Phages Without Borders: Distribution of Phage Nucleic Acids in 310 Metagenomes

Ramy K. Aziz\textsuperscript{1,2}, Mya Breitbart\textsuperscript{3}, and Robert A. Edwards\textsuperscript{1,4}

\textsuperscript{1}Department of Computer Science, San Diego State University, San Diego, CA, USA; \textsuperscript{2}Department of Microbiology and Immunology, Cairo University, Cairo, Egypt; \textsuperscript{3}University of South Florida, St. Petersburg, FL, USA, \textsuperscript{4}Argonne National Laboratory, Argonne, IL, USA

**ABSTRACT**

Background: Bacteriophages (phages) are the most abundant biological entities on our planet, with an estimated 10\textsuperscript{32} particles on Earth, ten times more than their bacterial host cells. Consequently, phage nucleic acids are expected to constitute a substantial proportion of all Earth’s genomes, both from free-living, lytic phages and from prophages embedded in bacterial and archaeal genomes. Whereas numerous studies have analyzed and estimated the phage abundance and diversity in specific habitats, to date there has not been a systematic analysis of the distribution of phage genes and genomes in different ecosystems.

Methods and Results: Here, we have developed a web tool (Phage Eco-locator, URL: http://www.phantome.org/eco-locator), which screens all publicly available sequenced metagenomes for particular phage sequences. Using this tool, we demonstrate a pervasive, yet uneven distribution of phage genes in metagenomes, reflecting phage abundance and phage module mosaicism. We also demonstrate an expected niche specificity, with phages overrepresented where their bacterial hosts are expected to thrive (e.g., cyanophages and virophages in marine ecosystems, and enterophages in human gut metagenomes). Strikingly, almost none of 550 fully sequenced phage genomes was entirely found in an ecosystem; however, a few phages had most of their genomes represented (e.g., some enteric bacteriophages were almost fully present in human and animal feces.)

Conclusions: We developed a tool that successfully helps trace and visualize phage distribution in ecosystems. Application of this tool to metagenomes from a wide range of environments will enhance our understanding of the factors controlling phage biogeography and environmental selection.

**INTRODUCTION**

• Viruses are the most abundant biological entities on Earth, with estimated population densities of 10\textsuperscript{32} per gram of soil, 10\textsuperscript{15} per ml of seawater, and 10\textsuperscript{12} in the planet. There are approximately 10 times as many viruses as the combined number of all cellular organisms, and most viruses are bacteriophages (phages), those viruses that infect bacteria and archaea.

• A phage either lyses its unicellular host or integrates its genome into the host’s chromosome, where it remains dormant but protected from outside competition and is called a prophage. This so-called lyogenic conversion process implicates that the actual abundance of phage DNA in the environment is even greater than what is calculated from phage particle enumeration because phage nucleic acids can be either packaged in free phage particles, or encoded within bacterial and archaeal chromosomes.

• Although an estimated 20% of nucleotides on Earth are viral in origin, only about 0.03% of nucleotides deposited in GenBank are from phages, and most of these are confined to limited host taxonomic ranges. In examining the biosphere from an anthropocentric perspective, it is easy to focus on the large plants and animals, or on microbes that positively or negatively impact human health and economy, but to dismiss their viruses, except those that infect hosts of obvious interest to humans.

• Metagenomics has become the technology of choice for an unbiased understanding of the biological world, dominated by living forms that remain undetected by culture-based methods. Whereas numerous metagenomic studies have analyzed and estimated the phage abundance and diversity in different habitats, to date there has not been a systematic analysis of the distribution of phage genes and genomes in sequenced Earth’s biomes. We recently found that phage genetic modules are conserved among the 20 most abundant protein-encoding genes in 187 ecosystems. Here, we conduct a systematic, comprehensive analysis of 310 publicly available metagenomic data sets from all over the globe to find the distribution of phages in different habitats. Using a web tool that we built, Phage Eco-locator, we show the abundance, diversity, pervasiveness, and habitat-specificity of about 600 phages, whose complete genomes have been sequenced and made available.

**CONCLUSION**

Current tools allow:

• tracing and visualizing phage distribution in ecosystems,
• quantifying abundance and ubiquity of phage-like sequences in metagenomes

Future development:

1. Developing more sophisticated metrics (e.g., weighted abundance indices) for evaluating phage distribution in ecosystems
2. Estimating the distribution of phage modules and motifs in nature