The ISME Journal (2013) 7, 1678–1695
© 2013 International Society for Microbial Ecology All rights reserved 1751-7362/13
www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes

Pascal Hingamp[1,11], Nigel Grimsley[2], Silvia G Acinas[3], Camille Clerissi[2], Lucie Subirana[2], Julie Poulain[4], Isabel Ferrera[3,12], Hugo Sarmento[3], Emilie Villar[1], Gipsi Lima-Mendez[5,6], Karoline Faust[5,6], Shinichi Sunagawa[7], Jean-Michel Claverie[1], Hervé Moreau[2], Yves Desdevises[2], Peer Bork[7], Jeroen Raes[5,6], Colomban de Vargas[2], Eric Karsenti[7], Stefanie Kandels-Lewis[7], Olivier Jaillon[4], Fabrice Not[2], Stéphane Pesant[8], Patrick Wincker[4] and Hiroyuki Ogata[1,10,11]

[1]CNRS, Aix-Marseille Université, Laboratoire Information Génomique et Structurale (UMR 7256), Marseille, France; [2]CNRS and Université Pierre et Marie Curie (Paris 06), UMR 7144, Station Biologique de Roscoff, Roscoff, France; [3]MARUM—Center for Marine Environmental Sciences, Universität Bremen, Bremen, Germany and [10]Education Academy of Computational Life Sciences, Tokyo Institute of Technology, Tokyo, Japan

Nucleo-cytoplasmic large DNA viruses (NCLDVs) constitute a group of eukaryotic viruses that can have crucial ecological roles in the sea by accelerating the turnover of their unicellular hosts or by causing diseases in animals. To better characterize the diversity, abundance and biogeography of marine NCLDVs, we analyzed 17 metagenomes derived from microbial samples (0.2–1.6 μm size range) collected during the Tara Oceans Expedition. The sample set includes ecosystems under-represented in previous studies, such as the Arabian Sea oxygen minimum zone (OMZ) and Indian Ocean lagoons. By combining computationally derived relative abundance and direct prokaryote cell counts, the abundance of NCLDVs was found to be in the order of 10^6–10^9 genomes ml^{-1} for the samples from the photic zone and 10^8–10^9 genomes ml^{-1} for the OMZ. The Megaviridae and Phycodnaviridae dominated the NCLDV populations in the metagenomes, although most of the reads classified in these families showed large divergence from known viral genomes. Our taxon co-occurrence analysis revealed a potential association between viruses of the Megaviridae family and eukaryotes related to oomycetes. In support of this predicted association, we identified six cases of lateral gene transfer between

remarkably wide taxonomic spectrum from microscopic unicellular eukaryotes to larger animals, including humans. Certain NCLDVs are known to have important roles in marine ecosystems. For instance, *Heterosigma akashiwo* virus (HaV) affects the population dynamics of their unicellular algal host, which forms seasonal harmful blooms in coastal areas (Tomaru *et al.*, 2004). Another well-known virus (*Emiliania huxleyi* viruses (EhV)) controls the population of the ubiquitous haptophyte *E. huxleyi*, which can form vast oceanic blooms at temperate latitudes and exerts complex influence on the carbon cycle (Pagarete *et al.*, 2011). Other NCLDVs cause diseases in fishes and can lead to economic damages in aquaculture industries (Kurita and Nakajima, 2012). NCLDVs include viruses with very large virion particles, which do not pass through 0.2-μm filters typically used in viral metagenomics to separate free viruses from other organisms (Van Etten, 2011). The prototype of such large viruses, also referred to as giruses (Claverie *et al.*, 2006), is the amoeba-infecting *Acanthamoeba polyphaga* Mimivirus with a 0.75-μm virion particle and 1.18-Mb genome (Raoult *et al.*, 2004). Since the discovery of the giant Mimivirus from fresh water samples, NCLDVs have become a subject of broader interest. This has led to several conceptual breakthroughs in our understanding of the origin of viruses and their links to the evolution of cellular organisms (Claverie, 2006; Forterre, 2006; Raoult and Forterre, 2008; Forterre, 2010; Legendre *et al.*, 2012). The sequencing of the Mimivirus genome prompted the discovery of many close homologs in environmental sequence data (Lopez-Bueno *et al.*, 2009; Cantalupo *et al.*, 2011). Most notably, Mimivirus gene homologs were detected in the Global Ocean Sampling (GOS) marine metagenomes (Ghedin and Claverie, 2005; Monier *et al.*, 2008a; Williamson *et al.*, 2008), suggesting Mimivirus relatives exist in the sea. Soon afterwards, two giant viruses related to Mimivirus were isolated from marine environments, These are *Cafeteria roenbergensis* virus (CroV; 750 kb) infecting a major marine microflagellate and *Megavirus chilensis*

The abundance of viruses (*Ostreococcus tauri* virus (OtV)s) infecting the smallest free-living green alga *O. tauri* could vary from undetectable levels to over 10^8 viruses ml^{-1} depending on the season and the distance from the shore (Balloc *et al.*, 2010). The abundance of EhVs could reach over 10^7 viruses ml^{-1} in rapidly expanding host populations in mesocosm experiments simulating host blooms (Schroeder *et al.*, 2003, Pagarete *et al.*, 2011). A typical observation in these studies was an episodic sudden increase (> several orders of magnitude) in virus concentration. These studies focused on specific viral species/strains and depended on the availability of host cultures for lysis evaluation or on relatively simple community compositions amenable to FC analysis. Currently, no direct method is available to assess the abundance of diverse NCLDVs in a complex microbial assemblage dominated by an overwhelming amount of bacterial cells and phages.

To better understand the diversity and geographical distribution of marine NCLDVs, we analyzed a subset of metagenomic sequence data (0.2–1.6 μm size fraction) generated by Tara Oceans, an international multidisciplinary scientific program aiming to characterize ocean plankton diversity, the role of these drifting microorganisms in marine ecosystems and their response to environmental changes (Karsenti *et al.*, 2011). Samples were collected during the first year of the expedition from the Strait of Gibraltar, through the Mediterranean and Red Sea, down to the middle of the Indian Ocean (Table 1). Some marine regions under-represented in previous metagenomic studies are included in this sample set, such as those from the Arabian Sea oxygen minimum zone (OMZ) and Indian Ocean lagoons. Most prokaryotic cells and many large virus particles are expected to be captured within the 0.2–1.6 μm size fraction used in the present metagenome study. Here we show that putative NCLDV sequences differ substantially from known reference genomes, suggesting a high diversity of giant marine viruses. The concentration of NCLDV genomes in the samples was estimated by factoring the metagenome data set with [illegible] determined by FC and micro-