Nucleotide Sequences of Giant Viruses Found in Soil Samples of the Mojave Desert, the Prairie, the Tundra and the Antarctic Dry Valleys

Csaba Kerepesi,[a], Vince Grolmusz [a,b1]
[a] PIT Bioinformatics Group, Eötvös University,
Pázmány Péter stny. 1/C, H-1117 Budapest, Hungary
[b] Uratim Ltd., H-1118 Budapest, Hungary

Abstract

The first giant virus was identified in 2003 from a biofilm of an industrial water-cooling tower in England. Later, numerous new giant viruses were found in oceans and freshwater reservoirs, some of them having even 2,500 genes. We have demonstrated their very likely presence in four soil samples taken from the Kutch Desert (Gujarat, India). Here we show that numerous other hot and cold desert soil samples as well as some tundra- and forest-soils also contain these viruses. Therefore, giant viruses could be frequent not only in aqueous habitats, but in a wide spectrum of soils on our planet.

1. Introduction

The mere existence of the giant viruses [1, 2, 3, 4, 5, 6, 7] still posts challenges to the definition of life: some authors argue that they should be considered as the members of the "fourth domain of life" [8, 9, 10], while some others are arguing that this is not the case [11, 12]. Nevertheless, the complex interactions of the genes of the amoeba hosts of giant viruses with the viral-, bacterial and eukaryotic-genes may be accounted for the genetic variations of numerous organisms [13, 14, 15].

Because the amoeba hosts of most of these viruses live in aqueous environments, almost all of these viruses were found in ponds, oceans and lakes and industrial water-cooling towers.

By analyzing the metagenomes of the soil samples of the Kutch Desert (Gujarat, India) [16], we have shown the presence of giant viruses in this periodically flooded, salty and hot environment [17]. In the present work we re-analyzed a dataset published with the article [18], describing the soil microbiota of 16 samples of diverse geographic locations, including the North-American prairie, the Chihuahuan- and the Mojave deserts in New Mexico and California, the

[1]to whom correspondence should be addressed

Antarctic dry valleys, the Alaskan tundra, and several forests in tropical and temperate regions. We have found DNA segments of the giant viruses in the samples, implying the very probable presence of giant viruses in these diverse soils, too.

The focus of the work of [18] was the thorough metagenomic analysis of 16 environmental samples for bacteria and archaea, enlightening phylogenetical and functional annotation of the genetic sequences found. No analysis was performed for viral genes.

We applied a two-step sequence search strategy, detailed on Figure 1 and in the Methods section, for finding DNA segments characteristic to giant viruses. We were looking for the DNA segments of nine giant viruses: *A. castellanii mamavirus Hal-V, A. polyphaga mimivirus, A. polyphaga mimivirus M4, A. polyphaga moumouvirus, C. roenbergensis virus BV-PW1, Megavirus chiliensis, Pandoravirus salinus, Pandoravirus dulcis, Pithovirus sibericum isolate P1084-T.*

Our main finding is that each sample collected from the 16 metagenomes [18] contained segments characteristic to at least one and at most six giant viruses from the nine virus genomes examined.

Consequently, we may conclude that giant viruses are, most probably, present in the most diverse soil microbiomes on our planet.

2. Results and discussion

We have examined the metagenomes collected and deposited with the article [18] for the presence of nucleotide sequences characteristic to giant viruses.

The summary of our results is given on Table 1, and the details are given in a large Excel table downloadable from http://uratim.com/Summary.zip.

Samples from the Konza prairie, Kansas and from the Garwood Valley, Antarctica, contained the most giant viral species, while a moist broadleaf forest in Misiones, Argentina and two hot (Mojave and Chihuahuan) one cold (Lake Hoare Valley) deserts the least giant viral species.

Pandoravirus salinus and *Pandoravirus dulcis*, with *A. castellanii mamavirus Hal-V* were the most frequently found species in the samples. We did not find any sequences characteristic to the *Pithovirus sibericum isolate P1084-T* in the metagenomes.

It is surprising that both hot and cold desert soils contain giant viruses; this finding is in line with our previous result concerning the presence of the giant viruses in the soil samples of the Indian Kutch saline desert [17].

3. Methods

The metagenomic data of the article [18] is deposited in the MG-RAST archive:
http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeProject &project=2997 . We downloaded and converted the files into fastq formats.

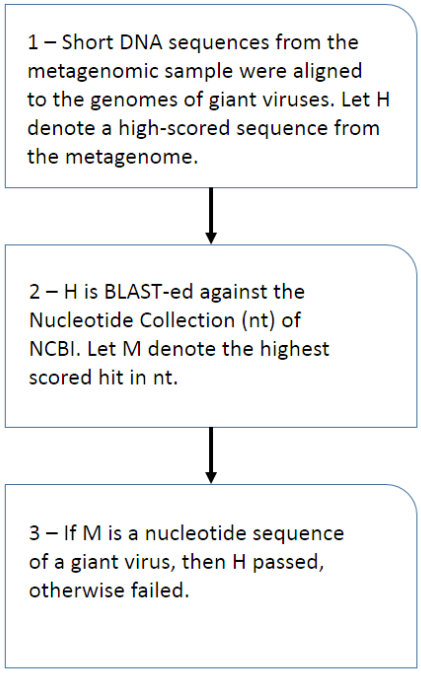

Figure 1: Summary of the two-phase search for nucleotide sequences, characteristic to giant viruses.

Next, with the stand-alone BLAST distribution [19] downloadable `makeblastdb` program we created 16 BLAST databases for each of the 16 metagenomes.

The source of the *A. castellanii mamavirus Hal-V, A. polyphaga mimivirus, A. polyphaga mimivirus M4, A. polyphaga moumouvirus, C. roenbergensis virus BV-PW1, Megavirus chiliensis, Pandoravirus salinus, Pandoravirus dulcis, Pithovirus sibericum isolate P1084-T* virus genomes was the NCBI Genome Database and the EBI Complete Genomes database. The exact sequences applied are as follows:

- ENA|JF801956|JF801956.1 Acanthamoeba castellanii mamavirus strain Hal-V, complete genome
- gi|311977355|ref|NC_014649.1|Acanthamoeba polyphaga mimivirus, complete genome
- ENA|JN036606|JN036606.1 Acanthamoeba polyphaga mimivirus isolate M4, complete genome
- gi|441431943|ref|NC_020104.1|Acanthamoeba polyphaga moumouvirus, complete genome

Source description *	Source location *	No. +	A	B	C	D	E	F	G	H	I
Deserts and xeric shrubland	Lake Bonney Valley, Antarctica	5	+	+	+	.	+	+	.	.	.
Temperate grasslands	Konza Prairie LTER, Kansas, USA	6	+	+	+	+	.	.	+	+	.
Deserts and xeric shrubland	Mojave Desert, California, USA	2	.	.	.	.	+	.	.	+	.
Tropical & subtropical moist broadleaf forest	Manu National Park, Peru	3	.	.	.	+	+	.	+	.	.
Deserts and xeric shrubland	Chihuahuan Desert, Galisteo, NM, USA	2	.	.	.	.	.	.	+	+	.
Deserts and xeric shrubland	Chihuahuan Desert, Sevilleta LTER, NM, USA	4	.	.	.	.	+	+	+	+	.
Tundra	Toolik Lake LTER, Alaska, USA	5	+	+	+	.	.	+	+	.	.
Tropical & subtropical moist broadleaf forest	Misiones, Argentina	1	.	.	.	.	.	.	.	+	.
Temperate coniferous forest	Bonanza Creek LTER, Alaska, USA	4	+	+	+	.	+	.	.	.	.
Temperate coniferous forest	Calhoun Experimental Forest, SC, USA	4	.	.	.	+	+	+	.	+	.
Temperate coniferous forest	Duke Forest, North Carolina, USA	3	+	.	.	.	.	.	+	+	.
Deserts and xeric shrubland	Garwood Valley, Antarctica	6	+	+	+	+	.	+	.	+	.
Deserts and xeric shrubland	Lake Bonney Valley, Antarctica	2	.	.	.	+	.	.	+	.	.
Deserts and xeric shrubland	Lake Fryxell Valley, Antarctica	5	+	+	.	+	.	+	+	.	.
Deserts and xeric shrubland	Lake Hoare Valley, Antarctica	2	.	.	.	.	.	.	+	+	.
Deserts and xeric shrubland	Wright Valley, Antarctica	4	+	.	.	+	.	.	+	+	.
		No. +	8	6	5	7	6	6	10	10	0

Table 1: The short summary of the results of the giant viral genome searches. The description of the sources follows the wording of [18]. The lettered columns correspond to the viruses as follows: A: *A. castellanii mamavirus* Hal-V, B: *A. polyphaga mimivirus*, C: *A. polyphaga mimivirus* M4, D: *A. polyphaga moumouvirus*, E: *C. roenbergensis* virus BV-PW1, F: *Megavirus chiliensis*, G: *Pandoravirus salinus*, H: *Pandoravirus dulcis*, I: *Pithovirus sibericum* isolate P1084-T. Plus sign means that at least one nucleotide sequence passed the conditions described in the Methods section that is, the presence of the giant virus is very probable in the sample.

- gi|310830989|ref|NC_014637.1|Cafeteria roenbergensis virus BV-PW1, complete genome
- gi|363539767|ref|NC_016072.1|Megavirus chiliensis, complete genome
- gi|531034792|ref|NC_022098.1|Pandoravirus salinus, complete genome
- gi|526118633|ref|NC_021858.1|Pandoravirus dulcis, complete genome
- gi|585299329|ref|NC_023423.1|Pithovirus sibericum isolate P1084-T, complete genome

The stand-alone program blastn was applied for aligning the nine giant virus genomes against the 16 BLAST databases, built from the metagenomic data. E-value was set to 0.01, all the other parameters and the scores and penalties were default for blastn. This step corresponds to the first box of Figure 1.

Next, the hits with better E-value than 0.01 were collected from each alignment, and were re-BLAST-ed against the whole Nucleotide Collection (nt) of the NCBI. This step corresponds to the second box on Figure 1.

In the evaluation step (corresponding to the third box on Figure 1) we have thrown out those sequence-hits from the metagenomes that could be aligned better to some non-giant viral genome from the nt collection than to the genome of some giant virus. Only those hits were passed that were specific to giant viruses.

As an example for the evaluation step let us consider the metagenome with the MG Rast accession number mgm4477803.3.050, corresponding to a soil sample from the cold desert of Lake Bonney Valley, Antarctica. In phase 1, we were searching for subsequences, similar to the genomes of the nine giant viruses. When we searched for similar sequences to the genome of *Megavirus chiliensis*, the best score was assigned to the subsequence HWI-EAS137R_0379:6:10:13644:6752#CTTGTA/1, with score of 78.7 and E-value of 1e-09 (c.f. the results at http://uratim.com/giant_oda-blast.zip). However, in phase 2, when we were examining specificity, it turned out that a genomic sequence from the cyanobacterium *Oscillatoria nigro-viridis* PCC 7112, has much better alignment than any giant virus examined (the E-value was 3e-22). Therefore, the sequence HWI-EAS137R_0379:6:10:13644:6752#CTTGTA/1 was discarded from the hit list. Another sequence from the metagenome, the HWI-EAS137R_0379:6:26:15491:12589#CTTGTA/1 passed, since from the NCBI Nucleotide Collection (nt) the best similarity score was attained by a *Megavirus courdo11* and a *Megavirus chiliensis* sequence, both are giant viruses (c.f. the table at http://uratim.com/giant_vissza-blast.zip).

Data availability: The metagenomes of the article [18] can be downloaded from:
http://metagenomics.anl.gov/metagenomics.cgi?page=MetagenomeProject
&project=2997 The results of the individual alignments (with nucleotide sequences, their alignment, scores and E-values can be downloaded from http://uratim.com/giant_oda-blast.zip (for the results of phase 1) and from http://uratim.com/giant_vissza-blast.zip (for the results of phase 2). The Excel table with the detailed summary of the results can be found at http://uratim.com/Summary.zip.

4. Conclusions

We have shown the very probable presence of giant viruses in diverse environmental soil samples by a two-phase search strategy in metagenomic samples and the NCBI Nucleotide Collection (nt). Consequently, such non-aqueous environments as Antarctic dry valleys, the Mojave desert, the prairie and several forest-soils most probably also contain these recently discovered viruses.

5. References

References

[1] Didier Raoult, Stephane Audic, Catherine Robert, Chantal Abergel, Patricia Renesto, Hiroyuki Ogata, Bernard La Scola, Marie Suzan, and Jean-Michel Claverie. The 1.2-megabase genome sequence of Mimivirus. *Science*, 306(5700):1344–1350, Nov 2004.

[2] Philippe Colson, Natalya Yutin, Svetlana A Shabalina, Catherine Robert, Ghislain Fournous, Bernard La Scola, Didier Raoult, and Eugene V Koonin. Viruses with more than 1,000 genes: Mamavirus, a new Acanthamoeba polyphaga mimivirus strain, and reannotation of Mimivirus genes. *Genome Biol Evol*, 3:737–742, 2011.

[3] Sheree Yau, Federico M Lauro, Matthew Z DeMaere, Mark V Brown, Torsten Thomas, Mark J Raftery, Cynthia Andrews-Pfannkoch, Matthew Lewis, Jeffrey M Hoffman, John A Gibson, and Ricardo Cavicchioli. Virophage control of antarctic algal host-virus dynamics. *Proc Natl Acad Sci U S A*, 108(15):6163–6168, Apr 2011.

[4] Mickaël Boyer, Natalya Yutin, Isabelle Pagnier, Lina Barrassi, Ghislain Fournous, Leon Espinosa, Catherine Robert, Saïd Azza, Siyang Sun, Michael G. Rossmann, Marie Suzan-Monti, Bernard La Scola, Eugene V. Koonin, and Didier Raoult. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. *Proc Natl Acad Sci U S A*, 106(51):21848–21853, Dec 2009.

[5] D Randy Garza and Curtis A Suttle. Large double-stranded dna viruses which cause the lysis of a marine heterotrophic nanoflagellate (bodo sp.) occur in natural marine viral communities. *Aquatic Microbial Ecology*, 9(3):203–210, 1995.

[6] Matthias G Fischer, Michael J Allen, William H Wilson, and Curtis A Suttle. Giant virus with a remarkable complement of genes infects marine zooplankton. *Proc Natl Acad Sci U S A*, 107(45):19508–19513, Nov 2010.

[7] Nadège Philippe, Matthieu Legendre, Gabriel Doutre, Yohann Couté, Olivier Poirot, Magali Lescot, Defne Arslan, Virginie Seltzer, Lionel Bertaux, Christophe Bruley, Jérome Garin, Jean-Michel Claverie, and Chantal Abergel. Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, 341(6143):281–286, Jul 2013.

[8] Jean-Michel Claverie, Hiroyuki Ogata, Stéphane Audic, Chantal Abergel, Karsten Suhre, and Pierre-Edouard Fournier. Mimivirus and the emerging concept of "giant" virus. *Virus research*, 117(1):133–144, 2006.

[9] Philippe Colson, Xavier de Lamballerie, Ghislain Fournous, and Didier Raoult. Reclassification of giant viruses composing a fourth domain of life in the new order megavirales. *Intervirology*, 55(5):321–332, 2011.

[10] Philippe Colson, Gregory Gimenez, Mickaël Boyer, Ghislain Fournous, and Didier Raoult. The giant cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of life. *PLoS One*, 6(4):e18935, 2011.

[11] Matthieu Legendre, Defne Arslan, Chantal Abergel, and Jean-Michel Claverie. Genomics of megavirus and the elusive fourth domain of life. *Communicative & integrative biology*, 5(1):102–106, 2012.

[12] Tom A Williams, T Martin Embley, and Eva Heinz. Informational gene phylogenies do not support a fourth domain of life for nucleocytoplasmic large dna viruses. *PLoS One*, 6(6):e21080, 2011.

[13] Matthias G Fischer and Curtis A Suttle. A virophage at the origin of large dna transposons. *Science*, 332(6026):231–234, 2011.

[14] Christelle Desnues, Bernard La Scola, Natalya Yutin, Ghislain Fournous, Catherine Robert, Saïd Azza, Priscilla Jardot, Sonia Monteil, Angélique Campocasso, Eugene V Koonin, et al. Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proceedings of the National Academy of Sciences*, 109(44):18078–18083, 2012.

[15] Pierre-Alain Jachiet, Philippe Colson, Philippe Lopez, and Eric Bapteste. Extensive gene remodeling in the viral world: new evidence for nongradual evolution in the mobilome network. *Genome biology and evolution*, 6(9):2195–2205, 2014.

[16] A. S. Pandit, M. N. Joshi, P. Bhargava, G. N. Ayachit, I. M. Shaikh, Z. M. Saiyed, A. K. Saxena, and S. B. Bagatharia. Metagenomes from the saline desert of Kutch. *Genome Announc*, 2(3), 2014.

[17] Csaba Kerepesi and Vince Grolmusz. Giant viruses of the kutch desert. *arXiv preprint arXiv:1410.1278*, 2014.

[18] Noah Fierer, Jonathan W. Leff, Byron J. Adams, Uffe N. Nielsen, Scott Thomas Bates, Christian L. Lauber, Sarah Owens, Jack A. Gilbert, Diana H. Wall, and J Gregory Caporaso. Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci U S A*, 109(52):21390–21395, Dec 2012.

[19] Stephen F. Altschul, John C. Wootton, E Michael Gertz, Richa Agarwala, Aleksandr Morgulis, Alejandro A. Schaffer, and Yi-Kuo Yu. Protein database searches using compositionally adjusted substitution matrices. *FEBS J*, 272(20):5101–5109, Oct 2005.