Supporting Information

Arslan et al. 10.1073/pnas.1110889108

SI Materials and Methods

Virus Isolation. Megavirus chilensis was isolated from coastal waters in front of the ECIM marine station from Las Cruces, Chile. One liter of seawater was supplemented with 4% of rice media (supernatant obtained after autoclaving 1 L of seawater with 40 grains of rice) and let to incubate for 1 mo in the dark at room temperature. The rationale of such procedure is to get rid of the phototrophic microorganisms while allowing the heterotrophic bacteria to grow for a while, when they then feed the phagocytic/heterotrophic protozoans that finally expand to a population allowing eventual viruses to multiply (1). Seawater with rice medium was then filtered first through a polycarbonate Isopore membrane filter of 1.2-μm pore size and then through 0.2-μm pore size membrane filter (RTTP04700, GTTP04700; Millipore). The 0.2-μm pore size membrane was then treated with gentamicin at 1 mg/mL final concentration, 10% penicillin/streptomycin and 5% fungizone for 3 d. Supernatant was inoculated to several acanthamoeba species cultured in microplates and monitored for cell lysis.

Giant Virus Naming. We believe it is useful and desirable that the name of a newly isolated microorganism convey some of its most distinctive properties. After the initial naming of Mimivirus (for "microbe mimicking"), already not a very good name because the prefix "mimi" does not convey a helpful scientific notion, newly isolated related viruses are receiving increasingly random/funny names such as "Mamavirus," "Moumouvirus," "Courdovirus," and "Terra" (2). Although it is traditionally the privilege of the first authors describing a new microbe to give it whatever name of their choosing, we believe the current trend is counterproductive and should give way to more informative names. With the few examples now at hand, it is clear that a distinctive feature of the above giant viruses (or of their close ancestors) is to possess genome in excess of a "megabase". Hence, the term "Megavirus," and the proposed family/genus "*Megaviridae*" that will be proposed to the International Committee on Taxonomy of Viruses. "Chilensis" then refers to the location where this virus was first isolated. Finally, we broke with tradition not incorporating the host's species to the virus name. This decision is justified by the fact that Megavirus and other Mimivirus relatives are capable of replicating in a variety of acanthamoeba species, whereas the phagocytic protozoan that is the natural host of *Megavirus chilensis* is not known, as will be the situation for most viruses isolated from the environment using the acanthamoeba coculture protocol.

Genome Assembly. The Megavirus genome was assembled by using a combination of 454-titanium and Illumina Hiseq paired-end reads. We first assembled the 42,288,396 Hiseq paired-end reads by using the Velvet assembler (3) with the following parameters: $k = 95$, ins_length = 280, cov_cutoff = 200 and exp_cov = 382. We next mapped the 278,663 454-titanium reads onto the assembled contigs by using Mira (4) to extend them. Gap5 (5) software was used to join the resulting overlapping contigs into a single one. We finally remapped the Hiseq reads at high stringency to correct sequencing errors. The 454 technology generated a large number of local errors due to the miscalling of homopolymeric sequences in the Megavirus A+T rich genome. Steep drops in the Illumina read coverage were used to guide the visual inspection of the sequence and its manual correction (usually a single A or T nucleotide insertion or deletion). The total Illumina data used for this finishing step corresponds to 1/10th of a flow cell channel used in a multiplexed fashion with nine other unrelated sequencing projects. A few positions were confirmed by PCR followed by Sanger sequencing. The final Megavirus genome sequence corresponds to a single 1,259,197-nt-long contig.

Gene Annotation. The Megavirus protein coding regions (CDSs) were identified by using the GeneMarkS algorithm (6). Transfer RNAs were searched by using tRNAscan-SE (7) with the general tRNA model. The functional assignment of these predicted Megavirus genes was performed by using a combination of BlastP searches against public databases using an *e* value threshold of 10^{-5} and protein motif identification using Interproscan (8). Megavirus/Mimivirus orthologous gene pairs were defined based on the best-reciprocal blast hit criterion between the two proteomes, again using BlastP at an *e* value threshold of 10^{-5}. Megavirus (respectively Mimivirus) "paralogues" correspond to predicted proteins exhibiting BlastP similarity within the Mimivirus (respectively Megavirus) proteome at the same threshold but failing the reciprocal best match criterion. These correspond to Megavirus/Mimivirus specific gene duplications. The last category of CDSs, specific to each virus, corresponds to those not exhibiting a BlastP hit at the conservative *e* value threshold of 10^{-5}.

Phylogenetic Analysis. The most similar homologs of Megavirus aminoacyl tRNA synthetases were first identified by using the Blast-Explorer tool (9) on the Phylogeny.fr (10) server. A subset of sequences was selected based on the alignment quality (preserving enough informative positions) and their phylogenetic distribution among the main domains (Archaea, Eukarya, Eubacteria). An optimal multiple alignment was then computed by using MAFFT version 6 (11) on the CBRC-AIST server (mafft.cbrc.jp/alignment/server/). Several trees were then reconstructed from this alignment by using a simple neighbor-joining algorithm (with the JTT model) or PhyML (with the WAG model) (10). The topology of the reconstructed trees and the confidence values were very similar for both methods.

1. Massana R, del Campo J, Dinter C, Sommaruga R (2007) Crash of a population of the marine heterotrophic flagellate Cafeteria roenbergensis by viral infection. *Environ Microbiol* 9:2660–2669.
2. La Scola B, et al. (2010) Tentative characterization of new environmental giant viruses by MALDI-TOF mass spectrometry. *Intervirology* 53:344–353.
3. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829.
4. Chevreux B, Wetter T, Suhai S (1999) Genome sequence assembly using trace signals and additional sequence information. Computer science and biology. *Proceedings of the German Conference on Bioinformatics* 99:45–56.
5. Bonfield JK, Whitwham A (2010) Gap5—editing the billion fragment sequence assembly. *Bioinformatics* 26:1699–1703.
6. Besemer J, Lomsadze A, Borodovsky M (2001) GeneMarkS: A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29:2607–2618.
7. Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33(Web Server issue):W686–W689.
8. Hunter S, et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res* 37(Database issue):D211–D215.
9. Dereeper A, Audic S, Claverie JM, Blanc G (2010) BLAST-EXPLORER helps you building datasets for phylogenetic analysis. *BMC Evol Biol* 10:8.
10. Dereeper A, et al. (2008) Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 36(Web Server issue):W465–W469.
11. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9:286–298.

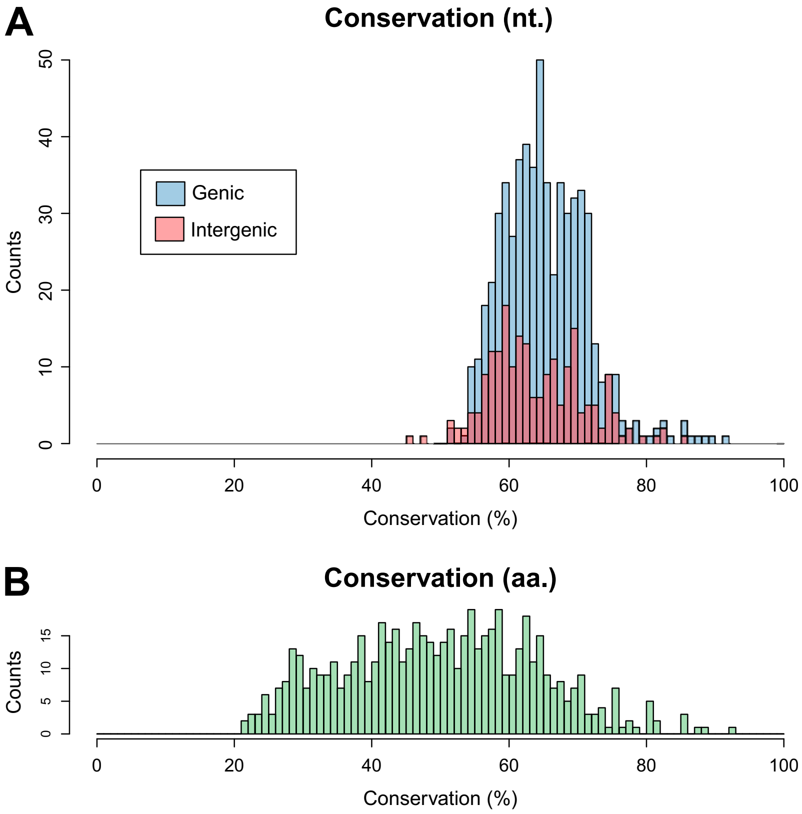


Fig. S1. Sequence divergence between Megavirus and Mimivirus. (*A*) Orthologous genes and intergenic nucleotide sequences. (*B*) Orthologous protein sequences. Notice that the nucleotide sequences exhibit more similarity than the amino acid sequences, in part due to the large nucleotide composition bias (75% A+T).

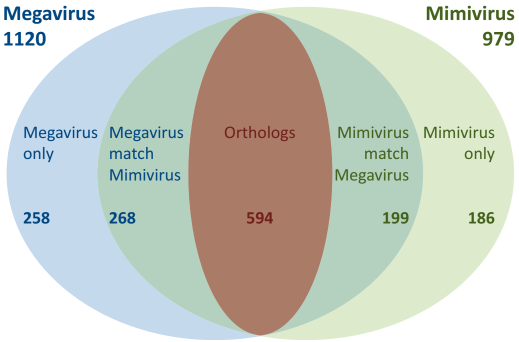


Fig. S2. Comparison of Mimivirus and Megavirus gene contents. The various subsets in this Venn diagram are not represented to scale.

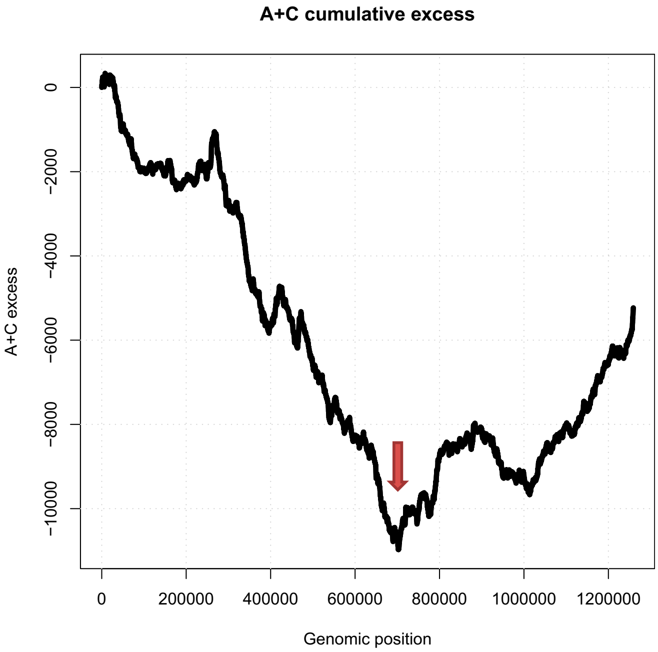


Fig. S3. A+C excess profile of the Megavirus genome. The slope reversal (red arrow) approximately coincides with one of the boundaries of the large inverted segment disrupting the colinearity between the Megavirus and Mimivirus genomes (Fig. 2).

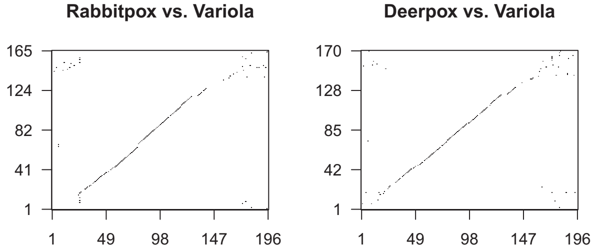
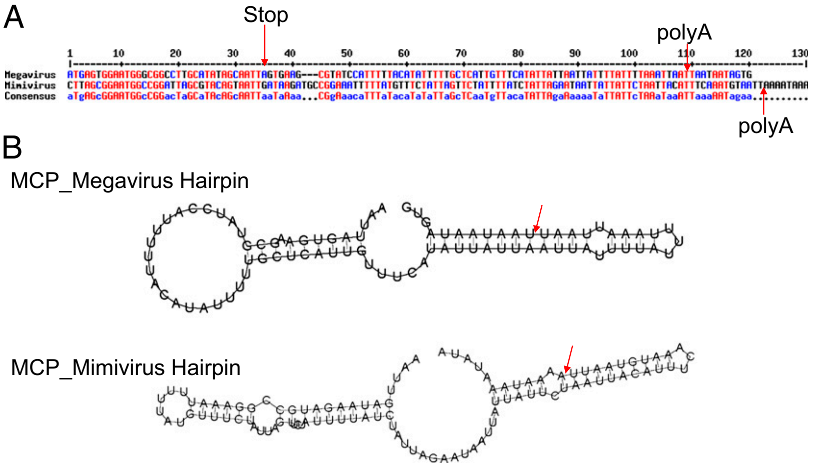


Fig. S4. Diagonal similarity plots (dot plots) of pairs of poxvirus genomes. These pairs of viruses exhibit global sequence similarity levels comparable to the one exhibited by the Mimivirus/Megavirus pair (DNA polymerase sharing ≈65% identical residues). The colinearity is conserved in the central region of the genomes and abruptly vanishes at both ends of the chromosomes.



```
           1        10        20        30        40        50        60        70        80        90        100       110       120       130
Megavirus  ATGAGTGGAATGGGCGGCCTTGCATATAGCAATTAGTGAAG---CGTATCCATTTTTACATATTTTTGCTCATTGTTTCATATTATTAATTATTTTATTTTAAATTAATTAATAATAGTG
Mimivirus  CTTAGCGGAATGGCCGGATTAGCGTACAGTAATTGATAAGATGCCGGAAATTTTTATGTTTCTATTAGTTCTATTTTATCTATTAGAATAATTATTATTCTAATTACATTTCAAATGTAATTAAAATAAA
Consensus  aTgAGcGGAATGGcCGGacTaGCaTAcAGcAATTaaTaAaa...CGgAaacaTTTaTacaTaTaTTaGcTCaaTgTTacaTATTAgaAaaaaTaTTATTcTAAaTaaATTaaaAATagaa..........
```

Fig. S5. (*A*) Optimal alignment of the 3′ UTR regions of the major capsid protein transcripts in Mimivirus and Megavirus. The two sequences only share 49% identical nucleotides. (*B*) Predicted hairpin structures and experimentally validated polyadenylation sites (red arrows).

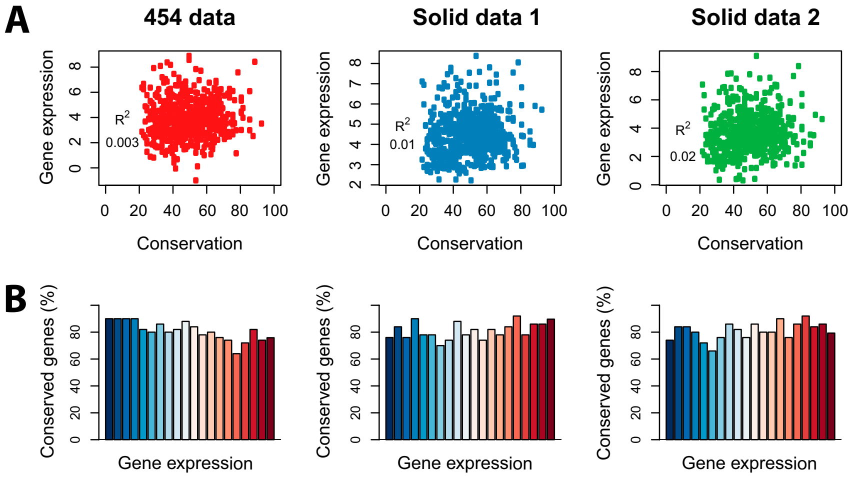


Fig. S6. Absence of correlation between the level of expression of Mimivirus genes and their conservation in Megavirus. (*A*) Shown is gene expression level (log scale) vs. % of identical residues between orthologs in three independent transcriptome datasets. Mimivirus gene expression (red; *Left*) was measured based on 454 mRNA sequence reads (1) and a total RNA dataset by using Solid sequencing technology (green; *Right*) (2). A third Solid dataset (finer grid through the infection cycle) with 196 million reads from total RNA is also shown (blue; *Center*). (*B*) Percentage of genes with a Megavirus ortholog vs. their expression level in Mimivirus distributed in 20 bins from the lowest (blue) to the highest (red).

1. Legendre M, et al. (2010) mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res* 20:664–674.

2. Legendre M, Santini S, Rico A, Abergel C, Claverie JM (2011) Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing. *Virol J* 8:99.

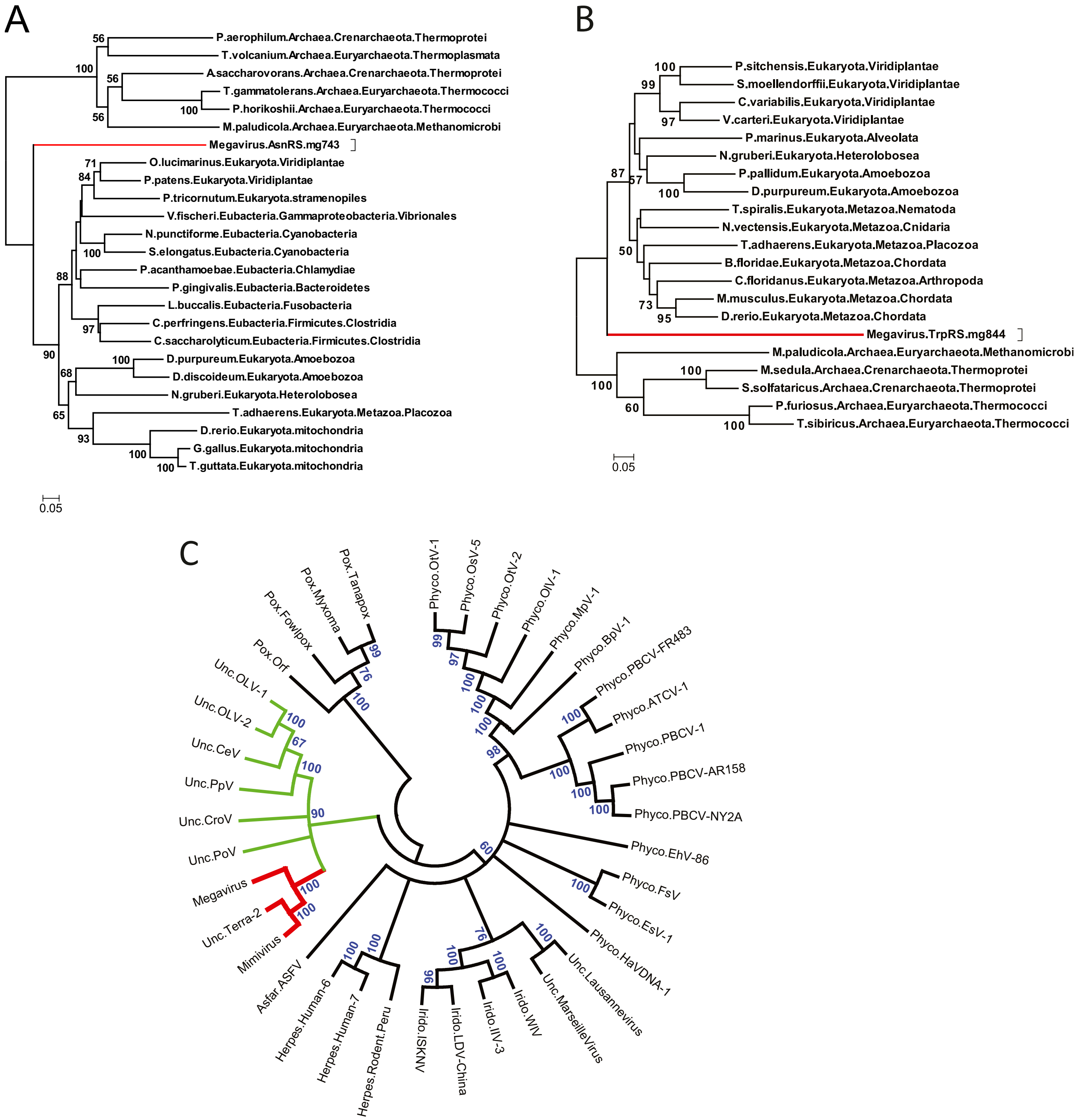

A
P.aerophilum.Archaea.Crenarchaeota.Thermoprotei
T.volcanium.Archaea.Euryarchaeota.Thermoplasmata
A.saccharovorans.Archaea.Crenarchaeota.Thermoprotei
T.gammatolerans.Archaea.Euryarchaeota.Thermococci
P.horikoshii.Archaea.Euryarchaeota.Thermococci
M.paludicola.Archaea.Euryarchaeota.Methanomicrobi
Megavirus.AsnRS.mg743
O.lucimarinus.Eukaryota.Viridiplantae
P.patens.Eukaryota.Viridiplantae
P.tricornutum.Eukaryota.stramenopiles
V.fischeri.Eubacteria.Gammaproteobacteria.Vibrionales
N.punctiforme.Eubacteria.Cyanobacteria
S.elongatus.Eubacteria.Cyanobacteria
P.acanthamoebae.Eubacteria.Chlamydiae
P.gingivalis.Eubacteria.Bacteroidetes
L.buccalis.Eubacteria.Fusobacteria
C.perfringens.Eubacteria.Firmicutes.Clostridia
C.saccharolyticum.Eubacteria.Firmicutes.Clostridia
D.purpureum.Eukaryota.Amoebozoa
D.discoideum.Eukaryota.Amoebozoa
N.gruberi.Eukaryota.Heterolobosea
T.adhaerens.Eukaryota.Metazoa.Placozoa
D.rerio.Eukaryota.mitochondria
G.gallus.Eukaryota.mitochondria
T.guttata.Eukaryota.mitochondria
0.05
B
P.sitchensis.Eukaryota.Viridiplantae
S.moellendorffii.Eukaryota.Viridiplantae
C.variabilis.Eukaryota.Viridiplantae
V.carteri.Eukaryota.Viridiplantae
P.marinus.Eukaryota.Alveolata
N.gruberi.Eukaryota.Heterolobosea
P.pallidum.Eukaryota.Amoebozoa
D.purpureum.Eukaryota.Amoebozoa
T.spiralis.Eukaryota.Metazoa.Nematoda
N.vectensis.Eukaryota.Metazoa.Cnidaria
T.adhaerens.Eukaryota.Metazoa.Placozoa
B.floridae.Eukaryota.Metazoa.Chordata
C.floridanus.Eukaryota.Metazoa.Arthropoda
M.musculus.Eukaryota.Metazoa.Chordata
D.rerio.Eukaryota.Metazoa.Chordata
Megavirus.TrpRS.mg844
M.paludicola.Archaea.Euryarchaeota.Methanomicrobi
M.sedula.Archaea.Crenarchaeota.Thermoprotei
S.solfataricus.Archaea.Crenarchaeota.Thermoprotei
P.furiosus.Archaea.Euryarchaeota.Thermococci
T.sibiricus.Archaea.Euryarchaeota.Thermococci
0.05
C
Pox.Tanapox
Pox.Myxoma
Pox.Fowlpox
Pox.Orf
Unc.OLV-1
Unc.OLV-2
Unc.CeV
Unc.PpV
Unc.CroV
Unc.PoV
Megavirus
Unc.Terra-2
Mimivirus
Asfar.ASFV
Herpes.Human-6
Herpes.Human-7
Herpes.Rodent.Peru
Irido.ISKNV
Irido.LDV-China
Irido.IIV-3
Irido.WIV
Unc.MarseilleVirus
Unc.Lausannevirus
Phyco.HaVDNA-1
Phyco.EsV-1
Phyco.FsV
Phyco.EhV-86
Phyco.PBCV-NY2A
Phyco.PBCV-AR158
Phyco.PBCV-1
Phyco.ATCV-1
Phyco.PBCV-FR483
Phyco.BpV-1
Phyco.MpV-1
Phyco.OlV-1
Phyco.OtV-2
Phyco.OsV-5
Phyco.OtV-1

Fig. S7. Phylogenetic analysis of selected megavirus protein sequences. All alignments were computed by using the default option of the MAFFT server (1). (*A*) AsnRS (mg743): This neighbor-joining tree was computed from the 300 conserved positions. The tree is rooted on the Archaea branch. The nodes are labeled with their bootstrap values when >50. A very similar tree was computed by using the default option of the Phylogeny.fr server (alignment with MUSCLE and tree reconstruction by PhyML) (2). Despite being one of the least canonical of the AARS (sensu Woese et al.; see ref. 3), the Megavirus AsnRS nicely separates the archeal enzymes from all of the eukaryotic (including mitochondrial) types known to intermix with bacterial enzymes. (*B*) TrpRSs (mg844): This PhyML tree (rooted on the Archaea branch) was computed on the Phylogeny.fr server (2) from 327 conserved positions. The nodes are labeled with their bootstrap values when >50. A similar tree was computed by using the default option of the MAFFT server (tree reconstruction by neighbor joining). The Megavirus TrpRS is branching off the eukaryotic domain before the radiation of all clades. (*C*) DNA polymerase (mg582). This neighbor-joining tree was computed from the 523 ungapped positions of an alignment of 38 DNA polymerases sequences from the main large DNA virus families: Poxviridae, Iridoviridae, Herpesviridae, Asfarviridae, Marseilleviridae, and Phycodnaviridae. Megavirus, Mimivirus, and Terra-2 form a tight cluster (in red) within a larger well supported group of unclassified (unc) aquatic viruses including those with the largest known genome sizes. This group (red and green) shares a number of distinctive features and is proposed to constitute a new family: the Megaviridae. PoV, *Pyramimonas orientalis* virus (560 kb); CroV, *Cafeteria roenbergensis* virus (730 kb); PpV, *Phaeocystis pouchetii* virus (485 kb); CeV, *Chrysochromulina ericina* virus (510 kb); OLV, Organic Lake virus (1 and 2); HaVDNA, *Heterosigma akashiwo* DNA virus; PBCV, *Paramecium bursaria* Chlorella virus; ATCV, *Acanthocystis turfacea* Chlorella virus; BpV, *Bathycoccus* sp. RCC1105 virus; OsV, *Ostreococcus* virus;

Legend continued on following page

OtV, *Ostreococcus tauri* virus; MpV, *Micromonas* sp. RCC1109 virus; OlV, *Ostreococcus lucimarinus* virus; EhV, *Emiliania huxleyi* virus; FsV, *Feldmannia* species virus; EsV, *Ectocarpus siliculosus* virus; WIV, *Wiseana* iridescent virus; IIV, Invertebrate iridescent virus; LDV, Lymphocystis disease virus; ISKNV, Infectious spleen and kidney necrosis virus; ASFV, African swine fever virus.

1. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform* 9:286–298.
2. Dereeper A, et al. (2008) Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res* 36(Web Server issue):W465–W469.
3. Woese CR, Olsen GJ, Ibba M, Söll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol Mol Biol Rev* 64:202–236.

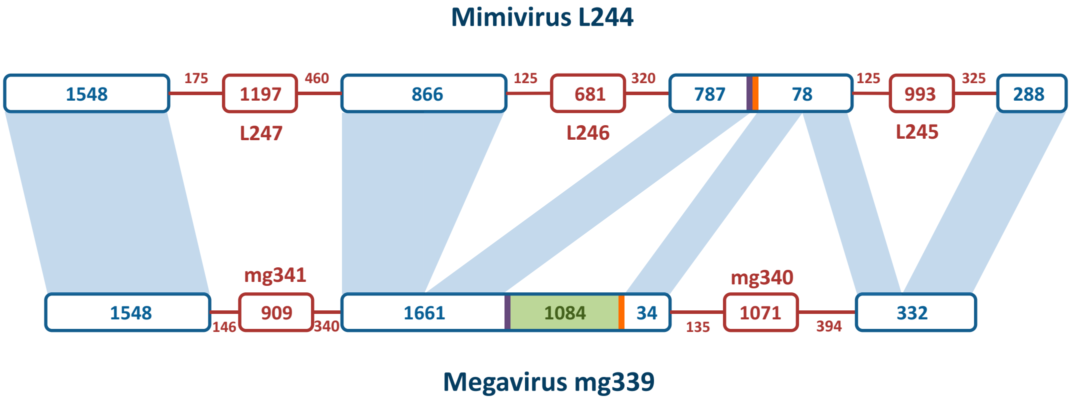


Fig. S8. Complex reorganization of the RPB2 gene in Megavirus. Exons are shown in blue, introns in brown, and the intein in green. Numbers correspond to DNA segment sizes in base pairs. The first exon is the only gene segment for which there is a one-to-one correspondence between Mimivirus and Megavirus. The second Megavirus exon incorporates most of the coding sequence of Mimivirus second and third exons, in addition to a 1,084-bp intein (1). The purple and orange boxes correspond to the last (H) and the first (S) amino acid of the N-terminal and C-terminal exteins, respectively. The third Megavirus exon corresponds to the end of the Mimivirus third exon and its entire fourth exon.

1. Perler FB (2002) InBase: The intein database. *Nucleic Acids Res* 30:383–384.

Other Supporting Information Files

Table S1 (DOC)