DATA MANAGEMENT PLAN

This project will generate a significant amount of primarily sequence data. In addition, it is likely that new computational tools or scripts will be developed to aid in processing and analyzing these data. We aim to make these data and software available to the broader scientific community as quickly as is possible while assuring a high level of quality.

Biological sequence data

The Illumina platform will be used to generate sequence data from both pure cultures of vibrios and viruses, and from size fractionated daily samples (i.e., metagenomes). All assembled sequence data will be submitted to Genbank (http://www.ncbi.nlm.nih.gov/genbank/) and the CAMERA database. Gene identifiers will be linked to the MicrobesOnline (http://www.microbesonline.org/) database, which provides extensive annotation related to gene structure, function, and phylogeny.

Data archives and integration

While we will submit our data to large, public resources such as Genbank, which enables the construction of linked Bioprojects for the complete set of data generated as part of this project, the NCBI Short Read Archive (metagenomic data), and the CAMERA portal (metagenomic data and genomes), we will also be producing data, such as phage/host interaction networks and time course data (Fig. 1), for which there is no currently available database sufficient for this level of data integration.

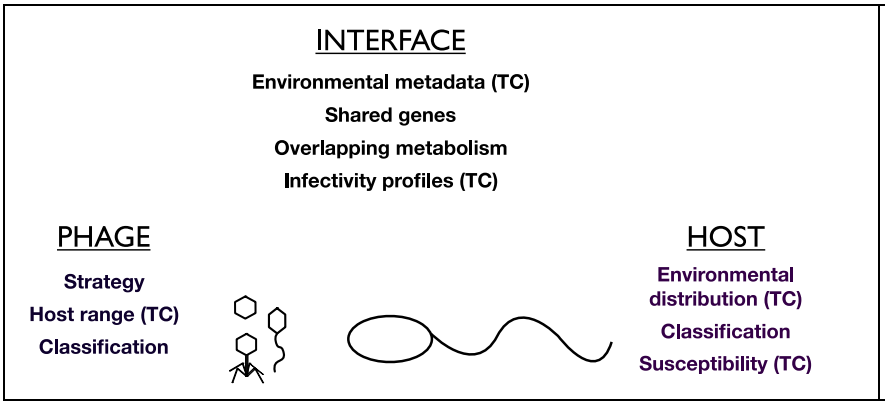

Figure 1: Diverse data sources in the *Vibrio* model system. "TC" indicates time course data is available for a particular source.

We therefore propose to develop a resource that will allow for integration of phage and host genomic data, host range data, and time course data as well as the extensive metadata available for our samples to address the connections between genomics, taxonomy, interactions, and function in our model system. The Kelly lab was previously involved with the development of a public web resource that linked disparate data sources on the cyanobacterium *Prochlocococcus* and its associated phage. The lessons learned from this implementation will be applied to development of an internal database, initially, for the *Vibrio* system.