Comparative genomics of the mycobacteriophages: insights into bacteriophage evolution
Hatfull GF, Cresawn SG, Hendrix RW (2008). Research in Microbiology 159:332-339
A tour (Part I)

It is essential to realize that this article is not a research article. It is a review article. You can distinguish the two because the latter does not provide a description of how experiments were performed. Take a second to leaf through the article. No Methods section.* Review articles often present the same unavoidable difficulties of research articles – field-specific jargon, unfamiliar concepts – but without the benefit of providing the experiments that might enable you to fight through the thicket to clarity.

Still, review articles have their uses, so long as you don't ask them to do what they cannot do. Here's what they ***can*** do:

- provide an overview of questions and findings that the authors think are important to a specific area of inquiry.
- provide an annotated list of references to research articles. This is often worth the price of admission by itself, saving you loads of time in searching for articles yourself.
- (sometimes) provide a synthesis, combining results from different research articles.
- A very rare subset (e.g. *Scientific American* articles) is aimed at a general audience and tries to help you understand of general concepts.

These are certainly benefits, often sufficient to justify the time required to read a review article. But ***don't*** expect the article to give you an understanding of what is really going on. That's the province of research articles.

As usual, you should approach an article with a goal. What do you hope to find out? Having such a question in mind helps you to choose what to read carefully and what to gloss over. There is not enough time in the world to read and understand every word of every article you come across. Learning what Graham Hatfull and friends might advise you concerning the scope of your group's project is certainly a worthy goal. You probably can think of others. To that end, my first task would be to skim the paper, making note of what sections it has and what pretty pictures it might offer.

1. Introduction

Interesting for deep background. Or if not, that's OK.

2. The mycobacteriophage collection

I gather from the text that there are many mycobacteriophage genomes available, *Mycobacterium smegmatis* was the host used to isolate many of the phage (those of you working on projects that focus in part on phage-host relationships will want to keep this in mind). Apart from that, my attention is taken up primarily by the table (Table 1) that accompanies the text. It's good to know a place where I can find such a table, even if I don't understand half the terms in it. One term,

* Warning: Some misguided journals (most notably *Science* and *Nature*) often disguise their research articles by hiding experimental descriptions in figure legends, reference lists, and on-line supplemental material.

unfortunately undefined, you should become familiar with right away: ‘orfs’. I’ll wait for you to come back from Google.

SQ1. What does ‘orf’ mean:

a. Optical Rotary Force
b. Sound of a barking seal
c. A term of great interest to those analyzing genomes

SQ2. Approximately how many G nucleotides are in the genome of the phage called Halo?

3. Mycobacteriophage viral morphologies

I confess I have a very short attention span when it comes to diddlyviridae. But the time may come when the names take on special meaning for me, so I’ll note that there’s an overview on the subject and move on.

4. Mycobacteriophage genometrics

Sizes.

SQ3. Is there anything in the first paragraph of this section that is inconsistent with Table 1 or not readily evident from it?

GC%.

SQ4. What does GC% mean? Why GC% and not, say, AG%?

The first sentence is quite interesting, suggesting that there is a significant relationship between the GC% of a phage and its host. Why should that be? Does it hold for phage in general?

The third paragraph is also interesting,… I think. It claims that there is a wide range in tRNA gene content of mycobacteriophage.

SQ5. Maybe the difference amongst different phage is just due to their differences in sizes: larger phage have more tRNA. Is size enough to account for the observed difference?

But why should a phage have ***any*** tRNA genes? What use is one without all the rest? Surely their proteins are made from all 20 amino acids! And what use is tRNA without ribosomes? Do the phage have to carry the approximately 100 genes needed to make a ribosome? Of course not! That’s the whole point of being a phage! They rely in general on the molecular apparatus of the host cell. So back to the tRNA… if a phage can use host ribosomes and tRNA to make its protein, why do some carry tRNA genes (and a few carry ***lots***!)? The last sentence of the section “…the reason why they are acquired remains unclear [34]” is wholly inadequate, not scratching the surface of what the cited article (#34 in the reference list) has to say on the subject nor the many other ideas have been put forth to explain the presence of tRNA in phage genomes. Of course, this isn't an article about tRNA. The group concerned with codon bias will want to pursue this matter.

5. Clustering of mycobacteriophage genomes by nucleotide similarity

Members of the group concerned with phylogeny might well have jumped straight to this section and would immediately have been greeted with a provocative figure (Fig. 2) and a provocative conclusion. But what is the connection between the two?

SQ6. Consider Figure 2. What is the relationship between the X axis and the Y axis?

SQ7. The names under the X axis and to the right of the Y axis are not evenly spaced. Why?

SQ8. What do the parallel lines inside the boxes mean? What do the dots mean?

If you have any idea how to answer the last question, then either you've read something on the subject or you have a bright career ahead of you in bioinformatics! (And you still may, even if you couldn't answer the last question). You're given no help in understanding the basis for Fig. 2. You're just left to ooh and aah at the boxes that apparently designate groups of phages. This is a review article, not a research article.

But such plots are sufficiently common that we should spend some time so that you CAN understand what they mean. To do that, pay a visit to the following site, which may prove helpful (http://www.vivo.colostate.edu/molkit/dnadot/). It will take a few seconds for the page to load fully, but when it does, you'll see three boxes near the bottom of the page. This page enables you to make a dotplot (also called dot-matrix) from DNA that you provide.

Try it out! Enter 20 random nucleotides into the left box (DNA Number 1). Don't sit on the A key, I mean 20 ***random*** nucleotides. Having some trouble producing them? That's OK. Humans shouldn't be asked to be random. Get into BioBIKE and pull down the RANDOM-DNA function, selecting the LENGTH-Of option. After executing the function, click the result box at the bottom of the page, highlight the 20 nucleotides, and copy them. Then paste them into the left box. Finally, click the Copy DNA1 --> DNA2 button, and click Make Plot.

SQ9. Compare your graph to that in Figure 2. By analogy, what do you think would be the X and Y axes of your graph?

SQ10. Change the value in the Window Size box from 9 to 5 and click Make Plot again. Then repeat the procedure with Window Sizes of 3 and 1. Ideas on what it all means?

Now that you've seen what the program produces, you might like some enlightenment on how it works. Scroll to the top of the screen and click **Background information on….** With that instruction behind you…

SQ11. Copy/paste the sequence below into both left and right boxes, make sure Window Size is 9, and click Make Plot. Why is the diagonal accompanied by two shorter parallel lines?

```
GGCCACTGCCCAAGGCCACTGCCCAACCCTCCATCATAAAACTTGGGCTTGGG
```

SQ12. Click the first point in the lower parallel line. You'll see in the yellow box above the output box the positions in DNA1 and DNA2 represented by that point (of course DNA1 and DNA2 are identical). Identify the nucleotides at those positions. Move

down the parallel line, clicking as you go on each point. What is the meaning of the parallel line?

SQ13. Consider Figure 2 again, focusing on the phage Bxz1 and Catera (next to each other). The intersection of Bxz1/Catera on the X axis with Bxz1/Catera on the Y axis defines a box with a thick black outline. Inside that box is a diagonal and two parallel lines. What is the significance of the parallel lines? What is the significance that the next two phage (wildcat/barnyard) have a diagonal but no parallel lines?

SQ14. Change the Window Size to 3 and replot. Why the dots? What is the significance of the sentence in the figure legend to Figure 2 that says "*The Dotter output in the lower left triangle reports a more relaxed sequence comparison and the upper right is more stringent.*"?

SQ15. What do the numbers to the left and above the dotplot in Figure 2 signify?

SQ16. What evidence do you have for the authors' statement "*…these are not simply all minor variations of one or two common sequences*"?

SQ17. Take a look at the parallel lines in the box above Che9d in Fig. 2. They appear to be too short. What is the significance of that?

6. Mycobacteriophage genetic mosaicism

Members of the group concerned with mosaicism might well have jumped straight to this section, but if they had, they would have missed an important clue from their answers to SQ17. Similarity amongst phage genomes appears to be segmented: regions of great similarity juxtaposed or interspersed with regions of hardly any similarity (the same sentiment is voiced in the first sentence of this section, differing only in the emphasis on either process or result). How does this come about? The next sentence proposes "horizontal genetic exchange events" as an explanation. Those who pursue this subject probably ought to find out what that term is about. The first paragraph ends with an important point, that comparison of amino acid sequences of proteins might allow you to detect more distant shared ancestry than comparisons of the nucleotide sequences of the genes that encode the proteins.

SQ18. Does that sound familiar? (Think flies) Why is it that protein comparisons should be more sensitive than gene comparisons?

(*to be continued*)