved in its assembly) than are detected by
MS.

Topological coherence of the Giles interactome
To estimate the degree to which essential and non-essential proteins are over- or under-
represented within the network (Suppl. Fig. S5), we investigated the excess retention among
essential proteins in the network (.20). Essential proteins are all those identified by Dedrick et al.
(12) to be essential for the phage’s lytic cycle, whether by experimental observation or inferred
from predicted structural roles. Essential and non-essential Giles proteins appear to retain similar
topological characteristics for \(k<10\). For \(k>10\), essential proteins appear to be over-represented.
This trend decreases for \(k>17\), potentially due to the limited number of nodes present in these k-
cores. It is likely that many of the connections seen for very highly-connected (i.e., \(k>15\)) nodes
are false positives, though these results show that the most highly-connected nodes are generally
essential nodes. See Materials and Methods for more details.

Conserved Giles interactome
Giles is an unusual mycobacteriophage; it has only a few relatives and they are nearly identical.
As a consequence, only 19 Giles proteins have homologs outside its cluster (Q). We have
summarized homologies of Giles (Suppl. Fig. S6) along side its high-confidence interactions
(score> 30, Suppl. Fig. S7). Surprisingly, only two interaction pairs are conserved in other
phages outside of the Q cluster: gp5 and gp76 were found to interact in Giles, and homologs of both proteins are found in Cluster P phages (e.g. Fishburne). Similarly, homologs of the Giles interaction pair gp26 and gp31 are found in the cluster F phage Brocalys.

**Discussion**

More than half of the mycobacteriophage Giles proteome is without functional annotation, including many essential and non-essential proteins of unknown function. Thus, dissecting the Giles proteome can provide hints about putative protein functions and how these proteins form functional or structural associations in a tightly connected Giles PPIs network. We initially attempted a pooled screening approach by screening pools of 5 to 7 non-self-activating baits at a time. Pooling baits appeared to decrease assay sensitivity and reduced the number of observed interactions per bait. We suspect that diminished sensitivity is the result of a reduced number of yeast cells per bait available for mating with prey-containing yeast cells. Thus, an array based Y2H system was used to map the protein interactome of mycobacteriophage Giles.

More than 45% of the reproducible PPIs in our screens are of high confidence (see Results, Table S2C) and are shown in the final PPI network (Fig. 4A). The essential proteins in the Giles PPIs network are more tightly connected than non-essential proteins. Thus, essential proteins are over represented in the Giles interactome. About 86% of the non-structural PPIs or 42% of the total PPIs in this study involve proteins of unknown functions. A major part of the Giles interactome is between functionally unknown proteins, emphasizing the need to characterize and understand their role in phage biology. Multiple vectors in a Y2H screen mimics the PPI data detected by different methods (21). We used a MV-Y2H system to test each protein as N- and C
terminal bait and prey fusions. Thus, each protein was tested in 8 different configurations, which increase the chances of detecting an interaction 8 fold. The use of multiple vectors also minimizes the impact of self-activating baits. The baits found as self-activating in one vector combination did not show self-activation in other. Thus, the multi-vector Y2H can balance the quality and size of the interactome by reducing both false positives and false negatives. In this study, we took additional care to minimize false negatives and false positives by considering multiple parameters of each interaction across all four vector pairs.

The Giles gp76 protein interacts strongly with gp5, a large subunit of terminase enzyme required for DNA packaging and maturation. Gp76 has similarity to HNH endonucleases. HNH motif-containing endonuclease proteins may interact with phage terminase proteins to promote phage DNA packaging and maturation (22). For example, a recent report showed that gp74 (an HNH endonuclease) of phage HK97 interacts with its terminase protein (23) and that the HNH motif of endonuclease is essential for interacting with terminase and completion of DNA packaging. Also, lambda gpFI, an endonuclease, interacts with lambda terminase (16). Many phages encode an HNH protein (endonuclease) located adjacent to the phage DNA packaging enzyme terminase, suggesting roles in phage DNA packaging and maturation (24). It is important and interesting that all of the phages including mycobacteriophage Giles, harbouring ‘hnh’ genes are ‘cos’ phages with linear DNA containing cohesive ends. Also, the phages φ12 and the φSLT require HNH nuclease as well as TerL (terminase large subunit) for cos-site cleavage and ultimately for DNA packaging (25). A hypothetical protein of M. avium 104 (MAV_0815), found to be similar to Giles gp76, interacts with the large subunit of phage terminase [MAV_0813, STRING,(26)]. Thus, the existing evidence and the nature of the Giles gp76-gp5 interaction observed in this study suggest that Giles gp76 may be essential for DNA packaging
and maturation of the phage. Interestingly, both gp5 and gp76 appear to be conserved in some mycobacteriophages from the P cluster, suggesting this functional pairing is maintained/conserved in more distantly related phages. Somewhat unexpectedly, no interaction could be found with the capsid protein gp9, not even with itself, possibly because the capsid proteins have not been processed or because no scaffold proteins were present in our screens. The gp6 (portal) and gp5 (terminase) proteins, though expected to interact, did not produce reproducible interactions due to high background signal. This PPI was also not observed between portal and terminase in the *Streptococcus* phage Dp-1 interactome (15). Interactions between gp20 (tape measure protein) and other proteins were expected but were not tested in this study, though the observed interaction between tail protein gp15 and gp31 (LysA) may present a similar type of interaction.

Harnessing phage activities and phage-derived proteins may offer new venues for phage therapy with mycobacterial infections. Recent research into mycobacteriophage therapy for *M. ulcerans* (27) has shown promise. Similar studies with other mycobacterial infections and phages could prove similarly fruitful. Mycobacteriophages, with their mosaic genomes and great genetic diversity, offer a multitude of options for manipulating or controlling mycobacteria.

### Materials and Methods

**Bait and prey array construction**

All Giles ORFs were cloned into Gateway compatible bait (pGBGT7g, pGBKCG) and prey vectors (pGADT7g, pGADCg) as previously described (28). Bait and prey plasmids were isolated from *E. coli* and the yeast strains MATα (AH109) and MATα (Y187) were transformed with both bait and prey plasmids, respectively (29, 30) using a slightly modified Li-Ac method.
Briefly, the log phase cells were washed and suspended in 1 ml of 0.1M LiAc/TE buffer for 30 minutes before transformation. The cells were then suspended in a reaction mix of 40% PEG and 2.5 µl of 10 mg/ml boiled carrier DNA (Ambion sheared salmon sperm DNA, Life Technologies). At least 100 ng of plasmid DNA was added to each well/tube. After gentle shaking, tubes/plates were placed at 30°C for 45 minutes, then transferred to 42°C for 15-30 minutes. Plates were centrifuged and cells were resuspended in 50-100 µl of sterile dH2O and spread on selective media plates (e.g. -Leu or -Trp). All the bait and prey clones were grown in YPD supplemented with adenine in a 96 well plate overnight and were pinned onto selective media plates (-Leu and -Trp, respectively) to verify transformation and growth. All bait and prey arrays were constructed and stored on selective (-Trp or -Leu) media. Working arrays were kept on YPD plates with 0.54 mM adenine to increase the mating efficiency.

A combination of N-terminal (pGBGT7g and pGADT7g) and C-terminal (pGBKCG and pGADCG) protein fusions (28) was used to enhance coverage and produce a reliable and credible interactome data set. Here, an N-terminal or a C-terminal protein fusion refers to the bait or prey containing the DNA binding domain or activation domain at its N-terminus or C-terminus, respectively. Each protein was tested both as both bait and prey, with the exception of gp20 (the tape measure protein) and gp22 (a predicted minor tail subunit), two constructs we were unable to produce in any vector. The constructs of gp38 (a protein of unknown function) in pGADCG and gp50 (also a protein of unknown function) in pGBKCG could also not be produced but were each used with the other three members of the vector set. With the missing preys, 75 Giles preys (73 of which were used as both N- and C-terminal protein fusions) were available out of 77 protein-coding genes. In total, the use of two different protein fusion variants and two different bait vs. prey arrangements yielded eight different configurations for each protein pair.
19

403

**Auto-activation tests**

To ensure that our bait collections (in pGBGT7g and pGBKCg) were not acting as auto-
activators, we mated them with yeast carrying empty prey vectors (pGADT7g, pGADCg). Giles
baits gp11, gp28, gp62 and gp11, gp25, gp77 in pGBGT7g and pGBKCg, respectively, showed
weak to strong auto-activation (Suppl. Table S1B). Interaction strength was titrated using
concentrations of 3-amino-triazole (3-AT) between 1 and 100 mM; (3-AT is used as a
competitive inhibitor of the HIS3 enzyme in yeast two hybrid screens, allowing titration of HIS3
eexpression levels and growth resulting from positive results (29, 32). The baits gp11, gp25 and
gp77 showed background growth with empty prey even at 100mM 3-AT concentration and were
screened only in the pGBGT7g Y2H vector system. The remaining auto-activator baits were
screened with the concentration of 3-AT found to minimize background growth.

Once the auto-activator baits were identified, we conducted our bait vs. prey Y2H screens.

**Array-based high-throughput MV-Y2H screening**

Yeast two hybrid (Y2H) array screens were performed as previously described (16). A flow chart
of the procedure is shown in Fig. 1. Briefly, each bait (DBD-X) was mated with each prey (AD-
Y) on rich media (YE PD + adenine) in 384-colony format for 36-48 hours at 30ºC. Diploid cells
were selected for by pinning cultures from mating plates onto selective agar plates (-Leu -Trp)
and growing them for 2-3 days. The diploids were then screened for interacting pairs by pinning
them onto selective screening media (-Leu -Trp -His) and incubating at 30ºC for another 4-7
days. All baits (including self-activating baits) were screened on -Leu -Trp -His plates containing
3-AT to suppress nonspecific background; at least two different 3-AT concentrations between 1
and 100 mM were used for each screen to avoid elimination of true positives. The plates were

19
monitored each day and positive colonies were evaluated with respect to the background growth on each plate. A representative screen is shown in Suppl. Fig. S1.

Completeness of the interactome space was calculated as follows:

The tested space is \((74 \times 73 / 2 = 2,701\) pairs).

The search space is \((77 \times 77 / 2 = 2,964\) pairs).

\[%\] Completeness is \((\text{Tested space}/\text{Entire space}) \times 100\), or

\[
\frac{2701}{2964} \times 100 = 91.12\%.
\]

Filtering and re-testing of raw interaction data

Array-based Y2H screens can reduce the number of false positives (which are the non-reproducible signals that arise by self-activation or because of the “sticky” nature of some preys). We filtered out non-specific raw Y2H data on the basis of prey count with a few exceptions. Prey count is defined as the number of times a defined prey protein is found to be an interacting partner for a bait. The preys found more than 12 times (an arbitrarily-defined value specific to our raw data set only) were predicted to be the result of non-specific interactions and were, with some exceptions, not included in the re-test Y2H data set. A sticky prey was included in the retest data set if it was found to interact specifically and strongly at a 3-AT concentration with no background growth visible on the same plate.

We used the filtered set of raw protein-protein interactions to form a retest set. These interactions were tested as above in 384-colony format in quadruplicate (each colony was plated four times on each plate) for each bait and prey combination in all different vector configurations. Fresh bait and prey arrays were prepared specifically for these retests. All protein-protein interactions were quantitatively titrated against background using a series of different concentrations of 3-AT between 0 and 50mM.
Quantitative assessment of results (% IScore)

A score, % (3-AT)S was calculated for each interacting bait-prey pair using the following formula:

\[
% \text{(3-AT)S} = \left( \frac{C_{\text{PPI}} - C_{B}}{C_{\text{PPI}}} \right) \times 100
\]

where \( % \text{(3-AT)S} \) is the % 3-AT score calculated for each PPI,
\( C_{\text{PPI}} \) is the highest concentration of 3-AT at which a PPI was scored, and
\( C_{B} \) is the concentration of 3-AT at which background was observed.

Thus, each interacting pair was assessed quantitatively and assigned a % (3-AT)S which was used to calculate an overall interaction score (% IScore).

Once we re-tested all PPIs, the % IScore was used to select high confidence PPIs. The % IScore was calculated as follows:

\[
\text{IScore} = \text{3-ATS} + \sum w_k
\]

where: 3-ATS is the 3-AT score assigned to each PPI as above
\( \sum w_k = w_1 + w_2 + w_3 \)

\( w_1 \), the weight value for PPIs detected in multiple vectors, directly proportional to the IScore; \( w_1 = 0 \) if a PPI was detected by only a single vector or 33 if detected by at least 2 vectors.

\( w_2 \), the weight value for reciprocal interactions, also directly proportional to the IScore; \( w_2 = 0 \) if not found in a reciprocal set of interactions (e.g A-B and B-A) or 50 if it is a reciprocal interaction.

\( w_3 \), the weight value for the prey count, inversely proportional to the IScore; \( w_3 = 0, -5, -10, -15, -20, -25, \) or \(-30\) for prey count 1, 2-5, 6-10, 11-15, 16-20, 21-25 or 26-30, respectively.
% IScore = (Actual IScore for a given interacting pair/Highest IScore observed for any interacting pair) X 100. See Suppl. Fig. S5 for a flowchart of this calculation and Suppl. Table S2C for a list of interactions with corresponding IScores.

Giles essential and non-essential proteins

Each Giles protein was assigned an essentiality value based on that determined by Dedrick et al.(12). All proteins determined to be likely essential for the phage lytic cycle, whether by experimental observation or by their role as phage structural components, were labeled ‘essential’. All other gene products were labeled ‘non-essential’.

Excess retention

Excess retention values were calculated as per Wuchty and Almaas 2005 (20). These values, when used in a comparison of essential vs. non-essential nodes in the protein-protein interaction network, correspond to the degree to which essential proteins are over- or under-represented relative to the full network (Suppl. Fig. S5). Values are provided for each k-core, that is, the subnetwork in which all nodes have a degree of at least k. In short, excess retention is defined for a particular k-core as

\[
ER^k = \frac{E^A_k}{E^A}
\]

Where \(E^A_k\) is the fraction of proteins with property A for a k-core of N_k nodes and \(E^A\) is the fraction of proteins with property A for the whole network. Because essentiality of each node in the network has been defined in a binary fashion such that each node is either essential or non-essential, a lower value in one category results in higher values for the other.
Protein interaction networks and bioinformatics analysis

The PPI networks were constructed and analyzed using Cytoscape 3.1 (33), http://www.cytoscape.org/. Giles protein sequences were analyzed further using HHpred secondary structure prediction software (34) and the STRING protein association network tool (26).

Phamerator software (35) was used to create a map for illustration of protein-protein interactions.

Twenty-two Giles proteins have homologs present in phages outside of cluster Q (Mycobacteriophage_Draft database, only looking at non-Draft genomes).

Mass spectrometry of Giles particles and infected *Mycobacterium smegmatis*

Wild-type *Mycobacterium smegmatis* was infected with mycobacteriophage Giles at a multiplicity of infection (MOI) of 3. At 30 minutes and 2.5 hours post-infection, a 1ml aliquot was centrifuged, the supernatant removed, and cell pellet immediately frozen. A high-titer Giles lysate was cesium chloride band purified twice, and then submitted for Mass Spectrometry (MS) analysis along with the 30min and 2.5hr post-infection time points. The Mass Spectrometry was performed by the University of California at Davis Proteomics Core on a LC-MS/MS Q-Exactive as described in Pope *et al.*, 2014 (13). This study refers to three MS fractions: an early fraction (30 min post-infection), a late fraction (2.5 hr post-infection) and the phage particle (whole virion only). Individual proteins may be present in more than one MS fraction.
References


Figure Legends

**Figure 1.** (A) Overview of the bait and prey array construction for Y2H screens. This method includes use of a C-terminal and a N-terminal protein fusions for each bait and prey construct. (B) The size and coverage of the Giles interactome. 91.12% of all possible bait-prey combinations (100%) were screened in this study.

**Figure 2.** (A) Differences between Y2H vectors in dissecting the Giles interactome. The total PPIs detected by each vector pair at different 3-AT concentrations. Percentages represent the contribution of each vector pair to the Giles interactome. (B) Overlapping PPIs detected by different Y2H vector combinations (CC, CN, NN, NC).

**Figure 3.** (A) All binary PPIs in the high-confidence set in the context of transcription. Each black square in the heat map represents an interaction between two proteins, regardless of their role as bait or prey. Dashed lines indicate borders of putative transcriptional units as per Dedrick et al (2013). (B) Number of identified PPI within and between functional groups.

**Figure 4.** (A) The protein-protein interaction network of bacteriophage Giles. Proteins are represented as nodes; node color corresponds to general functional role as noted in the Key. Proteins found or predicted to be essential to the phage lytic cycle are denoted by node shape. Proteins found in any MS fraction of those obtained are denoted by black outlines; those not found are outlined in red. Interactions are represented as edges between nodes. All edges are
representative of high-confidence Y2H results; brighter green edges are those with %IScores approaching a maximum confidence value of 100 while those approaching zero are grey. Red edges are those with %IScores of zero. Solid edges are those found in both Y2H results and between proteins found to be present in the same mass spectrometry (MS) fraction (out of three possible fractions; see Methods for more details). Dashed edges represent predicted interactions found in Y2H results between proteins not found to be present in the same MS fraction.

(B) Protein interactions related to Giles virion structure. Node color corresponds to general functional role as noted in the Key; these groups differ from those in part A. Shapes represent the presence of a particular protein at specific infection stages as shown by MS data.
Table 1. Comparison of bacteriophage interactomes.

<table>
<thead>
<tr>
<th>Phage</th>
<th>Host</th>
<th>Genome size (bp)</th>
<th>ORFs</th>
<th>ORFs tested in Y2H</th>
<th>Final PPIs</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Giles</td>
<td><em>M. smegmatis</em></td>
<td>53,746</td>
<td>77</td>
<td>75</td>
<td>136</td>
<td>This study</td>
</tr>
<tr>
<td>(mycobacteriophage)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lambda</td>
<td><em>E. coli</em></td>
<td>48,490</td>
<td>73</td>
<td>68</td>
<td>93</td>
<td>(16)</td>
</tr>
<tr>
<td>T7</td>
<td><em>E. coli</em></td>
<td>39,937</td>
<td>55</td>
<td>55*</td>
<td>25</td>
<td>(17)</td>
</tr>
<tr>
<td>Dp-1</td>
<td><em>Streptococcus</em></td>
<td>56,506</td>
<td>72</td>
<td>72</td>
<td>156</td>
<td>(15)</td>
</tr>
<tr>
<td>Cp-1</td>
<td><em>Streptococcus pneumoniae</em></td>
<td>19,345</td>
<td>28</td>
<td>28</td>
<td>15</td>
<td>(18)</td>
</tr>
</tbody>
</table>

*This study used libraries of random phage genome fragments rather than full-length ORFs.
Fig. 2

A

3-AT Concentration in mM

B

Venn Diagram

PPIs

3-AT Concentration in mM

0 20 40 60 80 100 120

pGBGT7g
pGADT7g
(CN)

pGBKCg
pGADCg
(CC)

pGBGT7g
pGADT7g
(NN)

pGBKCg
pGADCg
(NC)

26.8%
13.9%
34.9%
24.4%

NC

NN

CC

97
13
1
1

82

40

65

1

2

3

1
Fig. 3

A

B

Key

- No PPI
- 1 PPI
- 2 PPIs
- >2 PPIs

DNA Packaging
Head Assembly
Head-Tail Connector
Integration
Lysis
Tail Assembly
Tail Proteins
Viron
No group

1 4 8 14 2 6 18 30 2 34