Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses

Mickaël Boyer[9], Mohammed-Amine Madoui[9], Gregory Gimenez, Bernard La Scola, Didier Raoult*

URMITE, Centre National de la Recherche Scientifique UMR IRD 6236, Faculté de Médecine, Université de la Méditerranée, Marseille, France

Abstract

The discovery of Mimivirus, with its very large genome content, made it possible to identify genes common to the three domains of life (*Eukarya*, *Bacteria* and *Archaea*) and to generate controversial phylogenomic trees congruent with that of ribosomal genes, branching Mimivirus at its root. Here we used sequences from metagenomic databases, Marseillevirus and three new viruses extending the *Mimiviridae* family to generate the phylogenetic trees of eight proteins involved in different steps of DNA processing. Compared to the three ribosomal defined domains, we report a single common origin for Nucleocytoplasmic Large DNA Viruses (NCLDV), DNA processing genes rooted between *Archaea* and *Eukarya*, with a topology congruent with that of the ribosomal tree. As for translation, we found in our new viruses, together with Mimivirus, five proteins rooted deeply in the eukaryotic clade. In addition, comparison of informational genes repertoire based on phyletic pattern analysis supports existence of a clade containing NCLDVs clearly distinct from that of *Eukarya*, *Bacteria* and *Archaea*. We hypothesize that the core genome of NCLDV is as ancient as the three currently accepted domains of life.

Citation: Boyer M, Madoui M-A, Gimenez G, La Scola B, Raoult D (2010) Phylogenetic and Phyletic Studies of Informational Genes in Genomes Highlight Existence of a 4th Domain of Life Including Giant Viruses. PLoS ONE 5(12): e15530. doi:10.1371/journal.pone.0015530

Editor: Darren P. Martin, Institute of Infectious Disease and Molecular Medicine, South Africa

Received August 4, 2010; **Accepted** October 11, 2010; **Published** December 2, 2010

Funding: This work was funded by the Centre National de la Recherche Scientifique (CNRS, crédits récurrents). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: didier.raoult@gmail.com

[9] These authors contributed equally to this work.

Introduction

Molecular sequence analyses have allowed to partially identify the origins of genes. A putative tree of life based on ribosomal analysis was postulated that includes three domains: *Eukarya*, *Bacteria* and *Archaea*. Viruses are excluded from this classification system due to a lack of evidence that they possess a core of genes and because they lack ribosomes [1]. The double jelly-roll motif that can be found in the capsids of viruses infecting *Archaea*, *Eukarya*, or *Bacteria*, however, is proposed as evidence for an ancestry of some viruses [2,3]. The description of the giant Mimivirus [4] that share similar features with cells has recently generated much debate about the nature of viruses in the living world [5–7]. In our study, we identify a common set of proteins present in all forms of life including cellular organisms and viruses, particularly giant viruses, and we used this set of viral and cellular proteins to perform phylogenetic reconstructions.

Results

In modern cells, biosynthesis of DNA precursors (dNTPs) is achieved in two steps by two essential enzymes (Figure 1): ribonucleotide reductase (RNR) and thymidylate synthase (TdS) [8]. First, reduction of RNA precursors (rNTP) into dNTP is catalyzed by RNR, then deoxythymidine 5′-monophosphate (dTMP) is produced from deoxyuridine 5′-monophosphate (dUMP) by thymidylate synthase. These two key enzymes are thought to have been involved in the transition from ancient RNA to the DNA world [9]. They have been identified in cells and in nucleocytoplasmic large DNA viruses (NCLDVs) and used for phylogenetic reconstructions. RNR proteins have been classified into three classes [10]; we focus on classes I and II as class III proteins share very little similarity with the other two classes [11]. The RNR phylogeny (Figure S1) shows prokaryotic and eukaryotic clades and supports the emergence of RNR of three NCLDV families separately from the eukaryotes. Two families of non-homologous TdS, ThyA and ThyX, have been identified in living organisms, but the thymidylate synthases reported in NCLDVs are mostly of the ThyA type [12]. The phylogenetic reconstruction of the ThyA type (Figure S2) showed two clusters, one that includes bacteria, bacteriophages and *Archaea*, and the other that includes eukaryotes and their viruses. In the second group, ThyA in all NCLDV lineages was scattered within the eukaryotic clade, except for ThyA from Mimivirus and Marseillevirus, which emerge on a separate clade. Phylogenetic trees inferred for these two proteins also show NCLDV branches within eukaryotic clades. We assumed, as others have [13], that multiple gene exchanges or non-orthologous gene displacements have occurred during the extensive evolution between eukaryotes and viruses, thus blurring the topology.

Four key proteins are involved in DNA replication and repair. Three enzymes, including the DNA polymerase family B (DNAP B), the topoisomerase II A (TopoIIA), the Flap endonuclease

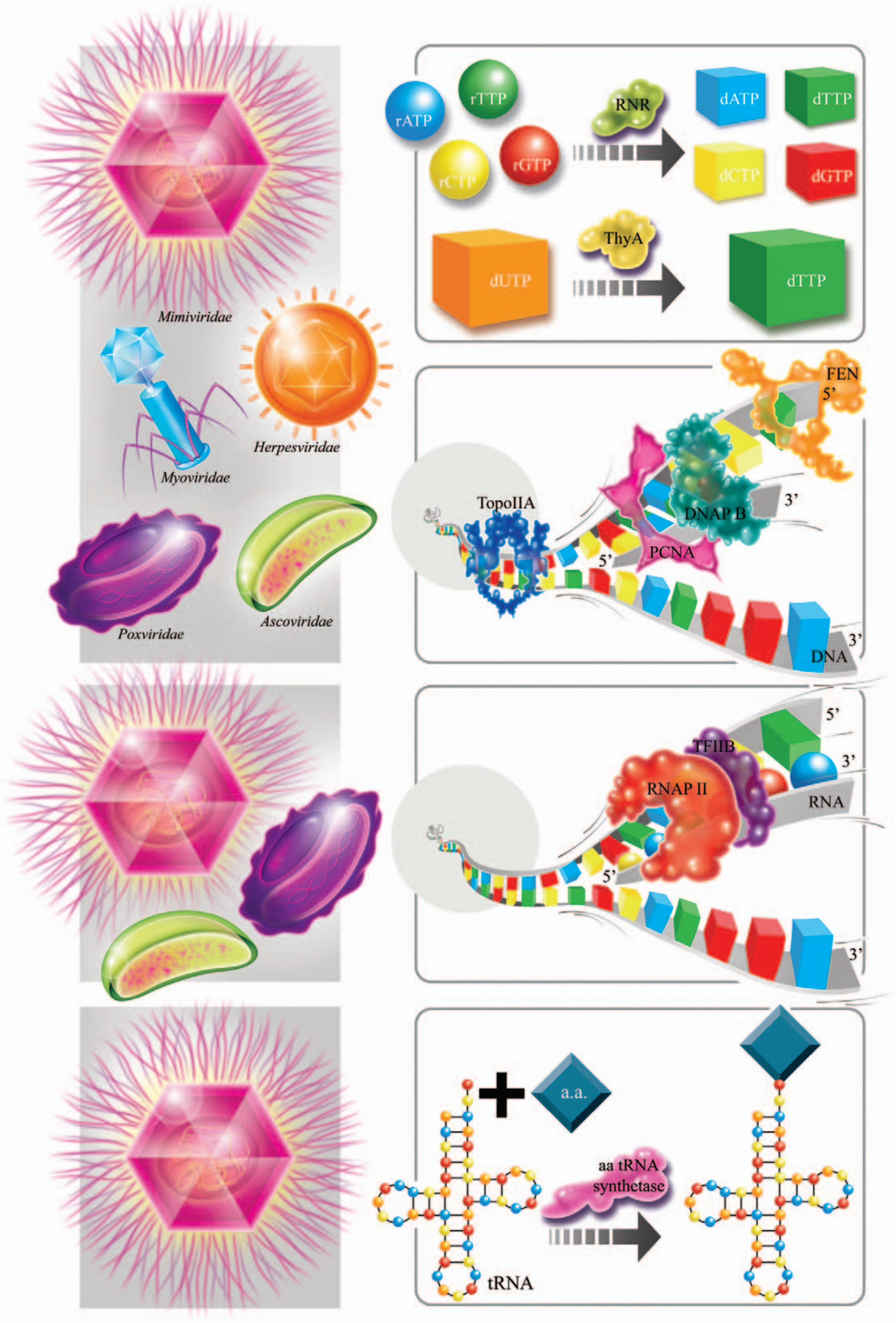

Figure 1. NCLDV proteins involved in DNA, RNA and protein biosynthesis. The frames on the right represent DNA processing steps with related enzymes found in NCLDVs, and the frames on the left with grey backgrounds show viruses that have the corresponding enzymes. The first frame from the top right corresponds to the dNTP biosynthesis catalyzed by RNR and ThyA. The second frame shows four enzymes involved in DNA replication: DNAP B, TopoIIA, PCNA and FEN. Proteins involved in these two DNA biosynthesis steps were found in diverse large DNA viruses, such as NCLDVs, *Herpesviridae* and *Myoviridae*. The third frame represents the DNA transcription step, in which the RNAP II and the transcription factor TFIIB are involved; viral homologs of these proteins are only found in NCLDVs. The last frame shows *Mimiviridae* amino-acyl tRNA synthetases (aa tRNA synthetase) involved in mRNA translation. Interestingly, viral lineages were consistently less well represented from nucleotide to protein biosynthesis, with the exception of NCLDVs, which were represented in each step. a.a., amino acid.
doi:10.1371/journal.pone.0015530.g001

(FEN), and the processing factor Proliferating Cell Nuclear Antigen (PCNA) (Figure 1) are largely distributed in NCLDVs [13] and are therefore used for phylogenetic reconstruction. The DNAP B tree (Figure S3) was built from archaeal, bacterial, viral and eukaryotic α, δ and ζ protein sequences' alignment. Sequences from NCLDVs and *Herpesviridae* were sister clades rooted between the δ and ζ eukaryotic subfamilies. The TopoIIA tree (Figure S4) was obtained from archaeal, bacterial, eukaryotic and NCLDVs' sequences' alignment. The TopoIIA phylogenetic reconstruction showed eukaryotic monophyly, whereas NCLDVs are not monophyletic but emerged as a bush between the eukaryotic and prokaryotic clades. PCNA and FEN were not found in *Bacteria* and their respective phylogenetic trees (Figures S5 and S6, respectively), supporting distinct *Archaea*, eukaryote and NCLDV monophylies. Therefore, these four protein phylogenies support, with a high degree of confidence, the existence of a viral clade with ancestral DNA replication machinery branching separately from *Eukarya* and *Archaea*.

In addition, we investigated the phylogenetic reconstruction of two proteins involved in transcription processing: the DNA-dependant RNA polymerase II (RNAP II) is a ubiquitous enzyme, and its phylogenetic tree reveals a clade containing NCLDVs clearly distinct from that of *Eukarya*, *Bacteria* and *Archaea* (Figure 2 and Figure S7). Transcription factor II B (TFIIB), absent in bacteria, is a general transcription factor that makes up the RNA polymerase II pre-initiation complex. Its phylogenetic analysis displayed similar results (Figure 3), with the exception that bacteria are absent. These results lead to the conclusion that highly conserved proteins involved in the RNA biosynthesis in NCLDVs have emerged in a clade as ancient as those of eukaryotes and *Archaea*.

The Mimivirus genome contains several genes encoding proteins involved in DNA translation [4]. We investigated a phylogenetic reconstruction for five proteins involved in translation, including four amino-acyl tRNA synthetases (Figures S8, S9, S10, S11) and the putative elongation factor EF-1 (Figure S12), which are also found in the draft genomes of three newly described members of the *Mimiviridae* family (Terravirus, Courdovirus and Moumouvirus) [14]. Arginyl-tRNA synthetase and cysteine-tRNA synthetase phylogenetic trees (Figures S8 and S10) show blurred topologies of *Mimiviridae* branching inside the eukaryotes clade, but the topology of the methyonyl-tRNA synthetase tree (Figure S11), which contains three distinct eukaryotic clades, supports the existence of a *Mimiviridae* clade branching before the eukaryotic clade containing *Amoebozoa* and opisthokonts. The topology of the tyrosyl-tRNA synthetase tree supports the hypothesis of a transfer from an ancient viral clade to an amoeba [15,16]. It was previously suggested that the amino-acyl tRNA synthetases are more commonly exchanged genes [17], and the amino-acyl tRNA synthetase trees, including that of *Mimiviridae*, support this hypothesis and show that viruses are also included in the confused evolutionary scenarios of these proteins. Proteins related to EF-1

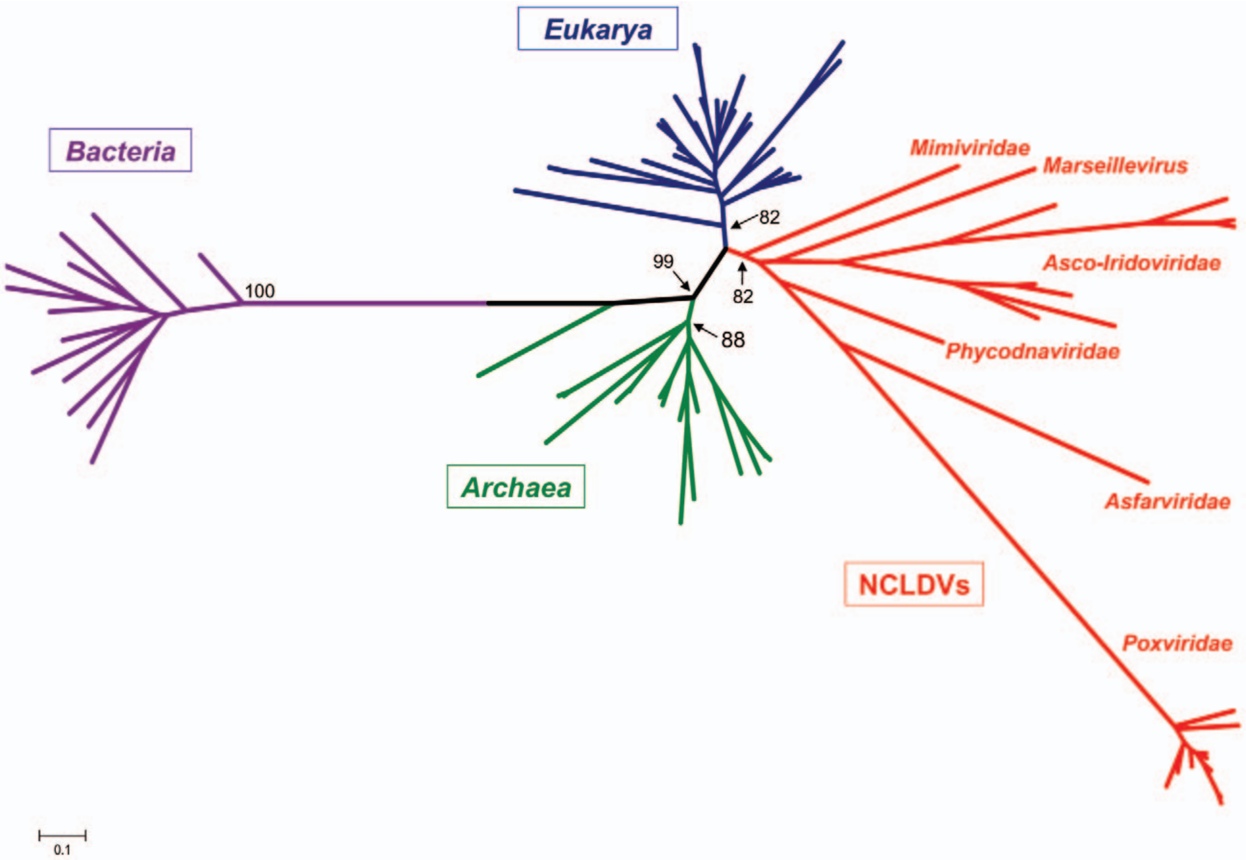

Figure 2. Phylogenetic tree of the RNA polymerase II beta subunit. The ML tree of RNAP II was inferred from a cured alignment of 80 sequences from the six supergroups of *Eukarya* (blue), *Bacteria* (purple), *Archaea* (green) and NCLDVs (red). The tree is unrooted, and values near branches are SH-like local supports computed by the FastTree program and are used as confidence values of tree branches. Scale bar represents the number of estimated changes per position for a unit of branch length.
doi:10.1371/journal.pone.0015530.g002

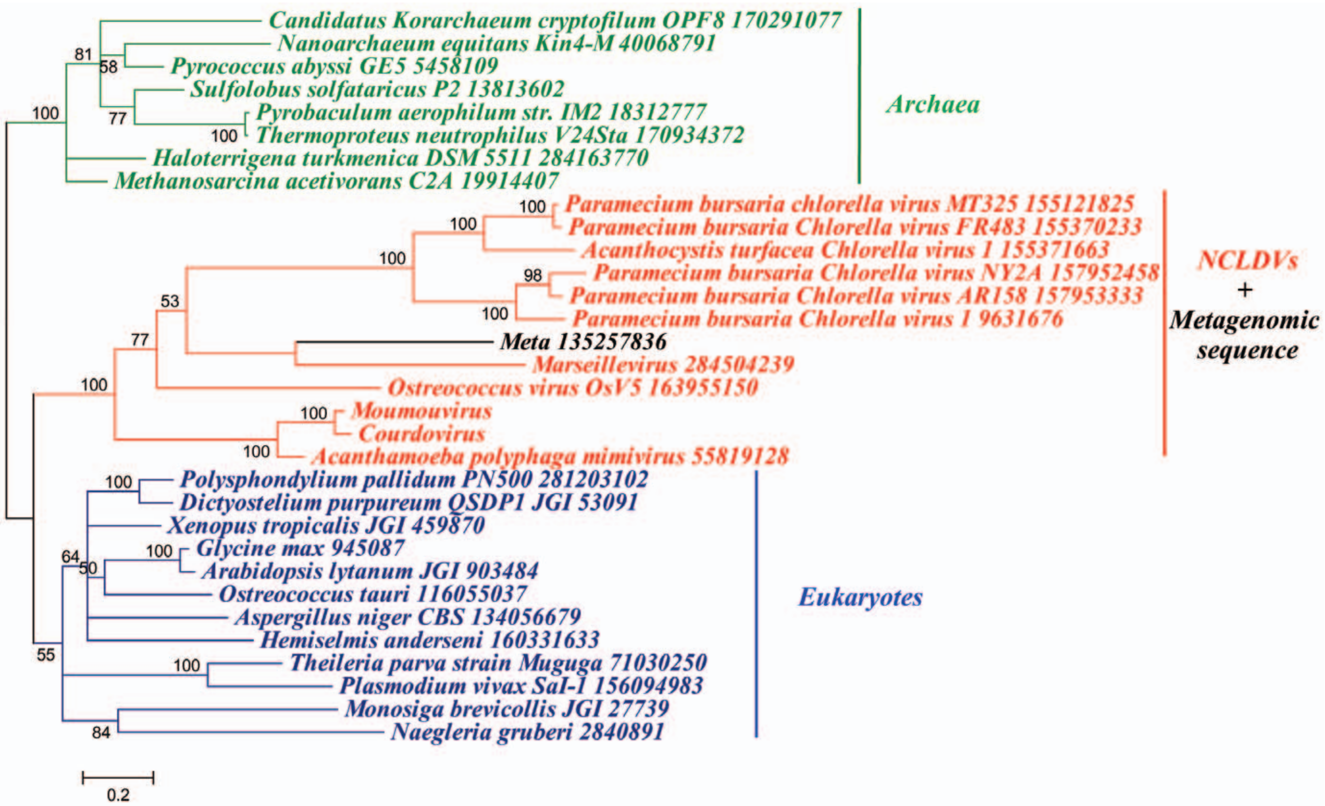
Candidatus Korarchaeum cryptofilum OPF8 170291077
Nanoarchaeum equitans Kin4-M 40068791
Pyrococcus abyssi GE5 5458109
Sulfolobus solfataricus P2 13813602
Pyrobaculum aerophilum str. IM2 18312777
Thermoproteus neutrophilus V24Sta 170934372
Haloterrigena turkmenica DSM 5511 284163770
Methanosarcina acetivorans C2A 19914407
Archaea
81
58
77
100
100
Paramecium bursaria chlorella virus MT325 155121825
Paramecium bursaria Chlorella virus FR483 155370233
Acanthocystis turfacea Chlorella virus 1 155371663
Paramecium bursaria Chlorella virus NY2A 157952458
Paramecium bursaria Chlorella virus AR158 157953333
Paramecium bursaria Chlorella virus 1 9631676
Meta 135257836
Marseillevirus 284504239
Ostreococcus virus OsV5 163955150
Moumouvirus
Courdovirus
Acanthamoeba polyphaga mimivirus 55819128
NCLDVs
+
Metagenomic
sequence
100
100
100
98
100
53
77
100
100
100
Polysphondylium pallidum PN500 281203102
Dictyostelium purpureum QSDP1 JGI 53091
Xenopus tropicalis JGI 459870
Glycine max 945087
Arabidopsis lytanum JGI 903484
Ostreococcus tauri 116055037
Aspergillus niger CBS 134056679
Hemiselmis anderseni 160331633
Theileria parva strain Muguga 71030250
Plasmodium vivax SaI-1 156094983
Monosiga brevicollis JGI 27739
Naegleria gruberi 2840891
Eukaryotes
100
100
64
50
55
100
84
0.2

Figure 3. Phylogenetic tree of the Transcription factor II B (TFIIB). The TFIIB phylogenetic tree is inferred with Bayesian approach from a cured alignment of 32 sequences (155 conserved positions) from the *Eukarya* (blue), *Archaea* (green), NCLDVs (red), and metagenomic databases (black). Bayesian posterior probabilities are mentioned near branches as a percentage and are used as confidence values of tree branches. Scale bar represents the number of estimated changes per position for a unit of branch length.
doi:10.1371/journal.pone.0015530.g003

are divided into different subfamilies according to their functional domains [18]. That of *Mimiviridae* displays the GTP-binding protein domain (GTPBP1) found in a protein family divergent from the bona fide EF-1 family [19]. The phylogenetic tree (Figure S12) exhibited three large clades corresponding to the archaeal and eukaryotic translation protein families related to EF-1, one clade corresponding to the divergent bacterial elongation factor EF-TU, and another corresponding to GTPBP1 proteins. A phylogenetic tree supported the emergence of the *Mimiviridae* elongation factor inside the GTPBP1 eukaryote clade. In conclusion, the phylogenetic trees based on translational proteins suggest that *Mimiviridae* acquired some of the genes at the root of the eukaryote clade.

As DNA processing genes appear to be the most conserved genes between cellular organisms and viruses, we further constructed a different type of tree, known as phyletic pattern [20], that is based on the comparison of the gene repertoire involved in the information storage and processing. A dendogram tree (Figure 4) was constructed from the presence/absence of the respective genes of each virus and cellular organisms in database of Clusters of Orthologous Groups (COGs) related to the DNA processing functions. The topology of this gene content tree supports the existence of a clade containing NCLDVs clearly distinct from that of *Eukarya*, *Bacteria* and *Archaea*. In accordance with the phylogenetic tree based on NCLDV core genes [13], viruses of the same family roughly grouped together inside the NCLDV clade. This tree based on informational gene repertoire is compatible with the above phylogenetic trees showing clearly four distinct clades and with that of the ribosomal phylogenetic trees (excluding viruses).

Discussion

Our results attest to the existence of a single and ancestral source of genes allowing for the biosynthesis of DNA in large DNA viruses. Based on the study of these genes, we speculate that they appeared simultaneously with or just after the emergence of the modern eukaryote lineages. This topology may be biased by the fact that the current eukaryotes, arising from the fusion of a proto-eukaryote with an alpha-proteobacteria, emerged a billion years ago. This was followed by a bottleneck and the disappearance of all proto-eukaryotes that did not successfully form a symbiotic fusion with bacteria [21]. Therefore, there are no traces of the viruses of these proto-eukaryotes. This might explain the branching of some NCLDVs inside the eukaryote clade. Finally, an analysis of the current genealogy of organisms showed that the group of genes associated with DNA synthesis has a common source that is conserved in three or four current domains of organisms (including viruses) and had a topology comparable to that of ribosomal phylogenetic trees (excluding viruses). Finally, if we consider these nine proteins, five were present in the three canonical domains and NCLDVs; the other four trees included only three clades: the ribosomal tree and three containing NCLDVs and two domains (PCNA, FEN, and TFIIB). We conclude that trees based on ribosomal proteins are not sufficient to represent all forms of life as they do not include viruses.

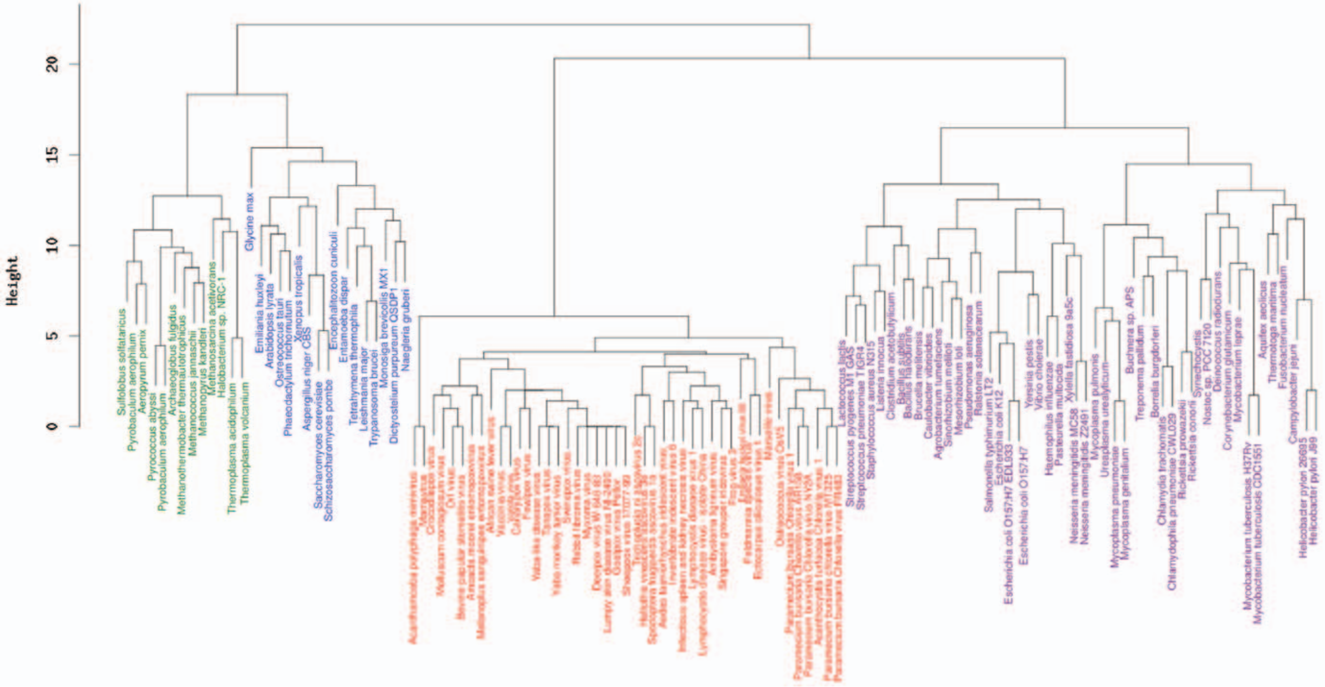

Figure 4. Hierarchical clustering of *Eukarya* (blue), *Bacteria* (purple), *Archaea* (green) and NCLDVs (red) by phyletic pattern. The phyletic patterns of the putative orthologous sets of informational genes indicating the presence/absence of the respective gene in each cellular organisms and virus were used for the construction of the dendogram tree.
doi:10.1371/journal.pone.0015530.g004

In addition, as previously suggested [22], it is possible that the ancestral cellular DNA machinery came from DNA viruses, which would have thus in turn provided genes allowing DNA replication. This work also confirms and extends the seminal work of Iyers *et al.*, who identified for the first time a core genome of the NCLDV with nine genes common to all species of this group of viruses [23]. Interestingly, the name NCLDV does not reflect the striking complexity of this group of viruses that may contain both DNA and RNA [4]. Finally in this study, we found that the informational gene repertoires, irrespective of their individual source, are also evidencing four major clades.

Because the set of genes associated with the DNA machinery is found in both viruses and cells, the origin of this gene set may be questionable, and some authors have suggested that large DNA viruses might have contributed to the formation of the nucleus of eukaryotes [24,25]. In this scenario, current forms of life would result from the fusion of the world of RNA and proteins (including ribosomes) with that of DNA; RNA and protein molecules have been preserved in modern cells (*Archaea*, *Bacteria* and *Eukarya*), whereas some viruses (i.e., *Mimiviridae*) containing both RNA and DNA have conserved translation-associated proteins such as tRNA synthetases, which are probably remnants of an ancestral translation apparatus. In any case, we speculate that DNA machinery is a clear remnant of a biological "Big Bang" generating at least two distinct forms of life: DNA-based and DNA-RNA-ribosome-based. Since this ancestral time, two forms of life have co-developed [6]; cells that are ribosome-encoding organisms (REOs), diversified in three domains (*Archaea*, *Bacteria*, and *Eukarya*), and capsid-encoding organisms (CEOs) (bona fide viruses, parasites of the REOs). Accepting this ancestry allows us to reinterpret the origin of two-tailed virus families (i.e., *Myoviridae* and *Siphoviridae*) infecting both bacteria and *Archaea*. One can suggest that they derive from ancestral viruses infecting the REO ancestor rather than being transferred from one domain to another [26]. This makes sense if one considers that *Siphoviridae* appear to be the most common viral family on Earth [27]. Finally, several previous studies focusing on the common genetic features in viruses from *Archaea* and from eukaryotes [28,29] suggest that there is a conserved endonuclease structure in influenza virus and in the endonucleases of *Archaea* and bacteria [30]. Evidence that the picorna-like superfamily of RNA viruses originated before the radiation of eukaryotes [31] and a phylogenetic tree of the guanylyl transferase domain branching poxviruses at the origin of eukaryotes [25] might also be reinterpreted in this light and are congruent with the hypothesis of CEO ancestry. Based on these data, we propose a scenario in which NCLDV emerged from the rhizome of life [32] with roots arising at the very beginning of life (Figure 5).

Materials and Methods

Groups of homologs common to the three domains of life and NCLDVs

Selected proteomes of organisms representing major phyla of the three domains of life (*Eukarya*, *Archaea* and *Bacteria*) were downloaded from NCBI and JGI. BLASTP [33] searches were performed with this set of proteins to identify conserved proteins in the complete proteomes that are available for the six NCLDV families (*Asfarviridae*, *Asco-Iridoviridae*, *Phycodnaviridae*, *Poxviridae*, Marseillevirus and *Mimiviridae*) and the draft proteomes of three new *Mimiviridae* [14]. Groups of homologs were built using the criteria that sequences were homologous if the Best Blast Hit (BBH) showed an alignment length of over 70 amino acids and with a percent identity over 20%. Four proteins matching these criteria that were also present in at least the three domains of life were identified: thymidylate synthase, RNR, DNAP B, TopoIIA (also called gyrase in bacteria), and RNAP II. The FEN, the PCNA and the TFIIB were found only in eukaryotes and *Archaea*.

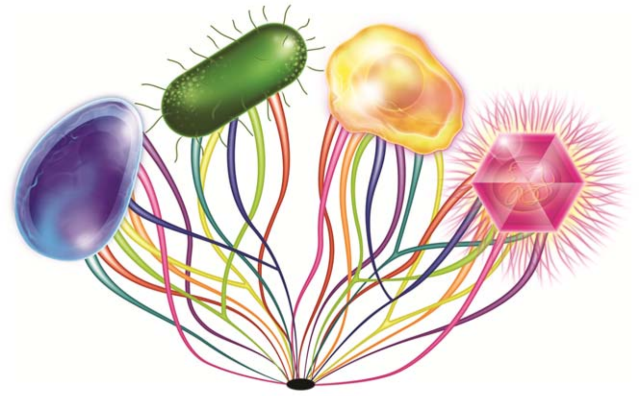

Figure 5. Scenario of NCLDVs emerging from the rhizome of life with roots appearing at the very beginning of life. This figure represents the living species in the four small pictures according to the current classification of organisms: eukaryotes (represented by yellow cell), bacteria (represented by green cell), *Archaea* (represented by blue cell) and viruses (represented by magenta colored Mimivirus). An organism classification based on ribosomal proteins allows the discrimination of cellular organisms, but viruses are excluded de facto from historical scenarios for life evolution. Our study shows that eight proteins involved in DNA processing and present both in viruses, particularly in NCLDVs, and in two or three other domains of life could be used for phylogenetic tree reconstruction, displaying schematic topologies for each protein represented by lines colored in light green for RNR, in yellow for ThyA, in blue-green for DNAP B, in blue for TopoIIA, in magenta for PCNA, in orange for FEN, in red for RNAP II and in purple for TFIIB. Topology representations were similar to that based on ribosomes (for RNAP II, FEN, PCNA and TFIIB proteins) and/or with bifurcations (for ThyA, RNR, DNAP B and Topo IIA proteins) representing probable lateral gene transfer. Thus, this figure illustrates that REOs and CEOs (bona fide viruses) share a common set of genes involved in DNA processing that evolved from a common ancestral source of genes. The phenomenon of long-branch attraction makes the identification of the exact deep rooting of each analyzed gene difficult and is therefore hidden by a black mark.
doi:10.1371/journal.pone.0015530.g005

Aminoacyl-ARNt synthetases and EF-1-like proteins were only identified in *Mimiviridae* and the three domains of life. For this set of proteins, homologs among all the viral proteomes together with homologs with the largest coverage relating to our inquiry in environmental sequence data (env_nr) were detected using BLASTP. The Courdovirus protein sequences have been deposited in GenBank under the following accession numbers: HQ223093 for cysteinyl-tRNA_synthetase, HQ223094 for tyrosyl-tRNA_synthetase, HQ223095 for methionyl-tRNA_synthetase, HQ223096 for arginyl-tRNA_synthetase, HQ223097 for GTPB1_like, HQ404368 for TFIIB. The Terravirus protein sequences have been deposited in GenBank under the following accession numbers: HQ223098 for cysteinyl-tRNA_synthetase, HQ223099 for tyrosyl-tRNA_synthetase, HQ223100 for methionyl-tRNA_synthetase, HQ223101 for arginyl-tRNA_synthetase, HQ223102 for GTPB1_like. The Moumouvirus protein sequences have been deposited in GenBank under the following accession numbers: HQ223103 for cysteinyl-tRNA_synthetase, HQ223104 for tyrosyl-tRNA_synthetase, HQ223105 for GTPB1_like, HQ404369 for TFIIB.

Multiple sequence alignments and phylogenetic reconstruction

T-Coffee [34] or Muscle [35] was used to construct multiple alignments, and conserved blocks were selected using Gblocks [36]. The corrected alignments were then used for maximum likelihood (ML) and Bayesian inference (BI). ML phylogeny inference was constructed using FastTree [37], with the JTT+CAT substitution model. Values near branches are SH-like local supports computed by the FastTree program and are used as confidence values of tree branches. For the Bayesian approach, the phylogeny was performed using MrBayes[38]; the WAG matrix was used, and model parameters (gamma shape and proportion invariant) were allowed to vary through the Markov Chain Monte Carlo Chain (MCMC). Four MCMC chains were run for 1 million generations and sampled every 100th generation. The first 100,000 trees were discarded, and the sumt command of MrBayes was used to compute clade posterior probabilities. Trees were displayed using MEGA 4 [39].

Construction of dendrogram gene content tree from phyletic patterns

Comparison of informational genes repertoire was performed using a phyletic pattern indicating the presence/absence of the respective informational genes of all species present in the corresponding functional categories of Clusters of Orthologous Group (COGs) based on 66 genomes [40,41]. Only COGs functional categories related to Information storage and processing ([J, A, K, L and B] COG categories) and Nucleotide transport and metabolism ([F] COG category) were used to build a matrix by assigning as a "1" if there is at least one ortholog contained in a genome and a "0" if not. As eukaryotes are weakly represented in initial COG database, we extended the initial matrix by assigning a COG functional category to proteins of 14 added eukaryotic proteomes using BLASTP against COG database (e-value <10-3). We also incremented the matrix of phyletic pattern with NCLDVs species by assigning NCLDVs proteins in corresponding COGs, which is performed by linking functional categories of NCVOGs database [13] with those of COGs using the same parameters. The euclidian distance matrix was computed from the 0/1 matrix and then a dendogram tree was built from hierarchical clustering using in house R language.

Supporting Information

Figure S1 Bayesian phylogenetic tree of RNR (31 sequences, 166 positions). GI or JGI numbers are listed next to the corresponding taxonomic name of each cellular organism and virus. A color code was used to represent taxonomic groups, *Bacteria* in purple, *Archaea* in green, *Eukarya* in blue, NCLDVs in red, other viruses and phages in pink and environmental sequences in black. Numbers at nodes are Bayesian posterior probabilities. Scale bar represents the number of estimated changes per position for a unit of branch length.
(PDF)

Figure S2 Bayesian phylogenetic tree of ThyA (44 sequences, 159 positions). Detailed legend is the same as in Figure S1.
(PDF)

Figure S3 Bayesian phylogenetic tree of DNAP B (62 sequences, 139 positions). Detailed legend is the same as in Figure S1.
(PPT)

Figure S4 Bayesian phylogenetic tree of TopoIIA (48 sequences, 140 positions). Detailed legend is the same as in Figure S1.
(PPT)

Figure S5 Bayesian phylogenetic tree of PCNA (40 sequences, 174 positions). Detailed legend is the same as in Figure S1.
(PPT)

Figure S6 Bayesian phylogenetic tree of FEN (37 sequences, 304 positions). Detailed legend is the same as in Figure S1.
(PPT)

Figure S7 Bayesian phylogenetic tree of RNAP II (80 sequences, 272 positions). Detailed legend is the same as in Figure S1.
(PPT)

Figure S8 Maximum-likelihood phylogenetic tree of arginyl-tRNA synthetase (32 sequences, 123 positions). Detailed legend is the same as in Figure S1 except that numbers at nodes are SH-like local supports.
(PPT)

Figure S9 Maximum-likelihood phylogenetic tree of tyrosyl-tRNA synthetase (33 sequences, 201 positions). Detailed legend is the same as in Figure S1 except that numbers at nodes are SH-like local supports.
(PPT)

Figure S10 Maximum-likelihood phylogenetic tree of cysteine-tRNA synthetase (38 sequences, 156 positions). Detailed legend is the same as in Figure S1 except that numbers at nodes are SH-like local supports.
(PPT)

Figure S11 Maximum-likelihood phylogenetic tree of methyonyl-tRNA synthetase (51 sequences, 204 positions). Detailed legend is the same as in Figure S1 except that numbers at nodes are SH-like local supports.
(PDF)

Figure S12 Bayesian phylogenetic tree of EF-1 like phylogenetic tree (65 sequences, 157 positions). Detailed legend is the same as in Figure S1.
(PPT)

Acknowledgments

We are grateful to Ghislain Fournous for help with bioinformatic analyses; Lina Barrassi, Angélique Campocasso and Isabelle Pagnier for their technical assistance in isolation of new viruses and Catherine Robert for technical assistance in genome sequencing. We thank Christelle Forzale for help with the figures 1 and 5.

Author Contributions

Conceived and designed the experiments: DR MB MAM. Performed the experiments: MB MAM GG. Analyzed the data: DR MB MAM GG BLS. Contributed reagents/materials/analysis tools: MB MAM GG. Wrote the paper: DR MB MAM.

References

1. Moreira D, Lopez-Garcia P (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7: 306–311.
2. Baker ML, Jiang W, Rixon FJ, Chiu W (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. J Virol 79: 14967–14970.
3. Krupovic M, Bamford DH (2008) Virus evolution: how far does the double beta-barrel viral lineage extend? Nat Rev Microbiol 6: 941–948.
4. Raoult D, Audic S, Robert C, Abergel C, Renesto P, et al. (2004) The 1.2-megabase genome sequence of Mimivirus. Science 306: 1344–1350.
5. Moreira D, Brochier-Armanet C (2008) Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. BMC Evol Biol 8: 12.
6. Raoult D, Forterre P (2008) Redefining viruses: lessons from Mimivirus. Nature Rev Microbiol 6: 315–319.
7. Raoult D (2009) There is no such thing as a tree of life (and of course viruses are out!). Nat Rev Microbiol 7: 615; author reply 615.
8. Forterre P (2002) The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5: 525–532.
9. Forterre P (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117: 5–16.
10. Torrents E, Aloy P, Gibert I, Rodriguez-Trelles F (2002) Ribonucleotide reductases: divergent evolution of an ancient enzyme. J Mol Evol 55: 138–152.
11. Jordan A, Reichard P (1998) Ribonucleotide reductases. Annu Rev Biochem 67: 71–98.
12. Filee J, Pouget N, Chandler M (2008) Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. BMC Evol Biol 8: 320.
13. Yutin N, Wolf YI, Raoult D, Koonin EV (2009) Eukaryotic large nucleo-cytoplasmic DNA viruses: Clusters of orthologous genes and reconstruction of viral genome evolution. Virol J 6: 223.
14. La Scola B, Campocasso A, N'Dong R, Fournous G, Barrassi L, et al. (2009) A burden of new giant viruses from environment and their tentative characterization by MALDI-TOF mass-spectrometry. Intervirology 53: 344–353.
15. Moliner C, Raoult D, Fournier PE (2009) Evidence of horizontal gene transfer between amoeba and bacteria. Clin Microbiol Infect 15: 178–180.
16. Forterre P (2010) Giant viruses: conflicts in revisiting the virus concept. Intervirology 53: 362–378.
17. Woese CR, Olsen GJ, Ibba M, Soll D (2000) Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. Microbiol Mol Biol Rev 64: 202–236.
18. Keeling PJ, Inagaki Y (2004) A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. Proc Natl Acad Sci U S A 101: 15380–15385.
19. Senju S, Iyama K, Kudo H, Aizawa S, Nishimura Y (2000) Immunocytochemical analyses and targeted gene disruption of GTPBP1. Mol Cell Biol 20: 6195–6200.
20. Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees and the tree of life. Trends Genet 18: 472–479.
21. Cavalier-Smith T (1987) The origin of eukaryotic and archaebacterial cells. Ann N Y Acad Sci 503: 17–54.
22. Villarreal LP, Witzany G (2010) Viruses are essential agents within the roots and stem of the tree of life. J Theor Biol 262: 698–710.
23. Iyer LM, Balaji S, Koonin EV, Aravind L (2006) Evolutionary genomics of nucleo-cytoplasmic large DNA viruses. Virus Res 117: 156–184.
24. McGeoch DJ, Davison AJ (1995) Origins of DNA viruses. In Molecular Basis of Virus Evolution, pp. 67–75. Edited by Adrian J. Gibbs, Charles H. Calisher & Fernando García-Arenal. Cambridge University Press.
25. Bell PJ (2001) Viral eukaryogenesis: was the ancestor of the nucleus a complex DNA virus? J Mol Evol 53: 251–256.
26. Prangishvili D, Forterre P, Garrett RA (2006) Viruses of the Archaea: a unifying view. Nat Rev Microbiol 4: 837–848.
27. Ackermann HW (2007) 5500 Phages examined in the electron microscope. Arch Virol 152: 227–243.
28. Blum H, Zillig W, Mallok S, Domdey H, Prangishvili D (2001) The genome of the archaeal virus SIRV1 has features in common with genomes of eukaryal viruses. Virology 281: 6–9.
29. Peng X, Blum H, She Q, Mallok S, Brugger K, et al. (2001) Sequences and replication of genomes of the archaeal rudiviruses SIRV1 and SIRV2: relationships to the archaeal lipothrixvirus SIFV and some eukaryal viruses. Virology 291: 226–234.
30. Dias A, Bouvier D, Crepin T, McCarthy AA, Hart DJ, et al. (2009) The cap-snatching endonuclease of influenza virus polymerase resides in the PA subunit. Nature 458: 914–918.
31. Koonin EV, Wolf YI, Nagasaki K, Dolja VV (2008) The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. Nat Rev Microbiol 6: 925–939.
32. Raoult D (2010) The post-Darwinist rhizome of life. Lancet 375: 104–105.
33. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
34. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol 302: 205–217.
35. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.
36. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56: 564–577.
37. Price MN, Dehal PS, Arkin AP (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.

38. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

39. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.

40. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.

41. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, et al. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29: 22–28.