

## Global Phage Diversity

Ten new mycobacteriophage genomes presented by Pedulla et al. (2003) show that most phage diversity remains uncharacterized. Extrapolation suggests that less than 0.0002% of the global phage metagenome has been sampled. The new genomes also contain a number of potential virulence factors that may be important in pathogenesis.

Phage are the most abundant group of biological entities on the planet. They may also be the most diverse. In this issue of Cell, Pedulla et al. (2003) present 10 new mycobacteriophage genomes. Over 50% of the open reading frames (ORFs) in the genomes are unrelated to anything in GenBank and only one of the new mycobacteriophage is significantly related to a previously sequenced phage. These findings are surprising, because all the new phage belong to the most thoroughly studied group of dsDNA phage, the Siphoviruses.

Many lines of evidence have already suggested that phage diversity is immense. Using culturing techniques, it is relatively easy to find multiple phage types that infect any microbial isolate. For instance, more than 50 phage types infect *E. coli* (Büchen-Osmond, 2002). Furthermore, most phage are host-specific and only infect certain species or even strains of bacteria. Together, these findings strongly suggest that phage are more diverse than their microbial prey, probably by a ratio of >10 phage per microbe. The report by Pedulla et al. (2003), for example, raises the number of sequenced phage that infect *Mycobacterium smegmatis* to 14. Globally there are an estimated 6 million free-living microbial species (Curtis et al., 2002). Moreover, most of 3–5 million species of single- and multi-cellular eukaryotes appear to have specific prokaryotic associates (Novotny et al., 2002). Taken together, there are roughly 10 million free-living and eukaryote-associated microbial species in the world. If each of these microbes is a host for at least 10 different phage, then phage species richness is immense with a predicted 100 million phage species.

High species richness does not necessitate that the global metagenome of phage be large. Relatively small sequence heterogeneities could create a new phage species (e.g., modification of a tail fiber that changes host specificity). This is not what Pedulla et al. (2003) found. Instead, most of the ORFs in the new mycobacteriophage are not similar to anything else previously reported or to each other. Shotgun sequencing of uncultured viral communities from marine water and sediment have also shown that ~75% of the phage sequences are novel (Breitbart et al., 2002). If we assume that all 100 million phage species in the world are 50% unknown, then there are 2.5 billion phage-encoded ORFs yet to be discovered (assuming that each phage encodes 50 ORFs). This ignores the fact that as more phage genomes are sequenced coverage of the metagenome will increase. Therefore, an independent estimation of the global phage metagenome was calculated using the non-parametric estimator Chao1 (Chao, 1984).

For this analysis, all the phage-encoded ORFs in GenBank were compared against every other and grouped together using a BLAST E value cutoff of  $10^{-4}$  (the same parameters used by Pedulla et al., 2003). Using these conditions, Chao1 predicts that 2 billion different phage-encoded ORFs remain to be discovered. The closeness of these two estimates suggests that less than 0.0002% of the global phage metagenome has been sampled.

Pedulla et al. (2003) also found many potential pathogenesis factors in the new mycobacteriophage genomes. Whether these genes are involved in virulence is unknown at this time. There is, however, precedence for phage-encoded genes being important in diseases. Cholera-toxin and most other exotoxins involved in human diseases are encoded by temperate phage (Davis and Waldor, 2002). Phage also carry antibiotic resistance genes and other genes involved in escaping the immune system.

Many major milestones of modern biology were made using phage as models (Cairns et al., 1992; Sanger et al., 1977). The work by Pedulla et al. (2003) shows that phage may also provide the key to understanding global diversity at the species and genome level. Since an average phage genome is only 50 kb long, phage are cheap to sequence and it may even be possible to determine the metagenome of free-living viral communities (Breitbart et al., 2002). Phage genomes also appear to be packed with interesting genes. This new report shows that we are just starting to appreciate phage diversity and how phage influence microbial phenotypes.

### Forest Rohwer

San Diego State University  
Department of Biology  
San Diego, California 92182

### Selected Reading

- Breitbart, M., Salamon, P., Andresen, B., Mahaffy, J.M., Segall, A.M., Mead, D., Azam, F., and Rohwer, F. (2002). Proc. Natl. Acad. Sci. USA 99, 14250–14255.
- Büchen-Osmond, C. (2002). ICTVdb: The Universal Virus Database. <http://www.ncbi.nlm.nih.gov/ICTVdb/Ictv/index.htm>.
- Cairns, J., Stent, G.S., and Watson, J.D. (1992). Phage and the Origins of Molecular Biology (Plainview, NY: Cold Spring Harbor Laboratory Press).
- Chao, A. (1984). Scand J Stat 11, 783–791.
- Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002). Proc. Natl. Acad. Sci. USA 99, 10494–10499.
- Davis, B.M., and Waldor, M.K. (2002). In Mobile DNA II (Washington D.C.: ASM Press), 1040–1059.
- Novotny, V., Basset, Y., Miller, S.E., Weiblen, G.D., Bremer, B., Cizek, L., and Drozd, P. (2002). Nature 416, 841–844.
- Pedulla, M.L., Ford, M.E., Houtz, J.M., Karthikeyan, T., Wadsworth, C., Lewis, J.A., Jacobs-Sera, D., Falbo, J., Gross, J., Pannunzio, N.R., et al. (2003). Cell 113, this issue, 171–182.
- Sanger, F., Air, G.M., Barrell, B.G., Brown, N.L., Coulson, A.R., Fiddes, C.A., Hutchison, C.A., Slocombe, P.M., and Smith, M. (1977). Nature 265, 687–695.