

# Bioinformatics Center – Chemical Life Science –

<http://cls.kuicr.kyoto-u.ac.jp>



Prof  
OGATA, Hiroyuki  
(D Sc)



Assoc Prof  
GOTO, Susumu  
(D Eng)



Assist Prof  
BLANC-MATHIEU, Romain  
(D Sc)



Program-Specific Res  
YOSHIZAWA, Akiyasu  
(D Sc)

## Researchers(pt)

SHIMIZU, Yugo  
YAMAMOTO, Rumiko

## Students

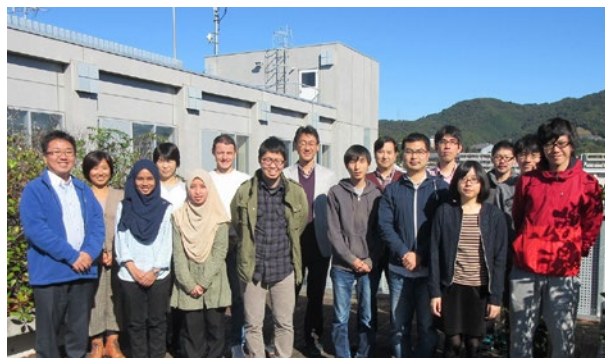
YOSHIKAWA, Genki (M2)      KURONISHI, Megumi (UG)  
NISHIYAMA, Hiroki (M2)      YAMASHITA, Shohei (UG)  
LI, Yanze (M1)                  NURSHAHIRA, Binti Yusuf (UG)  
PRODINGER, Florian (RS)      TUAN, Watie Binti Tuan Mat (UG)  
ARAMAKI, Takuya (UG)

## Scope of Research

We are interested in understanding the functioning and evolution of biological systems at varying scales from tiny microbes up to the Earth's environment, by leveraging rapidly accumulating big data in life science and bioinformatics approaches. We currently focus on 1) the evolution of viruses and their links to the origin of life, 2) microbial ecology in different ecosystems, and 3) the development of bioinformatics methods and biological knowledge resources for biomedical and industrial applications. To fuel these research activities, we take part in environmental sampling campaigns such as *Tara Oceans*. Our resources and developed tools are accessible through GenomeNet ([www.genome.jp](http://www.genome.jp)) to scientific communities and the public.

### KEYWORDS

GenomeNet                      Bioinformatics  
Environmental Genomics      Virology  
Molecular Evolution



## Selected Publications

Blanc-Mathieu, R.; Ogata, H., DNA Repair Genes in the Megavirales Pangenome, *Curr. Opin. Microbiol.*, **31**, 94-100 (2016).  
Mihara, T. *et al.*, Linking Virus Genomes with Host Taxonomy, *Viruses*, **8**, 66 (2016).  
Moriya, Y. *et al.*, Identification of Enzyme Genes Using Chemical Structure Alignments of Substrate-product Pairs, *J. Chem. Inf. Model.*, **56**, 510-516 (2016).  
Guidi, L. *et al.*, Plankton Networks Driving Carbon Export in the Oligotrophic Ocean, *Nature*, **532**, 465-470 (2016).

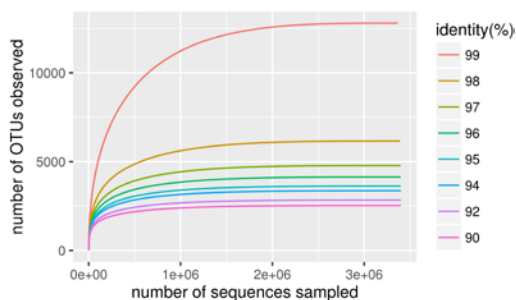
## Diversity of the Giant Virus Family Megaviridae

Megaviridae is a proposed family of eukaryotic viruses classified in the group of nucleocytoplasmic large double-stranded DNA viruses (NCLDVs). They possess large genomes and atypical gene contents as compared to other viruses. Genomic data suggest that the diversity of Megaviridae is vast and perhaps greater than that of cellular organisms. In the present work, we aimed at characterizing the diversity of Megaviridae at a single site using high-throughput sequencing analysis of PCR-amplified DNA polymerase family B (PolB) gene.

PolB was chosen as a marker gene for PCR amplification. PolB is encoded in all available Megaviridae genome sequences. Its high level of divergence makes it impossible to design a single PCR primer pair that would amplify all Megaviridae PolB sequences observed in current metagenomic data. To overcome this limitation we designed 82 degenerate primer pairs that target a conserved domain of PolB based on 923 PolB sequences from isolated Megaviridae and environmental Megaviridae. *In silico* tests demonstrated that this set of primer pairs specifically covers 97.4% of the 923 PolB sequences.

We first experimentally tested a subset of 34 primer pairs on environmental DNA extracted from water samples collected at the sea surface of Osaka bay. PCR-products ( $n=34$ ) were sequenced with Illumina MiSeq to produce paired-end reads. The reads were quality trimmed, merged and clustered by 100% identity threshold. Representative sequences of the clusters were then searched for homology against a reference database containing PolB sequences of cellular organisms and viruses. Then, PolB sequences from all primer pairs were pooled together and aligned. The alignment was trimmed so that the amplicon sequences were well aligned at their both ends. As a result, we obtained 3,400,247 high quality merged reads. All the PolB fragments were predicted to be of Megaviridae origin by a phylogenetic analysis using pplacer.

We clustered all PolB amplicon sequences into operational taxonomic units (OTUs) at 97% identity cutoff. The



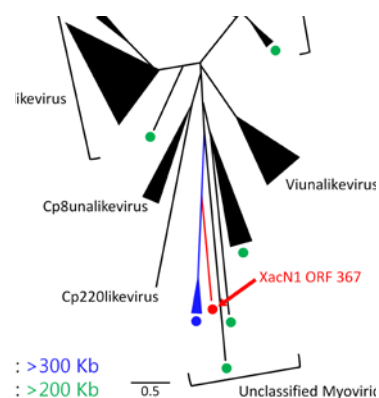
**Figure 1.** Rarefaction curve for the number of OTUs. Singleton OTUs are excluded.

number of OTUs reached near saturation at 5,000 OTUs (Figure 1), suggesting that even a single sea water sample contains diverse viruses of Megaviridae.

## A Characterization of the Genome of XacN1, a Jumbo Phage Infecting the Citrus Canker Agent *Xanthomonas Citri*

There is an increasing interest in the use of phages for controlling plant-pathogenic bacteria in agricultural activities. *Xanthomonas citri* is the bacterial pathogen of citrus canker disease, which is one of the serious citrus plant diseases and leads to significant economic damages worldwide. XacN1 is a bacteriophage that infects *X. citri*. Electron microscopic analysis revealed XacN1 is a large myovirus composed of an icosahedral head of 140 nm in diameter with a 145 nm contractile tail. XacN1 has the second largest double stranded DNA genome (384 kb) among sequenced phages. Its interesting structural feature is the presence of 65 kb direct repeats at the extremities of the linear genome, likely serving for genome replication.

We predicted 592 open reading frames (ORFs) in the XacN1 genome. Homology search revealed that the ORFs include important genes such as genes involved in replication, translation system, and those encoding structural proteins. The genome was found to encode a surprisingly large number of tRNA genes, i.e., 58 tRNAs corresponding to all the twenty amino acids and suppressor tRNAs. The tRNAs showed a significant tendency: codons highly used by the phage were less frequently used by the host bacterium and vice versa. Phylogenetic tree reconstructions of structural proteins indicate that four myoviruses and XacN1 with genome size >300 kb form a clade (Figure 2). Further analysis of the XacN1 genome will lead to uncover yet unidentified unique evolutionary and functional properties of giant phages.



**Figure 2.** Maximum likelihood phylogenetic tree of tail sheath proteins. Blue and green circles indicate phages with a genome greater than 300 kb and 200 kb, respectively.