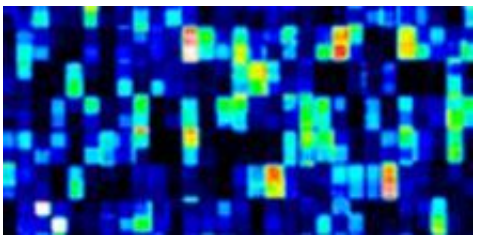
Environmental datasets help researchers double the number of microbial phyla known to be infected by viruses

August 17 2016



The number of microbes in, on, and around the planet - on the order of a nonillion, or 10^{30} - is estimated to outnumber the stars in the Milky Way. Microbes are known to play crucial roles in regulating carbon fixation, as well as maintaining global cycles involving nitrogen, sulfur, and phosphorus and other nutrients, but the majority of them remain uncultured and unknown. The U.S. Department of Energy (DOE) is targeting this "microbial dark matter" to better understand the planet's microbial diversity and glean from nature lessons that can be applied toward energy and environmental challenges.

Plumbing the Earth's microbial diversity, though, requires learning more about the poorly-studied relationships between microbes and the viruses that infect them, viruses that impact the microbes' abilities to regulate global cycles. Although the number of viruses is estimated to be at least

two orders of magnitude more than the microbial cells on the planet, there are currently less than 2,200 sequenced DNA virus genomes, compared to the approximately 50,000 bacterial genomes, in sequence databases. In a study published online August 17, 2016 in *Nature*, researchers at the U.S. Department of Energy Joint Genome Institute (DOE JGI), a DOE Office of Science User Facility, utilized the largest collection of assembled metagenomic datasets from around the world to uncover over 125,000 partial and complete viral genomes, the majority of them infecting microbes. This single effort increases the number of known viral genes by a factor of 16, and provides researchers with a unique resource of viral sequence information.

"It is the first time that someone has looked systematically across all habitats and across such a large compendium of data," said study senior author and DOE JGI Prokaryote Super Program head Nikos Kyrpides. "A key to uncover all these novel viruses was the sensitive computational approach we have developed along this work."

"A key to uncover novel viruses"

That approach, explained first author and postdoctoral fellow David Paez-Espino, involved using a non-targeted metagenomic approach, referencing both isolate viruses and manually curated viral protein models, and what he described as "the largest and most diverse dataset to date." The team analyzed over 5 trillion bases (Terabases or Tb) of sequence available in the DOE JGI's Integrated Microbial Genomes with Microbiome Samples (IMG/M) system collected from 3,042 samples around the world from 10 different habitat types. Their efforts to sift through the veritable haystack of datasets yielded over 125,000 viral sequences containing 2.79 million proteins.

The team matched viral sequences against multiple samples in multiple habitats. For example, one viral group they identified was found in 95

percent of all samples in the ocean's twilight zone - a region located between 200 and 1,000 meters below the ocean surface where insufficient sunlight penetrates for microorganisms to perform photosynthesis.

By analyzing a CRISPR-Cas system - an immune mechanism in bacteria that confers resistance to foreign genetic elements by incorporating short sequences from infecting viruses and phages - the team was able to generate a database of 3.5 million spacer sequences in IMG. These spacers, fragments of phage genetic sequences retained by the host, can then be used to explore viral and phage metagenomes for where the fragments may have originally come from. Also, using mainly this approach, the team computationally identified the host for nearly 10,000 viruses. "The majority of these connections were previously unknown, and include the identification of organisms serving as viral hosts from 16 prokaryotic phyla for which no viruses have previously been identified," they reported.

Beacons for CRISPR-Cas proteins

Jan-Fang Cheng, head of the DOE JGI's Functional Genomics group, said the work being done by Kyrpides' group in identifying new viral sequences will help the Synthetic Biology group develop novel promoters that can work in many bacterial hosts. "We are constantly searching for regulatory DNA parts that will work across many different phyla, and that would allow us to build genes and pathways that can express in many different hosts."

Cheng also anticipated that the expanded viral sequence space generated by Kyrpides' team will allow researchers to look for other genetic sequences known as proto-spacer adjacent motifs (PAMs). These sequences lie next to spacer sequencers in phages and are used as beacons by CRISPR-Cas proteins, triggering actions such as editing or

regulating a gene. "People are looking for new PAM sequences and new Cas9s, and with this new information, if you can map the spacer sequence back to the same phage and align them and see what's in common in neighboring sequences, then you could ID new PAM sequences."

"We believe that the finding of many large phages including the longest phage genome reported thus far points to the limitations of conventional virome enrichment and sequencing strategies which may bias the studies against the highly novel viruses with unusual properties", said Natalia Ivanova, group lead in the Super Program and co-author of this study.

"One of the most important aspects of this study is that we did not focus on a single habitat type. Instead, we explored the global virome and examined the flow of viruses across all ecosystems," said Kyrpides. "We have increased the number of viral sequences by 50x, and 99 percent of the virus families identified are not closely related to any previously sequenced virus. This provides an enormous amount of new data that would be studied in more detail in the years to come. We have more than doubled the number of microbial phyla that serve as hosts to viruses, and have created the first global viral distribution map. The amount of analysis and discoveries that we anticipate will follow this dataset cannot be overstated."

More information: David Paez-Espino et al, Uncovering Earth's virome, *Nature* (2016). DOI: 10.1038/nature19094

Provided by DOE/Joint Genome Institute

https://phys.org/news/2016-08-environmental-datasets-microbial-phyla-infected.html