The 1.2-Mb Genome Sequence of Mimivirus

Didier Raoult, *et al.*

Science **0**, 1101485v1 (2004);
DOI: 10.1126/science.1101485

Sciencexpress Research Article

The 1.2-Megabase Genome Sequence of Mimivirus

Didier Raoult,[1]* Stéphane Audic,[2] Catherine Robert,[1] Chantal Abergel,[2] Patricia Renesto,[1] Hiroyuki Ogata,[2] Bernard La Scola,[1] Marie Susan,[1] Jean-Michel Claverie[2]*

[1]Unité des Rickettsies, Faculté de Médecine, CNRS UMR6020, Université de la Méditerranée, 13385 Marseille Cedex 05, France. [2]Information Génomique et Structurale, CNRS UPR2589, IBSM, 13402 Marseille Cedex 20, France.

*To whom correspondence should be addressed. E-mail: Jean-Michel.Claverie@igs.cnrs-mrs.fr (J.-M.C.); Didier.Raoult@medecine.univ-mrs.fr (D.R.)

We recently reported the discovery and preliminary characterization of Mimivirus, the largest known virus, with a 400-nm particle size comparable to mycoplasma. Mimivirus is a double-stranded DNA virus growing in amoebae. We now present its 1,181,404-bp genome sequence, consisting of 1262 putative ORFs, 10% of which exhibit a significant similarity with proteins of known functions. In addition to exceptional genome size, Mimivirus exhibits many features that distinguish it from other nucleocytoplasmic large DNA viruses. The most unexpected is the presence of numerous genes encoding central protein translation components, including 4 amino-acyl tRNA synthetases, peptide release factor 1, translation elongation factor EF-TU, and translation initiation factor 1. The genome also exhibit 6 tRNAs. Other remarkable features include the presence of both type I and type II topoisomerases, of components of all DNA repair pathways, of many polysaccharide synthesis enzymes, and of one intein-containing gene. The size and complexity of Mimivirus genome challenge the established frontier between viruses and parasitic cellular organisms. This new sequence data might help shed a new light on the origin of DNA viruses and their role in the early evolution of eukaryotes.

Mimivirus, the sole member of the newly proposed Mimiviridae family of nucleocytoplasmic large DNA viruses (NCLDV) was recently isolated from amoebae growing in the water of a cooling tower of an hospital in Bradford, England, in the context of pneumonia outbreak (*1*). The study of Mimivirus grown in *Acanthamoeba polyphaga* revealed a mature particle with the characteristic morphology of an icosahedral capsid of at least 400 nm in diameter. Such a virion size comparable to that of a mycoplasma cell makes Mimivirus the largest virus identified so far. A phylogenetic study using preliminary sequence data from a handful of conserved viral genes tentatively classified Mimivirus in a new independent branch of NCLDV (*1*). The sequencing of the genome of Mimivirus was undertaken to determine its complete gene content, to predict some of its physiology, to confirm its phylogenetic position among known viruses, and to gain some overall insight on the origin of NCLDVs.

Overall Genome Structure

Mimivirus genome (Fig. 1) was assembled [see methods in (*2*)] into a contiguous linear sequence of 1,181,404 bp, significantly greater in size than our initial conservative estimate of 800 kbp (*1*). The size and linear structure of the genome were confirmed by restriction digests and pulsed-field gel electrophoresis. Two inverted repeats of about 900 nt are found near both extremities of the assembled sequence, suggesting that Mimivirus genome might adopt a circular topology via their annealing, as in some other NCLDVs. From transmission electronic microscopy pictures, we estimated the volume of the dark central core of the virion (approximated as a sphere) at about 2.6 10^{-21} m^3, thus 3.7 times larger than the core volume of *Paramecium bursaria chlorella virus* (PBCV-1) (*3*). This is quite consistent with the respective genome sizes (1180kb/331kb=3.56) of the two viruses, indicating similar physical constraints for DNA packing (i.e. a core DNA concentration of about 450 mg/ml).

The nucleotide composition was 72.0% A+T. The genome exhibited a significant strand asymmetry. Both the cumulative A+C excess and the cumulative gene excess plots (*2*) exhibit a slope reversal (around position 400,000, Fig. 1 and fig. S1) as found in bacterial genomes and usually associated with the location of the origin of replication. Mimivirus genes are preferentially transcribed away from this putative origin of replication. Despite this local asymmetry, the total numbers of genes transcribed from either strand are similar (450 "R" *vs* 461 "L" ORFs). Repeated sequences represented less than 2.2% of Mimivirus genome (*2*).

We identified a total of 1,262 putative non-overlapping ORFs of length ≥100 amino acid residues, corresponding to a theoretical coding density of 90.5%. Of these ORFs, 911 were predicted to be protein coding genes, based on their statistical coding propensity and /or their similarity with database sequences. The remaining ORFs have been downgraded to the "URF" category (Unidentified reading frame). We were able to associate 298 ORFs with functional attributes (*2*).

The overall amino acid composition of the predicted Mimivirus proteome exhibits a strong positive bias for residues encoded by A+T rich codons. For instance, isoleucine (9.87%), asparagine (8.89%) and tyrosine (5.43%) are twice as frequent in Mimivirus as compared to amoeba or human proteins. Alanine (encoded by A+T poor codons GCN), is half as frequent (3.06%) as in the other two organisms. Similar variations have been observed in the amino acid compositions of other A+T rich DNA viruses (*4*). For any given amino acid, the relative usage of synonymous codons is also biased by the A+T rich genome composition. For instance ATT is largely dominant for Ile, as is AAT for Asn, and TAT for tyrosine. In the opposite, GCG is rarely used for Ala, as are CGG for Arg, and GGG and GGC for Gly. The codon usage in Mimivirus is almost the exact opposite of the one exhibited by *Acanthamoeba castellanii*: the least frequent codon in the amoeba is systematically the dominant one for Mimivirus. The codon usage in human genes also differs from the one in Mimivirus, but to a lesser extent due to the more even vertebrate codon distribution.

NCLDV Core Genes Identified in Mimivirus Genome

Iyers et al. (*5*) identified a set of genes present in all or most members of the 4 main NCLDV families: Pox-, Phycodna-, Asfar-, and Irido- viridae. These core genes are subdivided into four classes, from the most to least evolutionary conserved: class I includes those found in all known NCLDV genome sequences, class II are found in all NCLDV clades but are missing in some species, class III are identified in 3 out of the 4 NCLDV clades, and class IV are found in two clades only (*5*). The pattern of presence/absence of Class I, II and III core genes in Mimivirus is summarized in Table 1. We identified homologs for all (9/9) class I genes, 6/8 class II genes, 11/14 class III genes, and 16/30 class IV genes (*2*) (table S2). Both class II genes missing in Mimivirus are relevant to the biosynthesis of dTTP: thymidylate kinase, and dUTPase, a paradox given its A+T rich genome. Ectocarpus Silicosus virus (ESV) also lacks these enzymes. However, Mimivirus exhibits homologs for the class IV core genes thymidylate synthase and thymidine kinase. Additional nucleotide synthesis enzymes include deoxynucleoside kinase (DNK) and cytidine deaminase, as well as the first nucleoside diphosphate kinase (NDK) identified in a dsDNA virus. Mimivirus also lacks an ATP-dependent DNA ligase (a class III core gene) apparently replaced by a NAD-dependent ATP ligase (class IV), as found in Iridoviruses (*5*). With the exception of RNA polymerase subunit 10, Mimivirus genome exhibits the same transcription-related core genes as found in Pox- and Asfar-viridae. This suggests that the transcription of at least some Mimivirus genes occurs in the cytoplasm. Overall, the pattern of presence/absence of core genes (class II–IV) in Mimivirus is unlike any of the established ones. This confirms our initial suggestion (*1*) that Mimivirus constitutes the first representative of a new distinct NCLDV class (the "Mimiviridae").

Global Gene Content Statistics

All predicted Mimivirus ORFs were compared to the Clusters of Orthologous Groups (COG) database (*6*) using the Reverse PSI-BLAST program (*7*). We found that 194 Mimivirus ORFs exhibited significant matches with 108 distinct COG families (*2*). This is more than twice the number of COGs represented in PBCV-1 virus (46 ORFs matching with 41 COGs). Compared to other NCLDVs, Mimivirus COG profile exhibits a significant overrepresentation in the functional categories of translation (COG category J), post translation modifications (COG category O) and amino acid transport and metabolism (COG category E) (X^2 test: $P<0.001$, $P=0.006$ and $P=0.08$ respectively) (*2*) (table S3).

Novel Features in Mimivirus Genome Unique Among ds-DNA Viruses

The detailed analysis of Mimivirus genome (*2*) revealed a number of unique features, including many genes never yet identified in a viral genome. Until now, some of these genes were though to be the trade-mark of cellular organisms. These novel/unique genes are listed in Table 2. They can be classified in 4 generic functional categories: protein translation, DNA repair enzymes, chaperonins, and new enzymatic pathways. In addition Mimivirus is the sole virus and one of the rare micro-organisms to simultaneously possesses type IA, type IB and type II topoisomerases.

Protein translation-related genes. The inability to perform protein synthesis independently from their host is one of the main characteristic distinguishing viruses from cellular ("living") organisms. However, tRNA-like genes are found in isolated ds-DNA viruses species such as bacteriophage T4 (*8*) and BxZ1 (*9*), herpes virus 4 (*10*) and Chlorella viruses (*11*). The Chlorella viruses are also the first ones found to encode a translation elongation factor (EF-3) (*12*). The genome analysis of Mimivirus now greatly expand the known repertoire of viral genes related to protein translation. In addition to 6 tRNA-like genes (3 Leu [TTA, TTG], Trp[TGG], Cys [TGC], His [CAC]), Mimivirus genome exhibits homologs to 10 proteins with functions central to protein translation: 4 aminoacyl-tRNA synthetase (aaRS), translation initiation factor 4E (e.g. mRNA cap–binding), translation factor eF-TU (GTP-binding translocation factor), translation initiation factor SUI1, translation initiation factor IF-4A (a helicase) and peptide chain release factor eRF1. Finally, Mimivirus genome encodes the first identified viral homolog of a tRNA modifying enzyme (tRNA (Uracil-5-)-methyltransferase). All these ORFs have significant sequence similarity with their eukaryotic homologs, and exhibit all the domains and specific signatures expected from functional representatives of these various gene families. Preliminary functional characterizations have been obtained for several of

these genes. For instance, Mimivirus tyrosyl-tRNA synthetase was produced in *E. coli*, purified, and its enzymatic activity measured (*2*) (fig. S2). Crystals of the protein have been obtained and its 3-D structure is been refined (Abergel *et al*., in preparation). In addition, mRNAs encoding Mimivirus tyrosyl-, cysteinyl-, and arginyl- tRNA synthetases are found associated with purified virus particles (*2*) (table S4), suggesting that they are involved in infection.

New DNA repair enzymes. Genomes are subject to damage by chemical mutagens (e.g. free radicals alkylating agents), UV light or ionizing radiations. Different repair pathways have evolved to prevent the lethal accumulation of the various types of DNA errors. They usually correspond to well-conserved protein families found in the 3 domains of life (Archaea, Eubacteria, Eukaria) but to a much lesser extent in viruses. The analysis of Mimivirus genome revealed several types of DNA repair enzyme homologs, including four never yet reported in ds-DNA viruses. For instance, we identified two genes (L315, L720) encoding putative formamidopyrimidine-DNA glycosylases, the role of which is to locate and excise oxidized purines. Mimivirus genome also exhibits a UV-damage endonuclease (UvdE) homolog (L687). Although this is the first report of such an enzyme in a ds-DNA virus, we identified an isolated UvdE homolog among the "hypothetical" proteins of the recently sequenced *Aeromonas hydrophila* phage Aeh1 (ORF111c, Genbank Accession: AAQ17773). The major mutagenic effect of methylating agents in DNA is the formation of O6-alkylguanine. The corresponding repair is performed by a DNA-[protein]-cysteine S-methyltransferase Mimivirus genome encodes the first viral 6-O-methylguanine-DNA methyltransferase (R693). In addition, Mimivirus R406 ORF is strongly homologous to a number of bacterial genes annotated as belonging to the same alkylated DNA repair pathways. Finally, ORF L359 was found to clearly belong to the MutS protein family, involved in DNA mismatch repair and recombination. Again, this is the first DNA repair enzyme of this family described in a ds-DNA virus. Besides the above DNA repair system components never yet reported in ds-DNA virus, Mimivirus ORF L386 and R555 encode homologs to the rad2 and rad50 yeast genes, respectively, both central to the repair of UV-induced DNA damage. Homologs for these genes are also found in Iridoviruses. Overall, Mimivirus appears uniquely well equipped to repair DNA mismatch and damages caused by oxidation, alkylating agent or UV light.

Topoisomerases. DNA Topoisomerases are the enzymes in charge of solving the topological (entanglement) problems associated with DNA replication, transcription, recombination and chromatin remodelling (*13*). Type I topoisomerases (ATP-independent) work by passing one strand of the DNA through a break in the opposite strand. Type II topoisomerases are ATPases, and work by introducing a double-stranded gap. Topoisomerases of various types are involved in relaxing or introducing DNA supercoils. With the notable exception of Poxviridae, many ds-DNA virus (including NCLDV and phages) encode their own type II A topoisomerase. Accordingly, Mimivirus exhibits a large ORF (> 1263 aa, R480) 41% identical to PBCV-1 topoisomerase IIA amino-acid sequence. Its best database match overall is with an homologous protein in the small eukaryote *Encephalitozoon cuniculi* (42% identical). More surprisingly, Mimivirus is the first ds-DNA virus found to also encode a Poxviridae-like topoisomerase (Topoisomerase IB). Mimivirus ORF R194 is 27% identical to *Amsacta moorei* entomopoxvirus topoisomerase IB (AMV052) and 25% identical to the well-studied vaccinia virus topoisomerase (H6R). In addition to encode both type IIA and type IB topoisomerases, Mimivirus exhibits the first type IA topoisomerase reported in a virus (*13*). ORF L221 best overall database match (37%) is with its homolog in *Bacteroides thetaiotaomicron* (a Gram-negative anaerobe colonizing the human colon) within a well defined subgroup of well-conserved type IA eubacterial topoisomerases the prototype of which is *E. coli* Omega untwisting enzyme. Among all available genome sequences, only a small number of microorganisms simultaneously exhibit topoisomerases of type IA, IB and IIA. They include yeast, *Deinococcus radiodurans*, and various environmental bacteria such as *Pseudomonas sp*, *Agrobacterium tumefaciens*, and *Sinorhizobium meliloti*.

Protein folding. The folding of many proteins, in particular those involved in large molecular assemblies, is guided toward their native structures by different families of protein chaperones. Mimivirus genome uniquely exhibits two ORFs entirely and highly homologous to chaperones of the HSP70 (DnaK) family. ORF L254 is 42% identical to DnaK protein 2 of *Thermosynechococcus elongates*, and ORF L393 is 59% identical to Bovine Heat-shock 70-kilodalton protein 1A. In addition, Mimivirus genome exhibits 3 ORFs (R260, R266, R445) with clear DnaJ domain signatures. Proteins containing a DnaJ domain are known to associate with proteins of the HSP70 family. The above Mimivirus ORFs might thus encode a set of proteins interacting to form a specific viral chaperone system, possibly required for the productive assembly of its huge capsid.

Besides its gene equipment related to protein folding, Mimivirus is the first to encode a homolog to the lon *E. coli* heat shock protein, an ATP-dependent protease thought to dispose of unfolded polypeptides. Mimivirus does also exhibits components of the Ubiquitin-dependent protein degradation pathway, already described in other NCLDVs. Finally, Mimivirus genome also encodes a putative peptidyl-prolyl cis-trans isomerase of the Cyclophilin family (ORF

L605). This type of enzyme – seen here in a virus for the first time - accelerates protein folding by catalyzing the cis-trans isomerization of proline imidic peptide bonds. Again, this new virally encoded function might be required for Mimivirus capsid to be assembled within physiological time limits.

New Metabolic pathways. The genome analyses of large phycodnaviruses and other NCLDVs already contributed the notion that large viruses posses significant metabolic pathways in addition to the minimal infection, replication, transcription and virion packaging systems. PBCV-1, for instance, exhibits enzymes for the synthesis of homospermidine, hyaluronan, GDP-fucose, and many other sugar-, lipid-, and amino-acid- related manipulations (*14*). With its larger genome, Mimivirus builds on this established trend by exhibiting previously described as well as new virally encoded biosynthetic capabilities.

For instance, Mimivirus genome encodes homologs to many enzymes related to glutamine metabolism: asparagine synthase (glutamine hydrolysing) (ORF R475), glutamine synthase (ORF R565), and GMP synthase (glutamine hydrolysing) (ORF L716). All are identified in a ds-DNA virus for the first time. In addition, Mimivirus exhibits a glutamine: fructose-6-P aminotransferase (i.e. glucosamine synthase) as previously described in PBCV-1. Mimivirus can proceed further along this pathway using its own encoded N-acetylglucosamine-1-phosphate uridyltransferase (the well studied GlmU enzyme) (ORF R689) to synthetize UDP-N-acetyl-glucosamine (UDP-GlcNAc). This metabolite is central to the biosynthesis of all types of polysaccharides in both eukaryotic and prokaryotic systems. Mimivirus genome encodes 6 glycosyltranferases: 3 from family 2, and one from each families 8, 10 and 25.

Glycosyltransferases form a complex group of enzymes involved in the biosynthesis of disaccharides, oligosaccharides and polysaccharides that are involved in the post-translational modification of proteins (N- and O-glycosylation), the synthesis of lipopolysaccharides, as well as of high molecular weight cross-linked periplasmic or capsular material. Among other NCLDVs, PBCV-1 has been well studied in that respect, and shown to encode an atypical N-glycosylation pathways, as well as hyaluronan synthesis (*14*). Other Chlorovirus promote the synthesis of chitin (*15*). Preliminary proteomic studies of Mimivirus particles (see below) indicate that several proteins are glycosylated, including the predicted major capsid protein. In addition, Mimivirus particles are positive upon standard Gram staining [1], suggesting the presence of a reticulated polysaccharide at their surface. It is likely that some of Mimivirus glycosyltranferases are involved in its synthesis. For instance, Mimivirus encodes (L136) a homolog to perosamine synthetase. Such an enzyme catalyzes the conversion of GDP-4-keto-6-deoxymannose to 4-NH2-4,6-dideoxymannose (perosamine), that is found in the O-antigen moity of the LPS of various bacteria. Another Mimivirus ORF (L230) is homologous to Procollagen-lysine, 2-oxoglutarate 5-dioxygenase. This enzyme catalyzes the formation of hydroxylysine in collagens and other proteins with collagen-like amino acid sequences, by the hydroxylation of lysine residues in X-lys-gly sequences. These hydroxyl groups then serve as sites of attachment for carbohydrate units and are also essential for the stability of the intermolecular collagen crosslinks. Given that Mimivirus also contains a large number of ORFs exhibiting the characteristic collagen triple helix repeat, it is tempting to speculate that the hairy-like appearance of the virion [1] might be due to a layer of cross-linked glycosylated collagen-like fibrils.

Among other enzymes never yet reported in a virus, Mimivirus includes a nucleoside diphosphate kinase (NDK) (EC:2.7.4.6) (ORF R418). NDK catalyzes the synthesis of nucleoside triphosphates (NTP) other than ATP. This enzyme may help circumvent a limited supply of NTPs for nucleic acid synthesis, UTP for polysaccharide synthesis and GTP for protein elongation.

Finally, Mimivirus is also encoding homologs to 3 lipid-manipulating enzymes: Acetylcholinesterase (L906), Lanosterol 14-alpha-demethylase (L808) and 7-dehydrocholesterol reductase (R807) the physiological roles of which - possibly the disruption of the host membrane- remain to be determined.

Intein and Introns. Inteins are protein-splicing domains encoded by mobile intervening sequences (IVS)(*16*). They self catalyze their excision from the host protein, ligating their former flanks by a peptide bond. They have been found in all domains of life (Eukaria, Archaea and Eubacteria) but their distribution is highly sporadic. Only few instances of viral inteins have been described, in *B. subtilis* bacteriophages (*17*), and in the ribonucleotide reductase alpha subunit of Chilo iridescent virus (CIV)(*18*). Mimivirus is then the second eukaryotic ds-DNA virus exhibiting an intein (*2*). In contrast with the one described for CIV (lacking a C-terminal Asn), Mimivirus intein is canonical and exhibits valid amino acids at all essential positions, as well as the DOD homing endonuclease motif (Ogata et al., in preparation). For reasons not yet understood, inteins are most often found associated with essential enzymes of the DNA metabolism. Inserted within DNA polB, Mimivirus intein is no exception to this rule.

Self-splicing type I introns are a different type of mobile IVS, self excising at the mRNA level. They are rare in viruses, and mostly found in phages. One type IB intron has been identified in several Chlorella virus species (*14*). Mimivirus exhibits 4 instances of self excising intron (*2*) all

in RNA polymerase genes: one in the largest and 3 in the second largest subunit.

Gene families or protein domains expanded in Mimivirus. The ankyrin-repeat signature is the most frequent motif, found in more than 30 distinct ORFs. This motif, about 33 amino acids long, is one of the most common protein-protein interaction motifs. It has been found in proteins with a wide diversity of functions. Another protein interaction domain, defined by the BTB signature, is found in 20 ORFs. This domain mostly mediates homomeric dimerisation. It is found in proteins that contain the KELCH motif such as Kelch and a family of pox virus proteins. We identified 14 different ORFs exhibiting the PFAM signature (*19*) of the catalytic domain of eukaryotic protein kinases (PFAM: pkinase motif, $p<0.05$). Four of them resemble known cell-division related kinases.

The collagen triple helix motif is another frequently represented motif, found in 8 ORFs. This motif is characteristic of extracellular structural proteins involved in matrix formation and/or adhesion processes. Like other collagens, the product of these collagen-like ORFs might be post translationally modified by the Procollagen-lysine, 2-oxoglutarate 5-dioxygenase homolog uniquely found in Mimivirus genome (see above). Mimivirus also contains 8 ORFs with significant similarity to helicases. Finally, Mimivirus exhibits 8 ORFs containing a specific glucose-methanol-choline (GMC) oxidoreductase motif. The role of these FAD flavoproteins is unknown.

Phylogeny

Relationship to other NCLDVs. Our preliminary study based on the protein sequences of ribonucleotide reductase small/large subunits and topoisomerase II (*1*) suggested an independent branching of Mimivirus in the phylogenetic tree of NCLDVs (*1*). This analysis was refined by using the concatenated sequences of the eight "class I" genes conserved in Mimivirus and all other NCLDVs. The resulting phylogenetic tree again suggested that Mimivirus defines an independent lineage of NCLDVs (Fig. 2) roughly equidistant from known Phycodnaviruses and Iridoviruses.

Relationship to the 3 domains of life. 63 COGs are common to all known unicellular genomes from the three domains of life: Eukarya, Eubacteria and Archaea. Seven of them are now identified in the genome of Mimivirus: three aminoacyl-tRNA synthetases [ArgRS (COG0018), MetRS (COG0143), TyrRS (COG0162)], the beta (COG0085) and beta' (COG0086) subunits of RNA polymerase, the sliding clamp subunit of DNA polymerase (3 PCNA paralogs; COG0592), and a 5'-3' exonuclease (COG0258). The unrooted phylogenetic tree build from the concatenated sequences of those proteins (*2*) is shown in Fig. 3. Mimivirus branches out near the origin of the Eukaryota domains. This is supported with a high bootstrap value and the Shimodaira-Hasegawa statistical test (*2*). The tree topology is also invariant to a variety of methodological changes (*2*) (figs. S3-S6). Consistently, scatter plots for the best BLAST scores against the three domains of life indicate that most Mimivirus ORFs exhibit higher sequence similarities to eukaryotic sequences than to prokaryotic sequences, and are equidistant from the 4 main eukaryotic Kingdoms: Protista, Animalia, Plantae and Fungi (*2*) (fig. S7). However, strictly speaking, the tree shown in Fig.3 can be rooted on any of the deepest branches, including the branch separating Mimivirus from eukaryotes, making its specific affinity with Eukaryota still uncertain.

Genome Complexity: Mimivirus *Versus* Parasitic Cellular Organisms

The number of Mimivirus COGs was compared to the ones found for representative of the 3 domains of life using the smallest known genomes: *Nanoarchaeum equitans* (490 kb), *Mycoplasma genitalium* (580 kb), and *Encephalitozoon cuniculi* (2.498 kb) (Fig. 4). Despite its comparable genome size, Mimivirus exhibits less identified COGs. However, there was no specific category in which it was significantly underrepresented, except for the translation category ($p<0.01$). In contrast, it possesses relatively more COGs in the replication, recombination, and repair categories than the others ($p<0.08$). By this standard, the absence of a functional protein translation apparatus is what distinguishes most Mimivirus from its parasitic cellular counterparts.

Preliminary Analysis of Mimivirus Particles

Detection of viral RNAs. Large viruses such as those of the *Herpesviridae* family, incorporate viral transcripts during the particle assembly process (*20*). We thus investigated whether viral RNAs could be found associated with Rnase-treated Mimivirus particles using RT-PCR and virus specific primers targeting several genes (*2*). Positive results were obtained for 3 aminoacyl tRNA synthetases (TyrRS, CysRS, ArgRS), DNA polymerase, transcription factor TFIIB, and the predicted major capsid protein gene (L425) (*2*) (table S4).

Virion proteomics. Constituent proteins of Mimivirus particles were extracted and analysed. In a preliminary set of experiments, 2D gel electrophoresis resolved 438 spots, many of them visibly corresponding to multiple isoforms of the same gene product (i.e. glycosylation, phosphorylation, etc.) (*2*) (fig. S8). The most abundant of the best resolved spots were eluted and characterized by mass spectrometry (Maldi-ToF and ion trap). Six predicted ORF products corresponding to proteins with homologs of known functions were unambiguously identified. As expected, they include the major capsid (L425) and core (L410) proteins, but also an mRNA capping enzyme (R382), thioredoxin (R548) and glutaredoxin (R195), and a GMC type oxidoreductase (R135).

Virion resistance to adverse conditions. Mimivirus particles remained infectious during 1 year when kept at 4°C,

25°C and 32°C in PAS buffer. Incubation of a suspension of 10^9 particles in PAS buffer at 55°C from 15 to 90 minutes reduced its titer by 100. By comparison, no viable *E. coli* are retrieved when submitted to the same treatment. No diminution in Mimivirus titre was observed after 48H desiccation. Mimivirus particles are thus quite resistant to adverse conditions. However, despite its many predicted DNA repair genes, Mimivirus is quickly killed by 35 kGy irradiation with gamma rays or exposure for 15 min (30 watts, 20 cm) to UV light (*2*).

Discussion

A common feature to all known viruses is their total dependency of the host translation machinery for protein synthesis. Surprisingly, Mimivirus genome sequence now reveal genes relevant to all key steps of mRNA translation: tRNA and tRNA charging, initiation, elongation and termination, with the exception of ribosome components themselves. Two main evolutionary scenarios may account for the presence of this partial complement of translation-related genes in Mimivirus. On one hand, they could be the relics of a more complete ancestral protein translation apparatus, gradually lost through a genome reduction process similar to the one governing the evolution of intracellular bacteria (*21*). On the other hand, these genes could have been individually acquired from cellular organisms and used to control the host translation apparatus in favour of Mimivirus mRNAs. The fact that our phylogenetic analysis did not support a recent acquisition of these genes, together with the low probability for these genes to have been acquired independently, is in favour of the loss rather than the gain scenario. By extrapolating this model, we could speculate that the Mimivirus lineage originated from a more complex ancestor possibly exhibiting an even more complete protein translation machinery.

By its particle size, and now by its genome complexity, Mimivirus significantly challenges our vision of viruses. Lwoff (*22*) proposed that viruses should have at least one dimension lower than 200 nm and speculated that viruses may possess only one type of nucleic acid. Both criteria are invalidated by Mimivirus. Lwoff also pointed out the lack of enzymes generating energy from substrates. This criteria is still valid, as very few genes of this category were detected in Mimivirus. Other criteria such as the strictly intracellular character and the inability to grow or undergo binary fission have not yet been challenged. By these three last criteria, Mimivirus remains a regular virus. However, by the unprecedented number of enzymes and putative metabolic pathways encoded by its 1.2 Mb genome, Mimivirus blurs the established frontier between viruses and the parasitic cellular organisms with small defective genomes such as *R. prowazekii* (*23*), Buchnera (*24*), *Nanoarchaeum* (*25*), Mycoplasma (*26*), and *Tropheryma whipplei* (*27*). As of today, the genome of Mimivirus is larger than the published genomes of 20 cellular organisms from two domains of life (e.g. Archaea and Eubacteria) and 5 main bacterial divisions: Proteobacteria, Firmicutes, Actinobacteria, Chlamydiae and Spirochaetes. The presence *vs.* absence of ribosomes remains, at the moment, a key property distinguishing these minimal cellular organisms from large DNA viruses.

Several independent studies have led to the hypothesis that DNA viruses may have a common origin, and a common ancestor, originating before the emergence of the three domains of life (*28*). Given the inherent uncertainty of phylogenetic reconstruction dating back 3 billion years ago, our results (Fig. 3) are consistent with the hypotheses that a lineage of large DNA viruses could have emerged before the individualization of cellular organisms from the 3 domains of life (*29*) or from an ancestor distinct of these 3 domains (*30*). The topology of this new "tree of life" is also consistent with the hypothesis that ancestral DNA viruses were involved in the emergence of Eukaryotes (*31–35*).

In conclusion, the serendipitous discovery of Mimivirus from samples initially thought to contains a new type of intracellular Gram positive bacterium – our main area of expertise -, allowed the characterization of the largest virus so far. The sequencing of its 1.2 Mb genome revealed a wealth of genes encoding functions never yet encountered in viruses, most probably due to its unprecedented size. The numerous new genes related to the protein translation apparatus challenge the established vision of viruses. With the addition of these new viral representative of universally conserved gene families, we could now build a tentative tree of life, within which Mimivirus appears to define a new branch distinct of the 3 other domains. We believe our work should prompt the search for more giant viruses the genome analysis of which could shed additional light on the origin of DNA viruses and their role in the evolution of cellular organisms.

References and Notes

1. B. La Scola *et al.*, *Science* **299**, 2033 (2003).
2. See supporting material on *Science* Online
3. N. Nandhagopal *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 14758 (2002).
4. C.L. Afonso *et al.*, *J. Virol.* **73**: 533 (1999).
5. L.M. Iyer, L. Aravind, E.V. Koonin, *J. Virol.* **75**, 11720 (2001).
6. R.L. Tatusov *et al.*, *BMC Bioinformatics* **4**, 41 (2003).
7. S.F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
8. E.S. Miller *et al.*, *Microbiol. Mol. Biol. Rev.* **67**, 86 (2003).
9. M.L. Pedulla *et al.*, *Cell* **113**, 171 (2003).
10. H.W. Virgin *et al.*, *J. Virol.* **71**, 5894 (1997).
11. J.L. Van Etten, R.H. Meints, *Annu. Rev. Microbiol.* **53**, 447 (1999).
12. T. Yamada, T. Fukuda, K. Tamura, S. Furukawa, P. Songsri,*Virology* **197**, 742 (1993).

13. J.J. Champoux, *Annu. Rev. Biochem.* **70**, 369 (2001).
14. J.L. Van Etten, *Annu. Rev. Genet.* **37**, 153 (2003).
15. T. Kawasaki *et al.*, *Virology* **302**, 123 (2002).
16. S. Pietrokovski, *Trends Genet.* **17**, 465 (2001).
17. V. Lazarevic *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1692 (1998).
18. S. Pietrokovski, *Curr Biol.* **8**, R634 (1998).
19. A. Bateman *et al.*, *Nucleic Acids Res.* **30**, 276 (2002).
20. W.A. Bresnahan, T. Shenk , *Science* **288**, 2373 (2000).
21. N.A. Moran, *Cell* **108**, 583 (2002).
22. A. Lwoff, *J. Gen. Microbiol.* **17**, 239 (1957).
23. S.G.E. Andersson *et al.*, Nature **396**, 133(1998).
24. I. Tamas I *et al.*, *Science* **296**, 2376 (2002).
25. E. Waters *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 100, 12984 (2003).
26. C.M. Fraser *et al.*, *Science* **270**, 397 (1995).
27. D. Raoult *et al.*, *Genome Res.* **13**, 1800 (2003).
28. W. Zillig *et al.*, *Extremophiles* **2**, 131 (1998).
29. C. Woese, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 6854 (1998).
30. D.P. Mindell, L.P. Villarreal, *Science* **302**, 1677 (2003).
31. L.P. Villarreal, V.R. DeFilippis, J. Virol. **74**, 7079 (2000).
32. P. Forterre, *C. R. Acad. Sci. III.* **324**, 1067 (2001).
33. M. Takemura, *J. Mol. Evol.* **52**, 419 (2001).
34. P.J.L. Bell, *J. Mol. Evol.* **53**, 251 (2001).
35. E. Pennisi, *Science* **305**, 766 (2004).
36. We thank our colleagues C. Fraisier, A. Honstettre, V. Arondel as well as N. Androvandi, S. Chenivesse and D. Moinier for technical help. Special thanks also to M. Drancourt and K. Suhre for helpful discussions. The IGS laboratory is partially supported by Aventis Pharma. Mimivirus genome sequence has been deposited to Genbank under accession number AY653733.

Supporting Online Material
www.sciencemag.org/cgi/content/full/1101485/DC1
Materials and Methods
SOM Text
Figs. S1 to S8
Tables S1 to S4

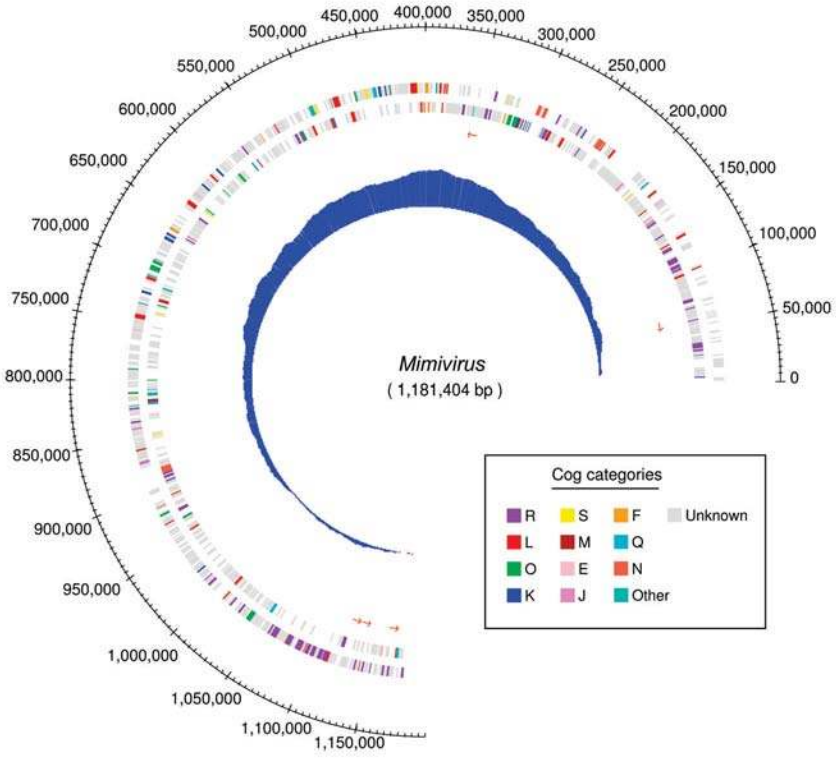
15 June 2004; accepted 22 September 2004
Published online 14 October 2004; 10.1126/science.1101485
Include this information when citing this paper.

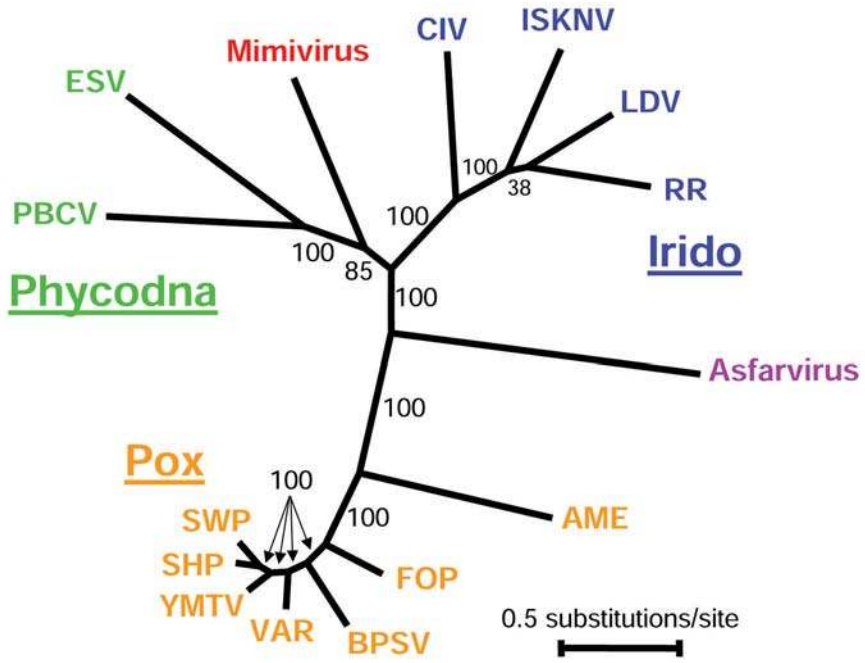
Fig. 1. Map of Mimivirus chromosome. The predicted protein coding sequences are shown on both strands, and colored according to the function category of their matching COG. Genes with no COG match are shown in grey. Abbreviations for the COG functional categories are as follows: E, Amino acid transport and metabolism; F, Nucleotide transport and metabolism; J, Translation; K, Transcription; L, Replication, recombination and repair; M, Cell wall/membrane biogenesis; N, Cell motility; O, Posttranslational modification, protein turnover, chaperones; Q, Secondary metabolites biosynthesis, transport and catabolism; R, General function prediction only; S, Function unknown. Small red arrow indicates the location and orientation of tRNAs. The A+C excess profile is shown on the innermost circle, exhibiting a peak around position 380,000 (*2*) (fig. S1).

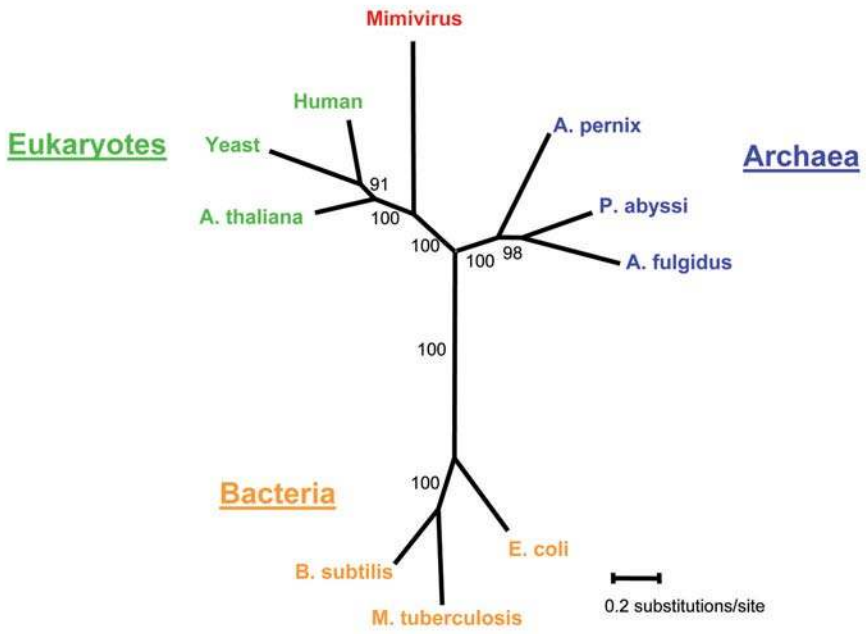
Fig. 2. Phylogenetic position of Mimivirus among established NCLDV families. Viral species representing the diverse families of NCLDV are included as follows: Mimivirus, *Phycodnaviridae* (*Paramecium bursaria* Chlorella virus 1 [PBCV] and *Ectocarpus siliculosus* virus [ESV]), *Iridoviridae* (*Chilo* iridescent virus [CIV], *Regina* ranavirus [RR], Lymphocystis disease virus type 1 [LDV], and Infectious Spleen and Kidney Necrosis virus [ISKNV]), *Asfarviridae* (African swine fever virus [ASFV]), and *Poxviridae* (*Amsacta moorei* entomopoxvirus [AME], Variola virus [VAR], Fowlpox virus [FOP], Bovine papular stomatitis virus [BPSV], Yaba monkey tumor virus [YMTV], Sheeppox virus [SHP], and Swinepox virus [SWP]). Fully sequenced viral genomes were analyzed to ensure the proper assessment of orthologous genes. This tree was build using maximum likelihood and based on the concatenated sequences of eight conserved proteins (NCLDV class I genes): VV D5-type ATPase, DNA polymerase family B, VV A32 virion packaging ATPase, capsid protein, thiol oxidoreductase, VV D6R helicase, serine/threonine protein kinase, A1L transcription factor. One of the class I genes (VV A18 helicase) being absent in LDV was not included. The alignment contains 1660 sites without insertions and deletions. A neighbor joining tree and a maximum parsimony tree exhibited similar topologies (*2*). Bootstrap percentages are shown along the branches.

Fig. 3. Phylogeny: overall position of Mimivirus among all kingdoms. A phylogenetic tree of species from the three domains of life (Eukaryota, Eubacteria, Archaea) and Mimivirus. The tree was inferred with the use of a maximum likelihood method based on the concatenated sequences of seven universally conserved protein sequences: arginyl-tRNA synthetase (COG0018), methionyl-tRNA synthetase (COG0143), tyrosyl-tRNA synthetase (COG0162), RNA polymerase II largest subunit (COG0086), RNA polymerase II second largest subunit (COG0085), PCNA (COG0592) and 5'-3' exonuclease (COG0258). The alignment contains 3164 sites without insertions and deletions. Bootstrap percentages are shown along the branches. Similar trees were obtained using a variety of other approaches (Supporting online text).

Fig. 4. Distribution of COG homologues in Mimivirus compared to the cellular organisms of the 3 domains of life with the smallest known genomes.







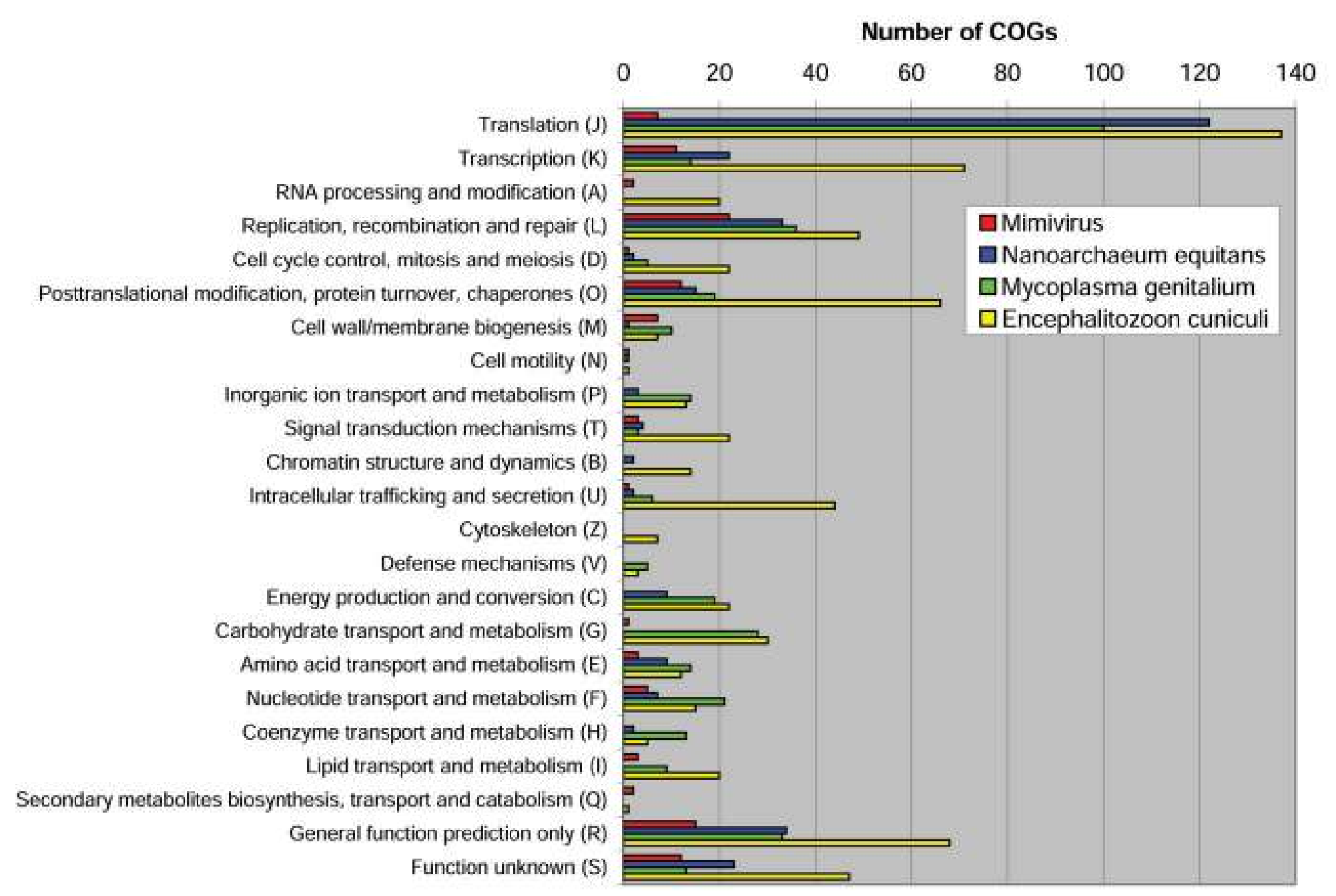

Number of COGs
0
20
40
60
80
100
120
140
Translation (J)
Transcription (K)
RNA processing and modification (A)
Replication, recombination and repair (L)
Cell cycle control, mitosis and meiosis (D)
Posttranslational modification, protein turnover, chaperones (O)
Cell wall/membrane biogenesis (M)
Cell motility (N)
Inorganic ion transport and metabolism (P)
Signal transduction mechanisms (T)
Chromatin structure and dynamics (B)
Intracellular trafficking and secretion (U)
Cytoskeleton (Z)
Defense mechanisms (V)
Energy production and conversion (C)
Carbohydrate transport and metabolism (G)
Amino acid transport and metabolism (E)
Nucleotide transport and metabolism (F)
Coenzyme transport and metabolism (H)
Lipid transport and metabolism (I)
Secondary metabolites biosynthesis, transport and catabolism (Q)
General function prediction only (R)
Function unknown (S)
Mimivirus
Nanoarchaeum equitans
Mycoplasma genitalium
Encephalitozoon cuniculi

Table 1. NCLDV "core" genes (group I, II, III) identified in Mimivirus. ■, best matching homologs; X, significant homolog detected in all available genomes; x, not in all in available genomes.

ORF #	Phycodna viridae	Pox viridae	Irido viridae	Asfar viridae	Gene group	Definition/putativefunction (Ref. 5)
L206	X	X	■	X	I	Helicase III / VV D5-type ATPase
R322	■	X	X	X	I	DNA polymerase (B family)
L437	X	X	■	X	I	VV A32 virion packaging ATPase
L396	■	X	x	X	I	VV A18 helicase
L425	■	X	X	X	I	Capsid protein D13L, (4 paralogs)
R596	■	X	X	X	I	Thiol oxidoreductase (e.g. E10R)
R350	X	■	X	X	I	VV D6R helicase, +1paralog
R400	■	X	X	X	I	S/T protein kinase (e.g. F10L)
R450	■	X	X	X	I	Transcription factor (e.g. A1L)
R339	■x	X	X	X	II	TFII-like transcription factor
L524	x	X	■X	X	II	MuT-like NTP pyrophophohydrolase
L323	x	X	■X	X	II	Myristoylated virion protein A
R493	■X	x	X	X	II	PCNA + 1paralog
R313	X	■x	X	X	II	Ribonucleotide reductase, large sub.
L312	X	■x	X	X	II	Ribonucleotide reductase, small sub.
Not found	x	x	X	X	II	Thymidylate kinase
Not found	x	X	X	X	II	dUTPase
R429	■	-	X	X	III	PBCV1-A494R-like (9 paralogs)
L37	X	■	X	X	III	BroA, KilA-N term
R382	X	X	-	■	III	mRNA Capping Enzyme
L244	-	X	■	X	III	RNA Pol subunit 2 (Rbp2)
R501	-	X	■	X	III	RNA Pol largest sub. (Rpb1)
R195	■	X	X	-	III	Glutaredoxin (e.g. ESV128)
R622	X	■	X	-	III	Dual spec. S/Y phosphatase
R311	-	x	X	X	III	BIR domain (e.g. CIV193R)
L65	-	■X	X	X	III	Virion-associated membrane protein
R480	■	-	X	X	III	Topoisomerase II
L364	X	■	X	-	III	SW1/SNF2 helicase (e.g. MSV224)
Not found	x	X	X	-	III	RuvC-like HJR (e.g. A22R)
Not found	x	x	-	X	III	ATP-dependent DNA ligase (e.g. A50R)
Not found	-	x	X	X	III	RNA polymerase subunit 10

Table 2. Major new features identified in Mimivirus genome.

ORF #	Definition/putativefunction	Comment
R663	Arginyl-tRNA synthetase	Translation
L124	Tyrosyl-tRNA synthetase	Translation
L164	Cysteinyl-tRNA synthetase	Translation
R639	Methyonyl tRNA synthetase	Translation
R726	Peptide chain release factor eRF1	Translation
R624	GTP binding elongation factor eF-Tu	Translation
R464	Translation initiation factor SUI1	Translation
L496	Translation initiation factor 4E (mRNA cap binding)	Translation
R405	tRNA (Uracil-5-)-methyltransferase	tRNA modification
L359	DNA mismatch repair ATPase MutS	DNA repair
R693	Methylated-DNA-protein-cysteine methyltransferase	DNA repair
R406	Alkylated DNA repair	DNA repair
L687	Endonuclease for the repair of UV-irradiated DNA	DNA repair
L315 L720	Hydrolysis of DNA containing ring-opened N7-methylguanine	DNA repair
R194 R480 L221	Topoisomerase I pox-like Topoisomerase II Topoisomerase I bacterial type	DNA accessibility
L254 L393	Heat shock 70kD HSP	Chaperonin
L605	Peptidylprolyl isomerase	Chaperonin
L251	Lon domain protease	Chaperonin
R418	NDK synthesis of nucleoside triphosphates	Metabolism
R475	Asparagine synthase (glutamine hydrolysing)	Metabolism
R565	Glutamine synthetase (Glutamate-amonia ligase)	Metabolism
L716	Glutamine amidotransferase domain	Metabolism
R689	N-acetylglucosamine-1-phosphate uridyltransferase	Polysaccharide synthesis
L136	Sugar transaminase dTDP-4-amino-4,6-dideoxyglucose biosynthesis	ExoPolysaccharide synthesis
L780	dTDP-4-dehydrorhamnose reductase	ExoPolysaccharide synthesis
L612	Mannose-6P isomerase	Glycosylation
L230	Procollagen-lysine,2-oxoglutarate 5-dioxygenase	Glycosylation, Capsid structure
L543	ADP-ribosyltransferase (DraT)	?
L906	cholinesterase	Host infection?
L808	Lanosterol 14-alpha-demethylase	Host infection?
R807	7-dehydrocholesterol reductase	Host infection?
R322	Intein insertion	in DNA PolB