Coevolutionary and Phylogenetic Analysis of Mimiviral Replication Machinery Suggest the Cellular Origin of Mimiviruses

Supriya Patil and Kiran Kondabagil *

Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra, India

*Corresponding author: E-mail: kirankondabagil@iitb.ac.in, kirankondabagil@gmail.com.

Associate editor: Crystal Hepp

Abstract

Mimivirus is one of the most complex and largest viruses known. The origin and evolution of Mimivirus and other giant viruses have been a subject of intense study in the last two decades. The two prevailing hypotheses on the origin of Mimivirus and other viruses are the reduction hypothesis, which posits that viruses emerged from modern unicellular organisms; whereas the virus-first hypothesis proposes viruses as relics of precellular forms of life. In this study, to gain insights into the origin of Mimivirus, we have carried out extensive phylogenetic, correlation, and multidimensional scaling analyses of the putative proteins involved in the replication of its 1.2-Mb large genome. Correlation analysis and multidimensional scaling methods were validated using bacteriophage, bacteria, archaea, and eukaryotic replication proteins before applying to Mimivirus. We show that a large fraction of mimiviral replication proteins, including polymerase B, clamp, and clamp loaders are of eukaryotic origin and are coevolving. Although phylogenetic analysis places some components along the lineages of phage and bacteria, we show that all the replication-related genes have been homogenized and are under purifying selection. Collectively our analysis supports the idea that Mimivirus originated from a complex cellular ancestor. We hypothesize that Mimivirus has largely retained complex replication machinery reminiscent of its progenitor while losing most of the other genes related to processes such as metabolism and translation.

Key words: **DNA replication, Mimivirus, giant viruses, evolution, phylogenetic, correlation analysis, MDS, coevolution, HGT, evolutionary selection, purifying selection, phylogenetic trees, NCLDVs, HGT, LUCA, LUCELLA.**

Article

Introduction

DNA replication is a highly complex and tightly regulated process. It plays a quintessential role in the transmission of genetic information from one generation to the next. Although the overall process of replication is conserved, replication, and error incorporation rates vary significantly between domains of life and viruses; and are linked to fitness and evolvability (Drake et al. 1998; Leipe et al. 1999; Loh et al. 2010; Lynch et al. 2016). The comparative genomics of proteins involved in DNA transcription and translation processes have shown that they are ubiquitous and conserved in all domains of life, whereas those involved in DNA replication are diverse (Leipe et al. 1999; Koonin 2003). Although DNA replication mechanisms exhibit substantial functional similarities, proteins involved in the bacterial and archaeal-eukaryotic machinery share less homology suggesting independent evolution from the last universal common ancestor (LUCA) (Leipe et al. 1999; Koonin 2003).

The discovery of giant viruses, their comparative genomics, and structural studies have further shed light on their role in the evolution of DNA replication machinery (Forterre 2006). Different theories on the origin of viruses have been proposed (Shackelton and Holmes 2004; Boyer et al. 2010; Koonin, Dolja, et al. 2015; Koonin and Yutin 2018) with some studies suggesting viral genes as a source of the cellular replication machinery (Villarreal and DeFilippis 2000; Forterre 2002). Although some viruses depend on their hosts for genome replication (Gelderblom 1996), viruses with large genomes like Mimivirus code for most of the informational proteins and can replicate their genomes independently (Raoult et al. 2004; Kazlauskas et al. 2016). Replication proteins of DNA viruses exhibit far more diversity between them than their hosts (Kazlauskas et al. 2016).

The origin and evolution of giant viruses are areas of intense research and several hypotheses have been proposed. As per some studies, either the viral world existed before the LUCA (Forterre and Prangishvili 2013), or viruses evolved from their cellular hosts by reductive evolution (Claverie 2006; Nasir et al. 2012a; Nasir and Caetano-Anollés 2015). Another study proposes to place viruses in the tree of life as a fourth domain and classified them as TRUC (things resisting uncompleted classification) (Boyer et al. 2010; Raoult 2013). A few studies have also suggested that the large DNA viruses evolved by host gene capture and gene

Table 1. Core Proteins of the DNA Replication Machinery of Bacteriophage (T4), Bacteria (*E. coli*), archaea, eukaryotes, and Mimivirus.

Function	Protein	Bacteriophage	Bacteria	Archaea	Eukaryotes	Mimivirus
Recognition of origin of replication	Origin binding protein	—	DnaA	Orc1	Orc1-6	gp1
Unwinding of duplex DNA helix	Helicase	gp41	DnaB	MCM	MCM2-7	gp229, gp8, gp612, gp635, gp132
	Helicase loader	gp59	DnaC	Cdc6	Cdc6, Cdt1	—
	Accessory proteins	—	—	GINS15, GINS23	Sld5, Psf1, Psf2, Psf3, Cdc45	—
Stabilization of unwound template strands	SSBP	gp32	SSBP	RPA	RPA70, RPA32, RPA14	gp544
RNA primer synthesis	Primase	gp61	DnaG	PriL, PriS DnaG[a]	Polα (A, B) Primase (Prim1, Prim2)	gp577, gp857, and gp229, gp8
DNA synthesis	Polymerase	gp43	Pol α, Pol ε, Pol θ	PolB1, PolD[b]	Pol δ (PolD1, PolD2, PolD3, PolD4), Polε (A, B, C)	gp351
Polymerase processivity factor	Sliding clamp	gp45	β	PCNA, PCNA3	PCNA	gp532, gp886, gp124
	Clamp loader	gp44, gp62	δ, δ′, ψ, χ, τ, γ	RFCL, RFCS	RFC 1-5	gp549, gp513, gp425, gp441, gp538
Joining Okazaki fragments	Ligase	gp30	LigA	LigI	LigI	gp331
Positive supercoil removal ahead of replication fork	Topoisomerase	gp52, gp39, gp60	Gyraseα, Gyraseβ (TopoII)	TopoI	Topo I, TopoII	gp243, gp515, gp216
Removal of RNA primers	RNase HI	RNase H1	RNaseHI	FEN1, RnaseH1	FEN1, RNaseH1	gp371, gp326, gp417

[a]Present in only crenarchaea.
[b]Present in all phyla except crenarchaea.

duplications events (Shackelton and Holmes 2004; Moreira and Brochier-Armanet 2008). Furthermore, recent studies have suggested that the large DNA viruses might have originated from small viruses with the acquisition of genes from eukaryotic hosts and other cellular organisms by horizontal gene transfer (HGT) (Koonin, Krupovic, et al. 2015; Koonin and Yutin 2018).

In this study, we have carried out the coevolutionary analysis of the putative components of the Mimiviral replication machinery to gain insights into its complexity and evolutionary origins. We also analyzed the correlation coefficient-based similarity matrix by multidimensional scaling analysis (MDS) to establish the network of coevolving proteins in Mimivirus by comparative analysis (Yin and Yau 2017). We have performed correlation and MDS analysis of DNA replication proteins from bacteriophage T4, *Escherichia coli*, partially characterized replication machinery of archaeon *Aeropyrum pernix*, and a set of highly conserved and interacting subunits of eukaryotic proteins (table 1) to validate the method. Our analyses of the mimiviral replication-related proteins show no evidence of recent HGTs, and that all the genes are under purifying selection. Based on the cues from phylogenetic, coevolution, and genome analyses, we conclude that the complexity of the mimiviral DNA replication machinery arose early suggesting a reductive evolution from a complex cellular ancestor.

Results

Mimiviral Replication Proteins Share Homology with Eukaryotic Components

To get a comprehensive picture of the phylogenetic tendencies of the mimiviral DNA replication machinery, sequence alignments, and phylogenetic trees were generated for the core replication proteins of Mimivirus (fig. 1). In addition to the cellular domains, we selected phage T4 for the analysis as both T4 and Mimivirus DNA polymerases belong to the B family and T4 codes for minimal but complete replication machinery (Karam and Konigsberg 2000; Raoult et al. 2004). A set of 20 representative sequences (four each from bacteria, archaea, eukaryotes, phage T4 family, and *Mimiviridae* family, supplementary table S1, Supplementary Material online) of replication proteins were selected as discussed in the Materials andMethods for sequence alignment and phylogeny construction (fig. 1). We constructed unrooted phylogenetic trees in four different ways, namely maximum-likelihood (ML), neighbor-joining (NJ), minimum evolution

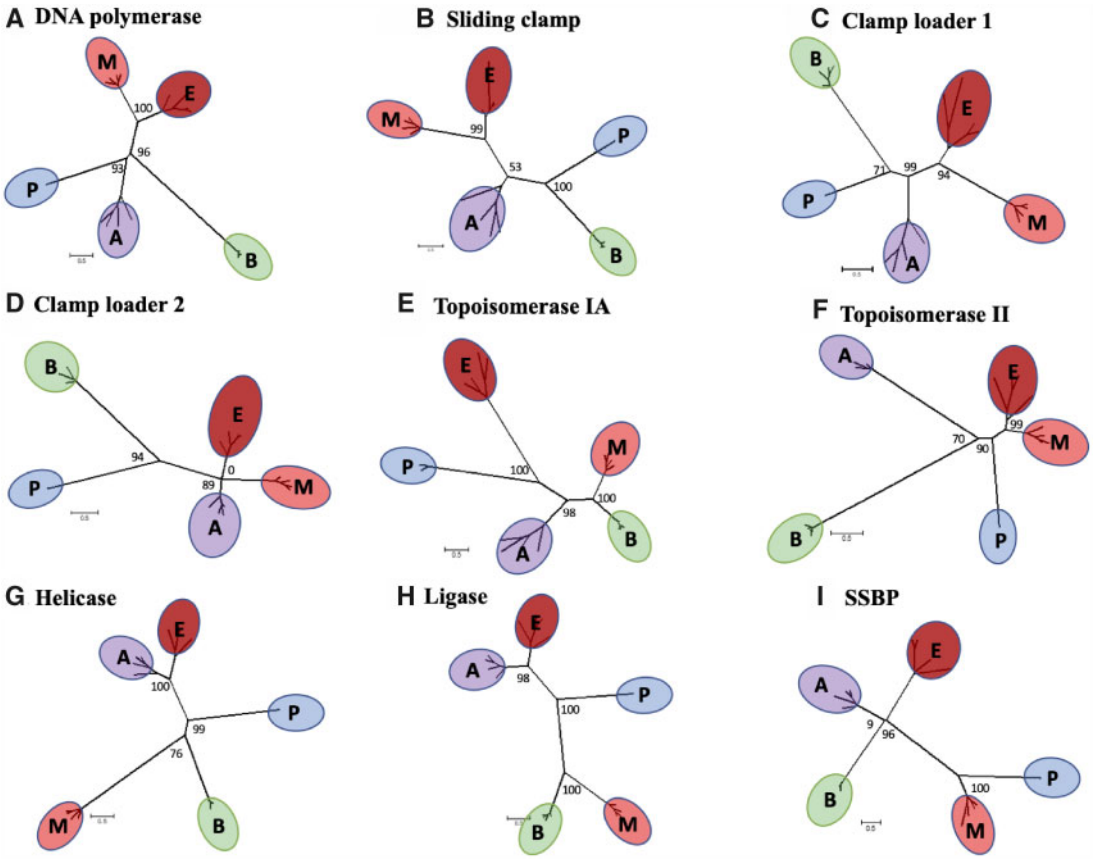

FIG. 1. Phylogenetic trees of proteins from the cellular domains (eukaryotes, bacteria, archaea) and phage T4 and Mimivirus. The multiple sequence analysis was performed by MUSCLE in MEGA6.0 and trees are built by FastTree v.2.1. The phylogenetic trees of (*A*) DNA polymerase, (*B*) sliding clamp, (*C*) clamp loader 1, (*D*) clamp loader 2, (*E*) topoisomerase IA, (*F*) topoisomerase II, (*G*) helicase, (*H*) ligase, and (*I*) SSBP are constructed using four sequences from each domain and viral families, accession number of protein sequences used to construct trees are included in the supplementary table S2, Supplementary Material online. Red, *Mimiviridae* (M); maroon, eukaryotes (E); green, bacteria (B); blue, T4-like phages (T); purple, archaea (A).

(ME), and unweighted pair group method with arithmetic mean (UPGMA) (supplementary fig. S1, Supplementary Material online). As all the phylogenetic trees displayed the same topology, we have shown only the trees constructed by the ML method (fig. 1).

Although eukaryotic, archaeal, and phage DNA polymerase considered in the study belong to the family B polymerase, mimiviral polymerase showed an evolutionary relationship with eukaryotic polymerase B rather than phage T4 polymerase (fig. 1*A*). Further, phylogenetic trees of the sliding clamp and two clamp loaders also exhibited similar topology (fig. 1*B*–*D*). Mimivirus genome encodes five clamp loaders akin to the eukaryotic replication system. The phylogenetic analysis suggested common ancestry for mimiviral and eukaryotic DNA polymerase and its processivity factors.

Mimivirus encodes three topoisomerases, namely, IA (gp243), IB (gp216), and II (gp515). In this study, based on their presence in all *Mimiviridae* family members, we have considered only IA and II for the analysis. Topoisomerase IA is annotated as a bacterial-type that has not been reported in a virus before (Raoult et al. 2004). Phylogenetic tree of topoisomerase IA showed that the *Mimiviridae* family shared a clade with bacteria, whereas topoisomerase II shared a clade with eukaryotic sequences (fig. 1*E* and *F*). Though we have not carried out a phylogenetic analysis of topoisomerase IB, it was earlier shown that the type IB topoisomerase of Mimivirus is functionally homologous to that of poxvirus despite their primary structural similarity with bacteria (Benarroch et al. 2006). Thus, it appears that the mimiviral topoisomerases have a complex evolutionary history.

Although the replicative helicase of the *Mimiviridae* family and all Nucleocytoplasmic Large DNA virus (NCLDVs) is of SF3 type associated with a primase domain of the archaeo-eukaryotic primase (AEP) family (Iyer et al. 2005; Kazlauskas et al. 2016; Gupta et al. 2017), their counterparts in archaea/eukaryotes and bacteria belong to SF6 and SF4 families, respectively. SF4 and SF3 helicases appear to be the most abundant (44% and 41%, respectively, of all replicative helicase) among viruses (Raoult et al. 2004). SF4 helicases are encoded mostly by phages that are homologous to *E. coli* DnaB or RepA, whereas SF3 helicases are encoded by some phages and eukaryotic viruses (Kazlauskas et al. 2016). The tree showed a divergent topology with little inclination of mimiviral helicase to bacterial helicases (fig. 1*G*). The mimiviral SF3 helicase might have been present in an ancestral NCLDV that was passed on vertically to all large DNA viruses.

DNA ligase is an important component of the DNA replication machinery. In the case of NCLDVs, whereas

mimiviruses, iridoviruses, and entomopoxviruses encode NAD-dependent ligase, other NCLDVs either encode an ATP-dependent ligase or do not encode ligase. Furthermore, although the NAD-dependent ligase of NCLDVs turned out to be monophyletic, the ATP-dependent ligases present in many other NCLDVs showed more diversity and exhibited different phylogenetic tendencies. Based on this observation, it was speculated that the NAD-dependent ligase is ancestral to NCLDVs (Koonin and Yutin 2010). Since our data set was restricted to mimiviruses that code for NAD-dependent ligase, as expected, they showed a cladistic relationship with bacteria (fig. 1*H*).

The presence of canonical OB-fold single-strand DNA-binding proteins (SSBP) in *Mimiviridae* (gp544 in Mimivirus) and other NCLDVs has been established by computational analysis (Kazlauskas and Venclovas 2012). A phylogeny of SSB proteins constructed from bacterial SSBP, archaeal, and eukaryotic replication protein A (RPA), phage T4 gp32-like and mimiviral SSBPs showed that *Mimiviridae* and phage proteins share common ancestry (fig. 1*I*). It was previously reported that Mimivirus possesses a phage T7-like SSBP (Kazlauskas et al. 2016).

The phylogenetic analysis here and reported elsewhere (Raoult et al. 2004; Benarroch and Shuman 2006; Kazlauskas et al. 2016) showed that the Mimivirus replication machinery is a composite of components rooted in eukaryotic, bacterial, and phage lineages or with their ancestor. Previous studies have suggested the horizontal acquisition of DNA ligase by eukaryotic DNA viruses from a bacterium (Benarroch and Shuman 2006) that is consistent with our observation. In addition, mimiviral topoisomerase IA which is earlier reported as a bacterial-type forms a clade with bacteria in our phylogeny. Our previous analysis of the primase-helicase bifunctional protein, which is one of the hallmark proteins of NCLDVs, suggested that it might have been present in an ancestral virus and passed on to NCLDVs by vertical transfer (Gupta et al. 2017). Furthermore, an earlier hypothesis suggested DNA viruses as a source of eukaryotic replication proteins (Villarreal and DeFilippis 2000) including topoisomerases IIA (Forterre and Gadelle 2009).

Out of the nine core mimiviral replication components we analyzed, five clades with eukaryotes, whereas two with bacteria, one with phage and a distinct protein. Although multiple loss and gain events further complicate inferring phyletic patterns of mimiviral replication-related proteins, our analysis, when taken together, compositionally, Mimivirus replication machinery shows a strong affinity toward eukaryote-like replication machinery.

Coevolution of the DNA Replication Machinery

Proteins that function together through direct or indirect interactions tend to coevolve to compensate for changes in the partner (Goh et al. 2000). DNA replication and its regulation demand a series of precise interactions between its components, and hence they tend to coevolve (Nossal 1992; Schaeffer et al. 2005). We have taken the coevolutionary analysis approach to assess the interdependencies of components of mimiviral replication machinery that appears to be assembled from different sources.

Establishing Correlation/MDS Analysis as a Method to Study Multiprotein Complexes

Since biochemistry, structure, and interactions of DNA replication machinery of bacteriophage T4 and *E. coli* have been very well characterized (Nossal 1992; Schaeffer et al. 2005), we have used these data sets as controls to validate our analysis. We have also extended our analysis to representative eukaryotes and archaea. Based on the literature, we have considered 13 components of the replication machinery from T4-like phage (fig. 2*A*, *B* and supplementary table S3, Supplementary Material online), 22 from bacteria (fig. 2*C*, *D* and supplementary table S4, Supplementary Material online), 18 from archaea (fig. 3*A*, *B* and supplementary table S5, Supplementary Material online), and 25 from representative eukaryotes (fig. 3*C*, *D* and supplementary table S6, Supplementary Material online) for the analysis. Thus, proteins from phage T4, *E. coli*, *A. pernix*, and *Homo sapiens* were used as seed sequences for PSI-BLAST to retrieve sequences from the respective family of proteins (supplementary table S2, Supplementary Material online).

The Pearson correlation coefficients were calculated from the distance matrices generated by the multiple sequence alignments (MSA) as described in the Materials and Methods section and represented by correlation matrix tables and heatmaps. The correlation coefficient (r) >0.7 suggests functional dependency and coevolution, whereas coefficients <0.7 represent divergence (Goh et al. 2000; Pazos and Valencia 2001). The MDS analysis is based on the similarity matrix of the Pearson correlation coefficients.

Phage T4. Phage T4 is one of the simple model replication systems and has been studied extensively to understand the DNA replication mechanism and interactions of the phage-coded components of the replication machinery (Nossal 1992). When all the known and characterized components of the T4 replication machinery were analyzed, the correlation coefficients between all combinations were found to be >0.84 (fig. 2*A* and supplementary table S3, Supplementary Material online). We have also illustrated the protein–protein interactions by correlation analysis, and correlation coefficients have been represented in two dimensions by MDS, which show proteins with high correlation values are placed closer to each other on the plot (fig. 2*B*). Distances of all phage proteins calculated by MDS show their functional dependencies and essential roles in the replication system. The group of highly conserved proteins with indispensable functions in the phage T4 genome replication was found to be highly coevolving.

Escherichia coli. The complexity of replication machinery increases with genome size. The phage T4 with a genome of ~169 kb (Miller et al. 2003) carries a complete but a minimal set of proteins whereas the replication machinery of bacteria that has a much larger genome (the size of *E. coli* genome, e.g., is 4,639 kb) (Blattner et al. 1997), requires the

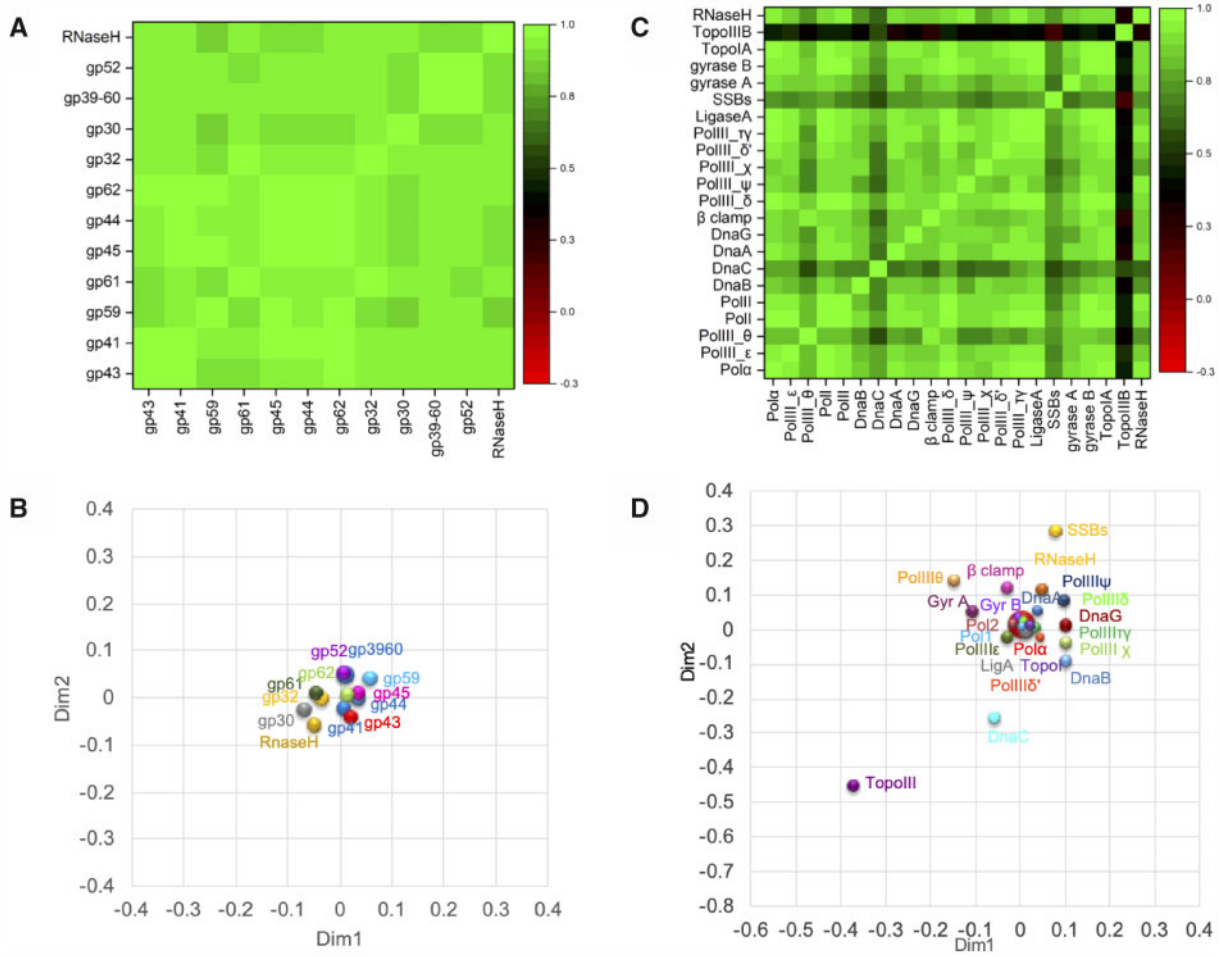

FIG. 2. Components of the phage T4 and bacterial replication machinery show evidence of coevolution. Coevolution of protein complexes has been analyzed by Pearson correlation coefficient and multidimensional scaling analysis (MDS). (*A*) A matrix of 13 replication proteins of phage T4 with correlation coefficients shown as a heatmap and (*B*) the MDS analysis of the same set of proteins showing the clustering of proteins in the 2D space. DNA polymerase (gp43), helicase (gp41), helicase loader (gp59), primase (gp61), sliding clamp (gp45), clamp loaders (gp44/gp62), SSBP (gp32), ligase (gp30), topoisomerase (gp39, gp52, and gp60), and RNase H. (*C*) A correlation matrix of 22 bacterial core replication proteins correlation coefficients displayed as a heatmap and (*D*) MDS analysis of the same set of proteins display distances of proteins in the 2D space. The goodness of fit of MDS assessed by Shepard diagram (supplementary fig. S3, Supplementary Material online) and smaller Kruskal's stress (1) suggests the better MDS representation (supplementary table S10, Supplementary Material online).

participation of at least 30 core proteins (table 1) (Schaeffer et al. 2005). Several interaction studies have been carried out to gain insights into the replication process of bacteria (Wickner and Hurwitz 1975; Maki et al. 1988; Kornberg and Baker 1992; Lu et al. 1996; Marceau et al. 2011; Bhardwaj et al. 2018). Most bacterial replication proteins exhibit high correlation coefficients with each other that support the experimental evidence of interaction (fig. 2C and supplementary table S4, Supplementary Material online). The interaction studies of gyrases (topoisomerase II), topoisomerase IA, and topoisomerase IIIB with other replication proteins have not yet been carried out but, gyrases A and B play a role in the removal of positive supercoils ahead of the replication fork (Champoux 2001). From our correlation analysis, topoisomerase I and II were found to be coevolving with other replication proteins, whereas type III shows less conservation, which suggests that it may not be part of the functional replication complex (fig. 2C).

The visualization of correlation coefficient-based distances between bacterial proteins in 2D space by MDS suggests their functional dependencies (fig. 2D). A cluster of DNA polymerase core complex with sliding clamp and clamp loaders of polymerase holoenzyme supports experimental analysis, which is represented by smaller distances. Primosome complex of DnaB and DnaG and initiator protein DnaA cluster with other proteins except for DnaB-interacting loader DnaC. MDS analysis further suggests that auxiliary proteins SSBPs, RNase H, and ligase are also an integral part of the replication process. Distances in the cluster represented in the 2D space of these proteins may correspond to their functional role and indicate their coevolution. Topoisomerase III is an outlier and shows no coevolution with other replication proteins. It appears that topoisomerase III may not take part directly in replication.

Archaea. The information system of archaea has properties of both bacterial and eukaryotic types. We have considered

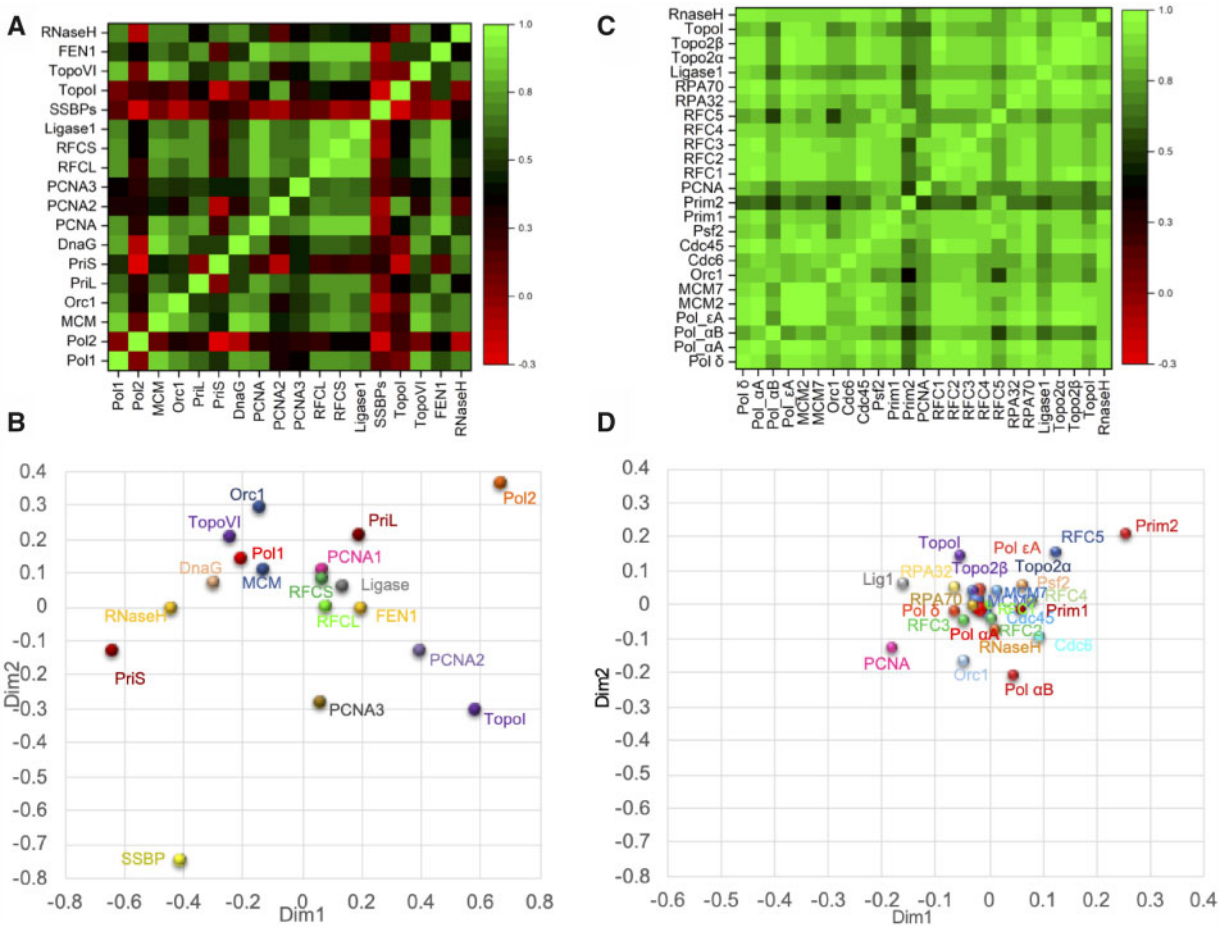

FIG. 3. Coevolutionary analysis of archaeal and eukaryotic replication machinery. Coevolution of proteins of the *A. pernix* archaeal replication machinery shows diversity among replication components. (*A*) A correlation matrix of 18 replication proteins of the archaeon is displayed as a heatmap and (*B*) the MDS analysis of the same set of proteins show distances of proteins in the 2D space with scattered pattern suggesting the diversity. The eukaryotic replication machinery components show coevolution analyzed by (*C*) a correlation matrix of replication proteins displayed as a heatmap and (*D*) the MDS analysis of the same set of proteins showing distances of proteins in the 2D space. The goodness of fit of MDS assessed by Shepard diagram (supplementary fig. S3, Supplementary Material online) and smaller Kruskal's stress (1) suggests the better MDS representation (supplementary table S10, Supplementary Material online).

Aeropyrum pernix, a crenarchaeon, as an example to understand the coevolution of archaeal DNA replication proteins. The archaeal DNA replication system is closely related to the eukaryotic system, but the replication proteins within the archaeal domain are highly diverse (Sarmiento et al. 2014). Although extensive biochemical characterization of the components of the archaeal replication machinery has not yet been carried out, a few studies have given insights into their functions and diversity (Cann et al. 1999; Daimon et al. 2002; Imamura et al. 2007; Atanassova and Grainge 2008; Lang and Huang 2015).

The archaeal replication system shows higher diversity, and only the core proteins were found to have coevolved (fig. 3*A* and supplementary table S5, Supplementary Material online). Our analysis shows that although the minichromosome maintenance (MCM), origin recognition complex (Orc1), polymerase B (Pol1), proliferating cell nuclear antigen, PCNA1 (but not PCNA2 and PCNA3), and both large and small replication factor C (RFC) proteins have coevolved, Pol2 may not have coevolved with this group of proteins. The SSBP surprisingly shows no evidence of coevolution despite being a part of the replication complex. Our analyses further show that only topoisomerase VI is part of the replication complex and coevolving, whereas topoisomerase I is not (fig. 3*A* and supplementary table S5, Supplementary Material online).

MDS analysis (fig. 3*B*) further suggests that unlike phage T4 and *E. coli*, only some components of the archaeal system are coevolving, which might reflect the significant mechanistic differences in DNA replication of bacteria and archaea. Biochemical evidence is lacking for the interaction of most replication-related proteins of *A. pernix* or any other archaeon. The MDS analysis shows that PCNA1, RFCL, RFCS, flap endonuclease1 (FEN1), and ligase1 with PriL form a cluster alongside MCM, Pol1, Orc1, DnaG, TopoVI (fig. 3*B*).

Eukaryotes. With much larger genomes and multiple origins of replication, the genome replication process in eukaryotes is quite complex and different from bacteria and requires many proteins with several subunits. For this study, we have considered a few representative subunits of the core replication

proteins from representative eukaryotes (supplementary table S2, Supplementary Material online). Our analysis suggests that eukaryotic replication-related proteins are coevolving (r ~0.7 to 0.9) (fig. 3C and supplementary table S6, Supplementary Material online). Most of the core replication proteins considered for the analysis (PolαA, MCM2, RAP70, PolεA, Topo2 α and β subunits, RFC1, RFC2, and Cdc45 with MCM7, RPA32, Pol δ, RFC3, RNase H, Cdc6, Prim1, RFC4, and Psf2) are closely placed in the MDS plot suggesting their functional dependencies (fig. 3D). The analysis further suggests that polymerases PolαA, Polδ, and Polε subunits might have coevolved with accessory replication proteins, clamp and clamp loaders, ligase, SSBP, and RNase H. The initiation complex of MCM2, MCM7, Cdc6, Cdc45, Orc1, Pif2, and Prim1, except Prim2, supports the functional conservation and coevolution (Koonin 1993; Labib and Diffley 2001; Zannis-Hadjopoulos et al. 2004; Frigola et al. 2017). Both topoisomerase II subunits appear to coevolve with the rest of the components (fig. 3D).

In sum, coevolutionary analysis is consistent with experimental findings and can be used as a tool to gain insights into the makings of large uncharacterized complexes such as the replication machinery of Mimivirus.

Applying the Established Method to Mimiviridae Proteins

For the coevolution analysis, we have considered a set of 15 viruses of the *Mimiviridae* family representing all subfamilies and lineages including Klosneuvirinae, except subfamily II as viruses of this subfamily do not carry all the genes required for genome replication (supplementary table S7, Supplementary Material online).

At least 21 of the annotated mimiviral proteins appear to play a role in DNA replication (supplementary table S8, Supplementary Material online). Except for DNA ligase and topoisomerase IB, none of the other proteins in supplementary table S8, Supplementary Material online, have been characterized (Benarroch and Shuman 2006; Benarroch et al. 2006). For the analysis, we have not considered the standalone primases, gp577 and gp857 as biochemical characterization showed that mimiviral gp577 is a primase-polymerase (PrimPol) that is involved in repair rather than replication and gp857 appears to be a paralog of gp577 (Gupta et al. 2019). We have also not considered topoisomerase IB (gp216) and one of the PCNAs (gp124) as they are absent in some viruses of the *Mimiviridae* family. We have taken a putative nuclease (gp456) of unknown function as a negative control for the correlation and MDS analysis.

Two of the NCLDV core proteins, DNA polymerase (gp351) and D5-like primase-helicase bifunctional protein (gp299), returned an r of 0.911 (fig. 4A and supplementary table S9, Supplementary Material online) indicating coevolution with other replication proteins. Furthermore, a high correlation coefficient ($r > 0.95$) of DNA polymerase with PCNA (gp532) and RFCs (gp549 and gp441) and an intermediate r (0.63–0.73) with the other PCNA (gp886) and RFCs (gp425, gp513, and gp538) were observed (fig. 4A and supplementary table S9, Supplementary Material online). The correlation coefficients among PCNA, gp532, and RFCs gp549 and gp441 were also significant (supplementary table S9, Supplementary Material online). The PCNA (gp886) and RFCs (gp425, gp513, and gp538) showed comparatively lesser r with DNA polymerase, among themselves and with the proteins of the gp532/gp549/gp441 complex. In terms of relatedness and overall configuration, the DNA replication system of Mimivirus seems to share more similarities with archaea/eukaryotes than bacteria.

The SSB protein gp544, which is thought to be derived from phage (Kazlauskas and Venclovas 2012), also exhibited a high correlation coefficient with DNA polymerase and processivity factors (gp532/gp549/gp441) (supplementary table S9, Supplementary Material online). Both RNase H proteins (gp371 and gp326) and FEN1 (gp417) appear to have coevolved with DNA polymerase and ligase1 (gp331). Topoisomerase I (gp515) and II (gp253) that are conserved in the *Mimiviridae* family returned significant correlation coefficients with other proteins (fig. 4A and supplementary table S9, Supplementary Material online). Mimivirus codes for 4 DNA helicases (gp8, gp612, gp635, and gp132) belonging to either SF1 or SF2 families. Interestingly, all of them, except gp8,exhibited significant correlation coefficients suggesting their coevolution with other replication proteins (supplementary table S9, Supplementary Material online). Although these are putative helicases, they might play a role in replication as well as repair (Fridmann-Sirkis et al. 2016).

We have identified a group of proteins represented by small distances in the space that includes D5 primase-helicase (gp299), DNA polymerase (gp351), SSBPs (gp544), PCNA (gp532), RFC (gp549, gp441), ligase (gp331), FEN1 (gp417), RNase H (gp371 and gp326), topoisomerase I and II (gp515, gp253), and helicases (gp635, gp612, and gp132) (fig. 4B). The second PCNA homolog (gp886) and RFC homologs (gp425, gp513, gp538) are represented on the MDS plot by larger distances (fig. 4B). Two proteins, namely, gp1 and gp8, are annotated as a putative origin binding protein and a putative helicase, respectively. The correlation analysis and MDS representation suggest that these two proteins might not be coevolving with the mimiviral replication complex (fig. 4A and B).

We also constructed phylogenetic trees of mimiviral proteins such as gp351, gp229, gp515, gp243, gp532, gp441, gp549, gp331, gp544, and gp1 proteins from the same data set used for the correlation analysis (fig. 4C, accession numbers are given in supplementary table S11, Supplementary Material online). The topology of each tree, except gp1, showed a cophylogenetic mirror pattern which agrees with coevolutionary studies carried out by correlation and MDS analyses.

As we have shown that the replication complex of Mimivirus is coevolving with few exceptions, we extended the coevolutionary study to putative proteins involved in other essential DNA processes. We considered putative proteins involved in DNA repair (gp389: putative DNA mismatch

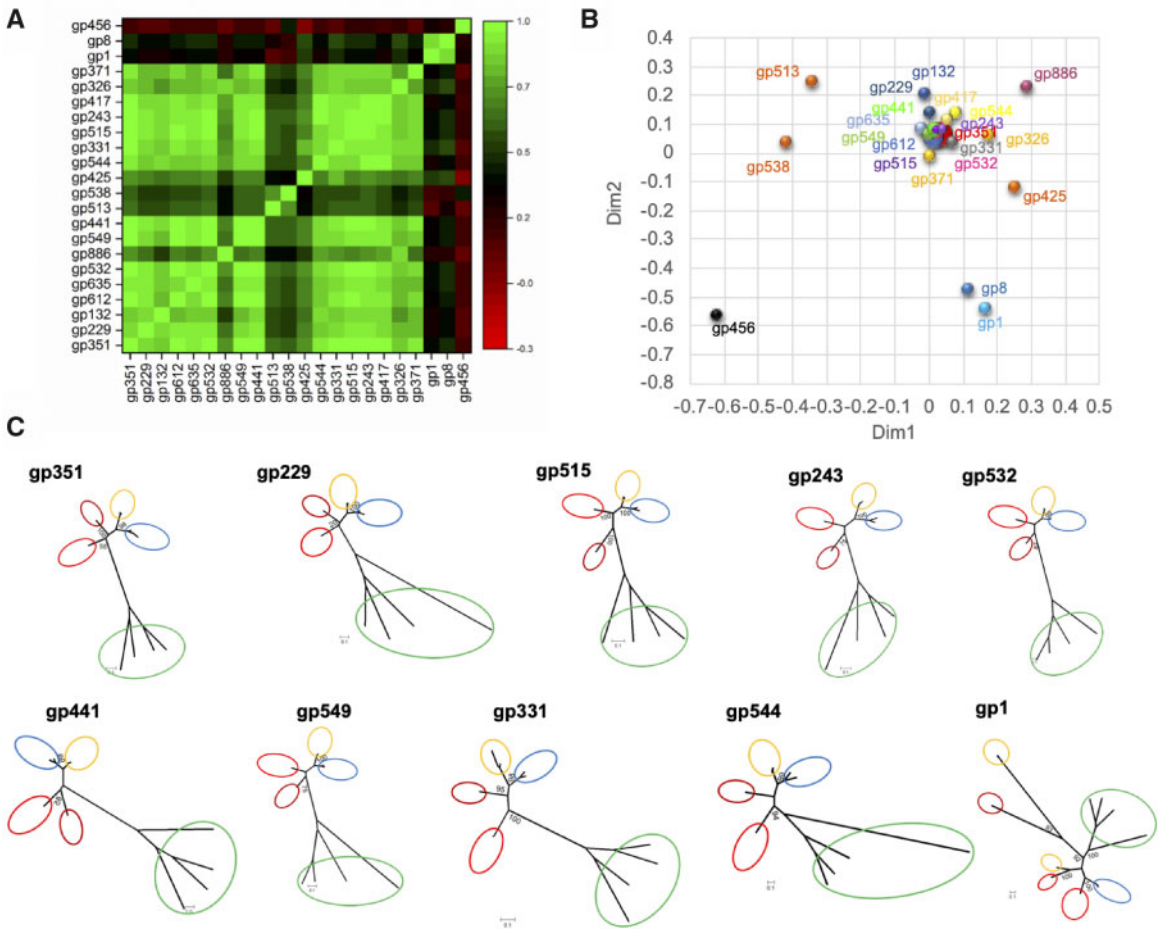

FIG. 4. Components of the *Mimiviridae* replication machinery show evidence of coevolution. (*A*) A matrix of 21 replication proteins of Mimivirus with correlation coefficients shown as a heatmap and (*B*) the MDS analysis of the same set of proteins showing the cluster of proteins in 2D space. The goodness of fit of MDS assessed by Shepard diagram (supplementary fig. S3, Supplementary Material online) and smaller Kruskal's stress (1) suggests the better representation of MDS analysis (supplementary table S10, Supplementary Material online). (*C*) Maximum likelihood phylogenetic trees of mimiviral proteins gp351, gp229, gp515, gp243, gp532, gp441, gp549, gp331, gp544 except gp1 showed cophylogenetic mirror pattern, accession numbers of respective proteins are given in supplementary table S10, Supplementary Material online, Red, *Mimiviridae* I lineage A; blue, *Mimiviridae* I lineage B; yellow, *Mimiviridae* I lineage C; maroon, *Mimiviridae* III; green, Klosneuvirinae. Putative helicase (gp229), putative replication origin-binding protein (gp1), putative ATP-dependent DNA helicase (gp612), putative helicase (gp635), putative helicase (gp8), putative helicase (gp132), DNA topoisomerase 1 (gp243), DNA topoisomerase 2 (gp515), DNA ligase (gp331), DNA polymerase (gp351), putative proliferating cell nuclear antigen (gp886), probable DNA polymerase sliding clamp (gp532), putative replication factor C small subunit (gp549), putative replication factor C small subunit (gp513), putative replication factor C small subunit (gp425), putative replication factor C large subunit (gp441), putative replication factor C small subunit (gp538), probable ribonuclease H protein (gp326), probable ribonuclease 3 (gp371), hypothetical protein-SSBPs (gp544), putative endonuclease of the XPG family (gp417), and putative nuclease (gp456).

repair protein MutS-like protein, gp271: probable uracil-DNA glycosylase, R555: Mre11/Rad50 complex), genome packaging (gp468: A32-like virion packaging ATPase), signal transduction pathways (gp254: putative serine/threonine-protein kinase), biosynthesis of deoxyribonucleotides (gp342: ribonucleoside-diphosphate reductase large subunit), transcription (gp540, gp266: DNA-directed RNA polymerase subunits 1 and 2), and translation (gp492: putative translation initiation factor 4a) in addition to DNA replication proteins (gp351: DNA polymerase, gp229: putative helicase, gp515: DNA topoisomerase II) (supplementary table S12, Supplementary Material online). The correlation analysis of these proteins returned an $r > 0.7$ and, the MDS analysis also reflected the distance proximity between proteins in a 2D plot (fig. 5*A* and *B*; supplementary table S13, Supplementary Material online). Hence, we found that proteins involved in diverse functions are also coevolving.

Horizontal Gene Transfer and Evolutionary Selection Analysis

Despite the chimeric nature of its components, Mimivirus replication machinery appears to be coevolving; and the predicted coevolutionary network is comparable to the other well-characterized systems. To identify the horizontally acquired genes, we calculated GC content, GC3s, Codon Adaptation Index (CAI), and the Effective Codon Number (Nc) of Mimivirus replication-related genes (table 2). It is proposed that the nucleotide composition of a gene acquired from other sources might vary from that of the recipient genome, and it has been widely studied in the case of bacterial

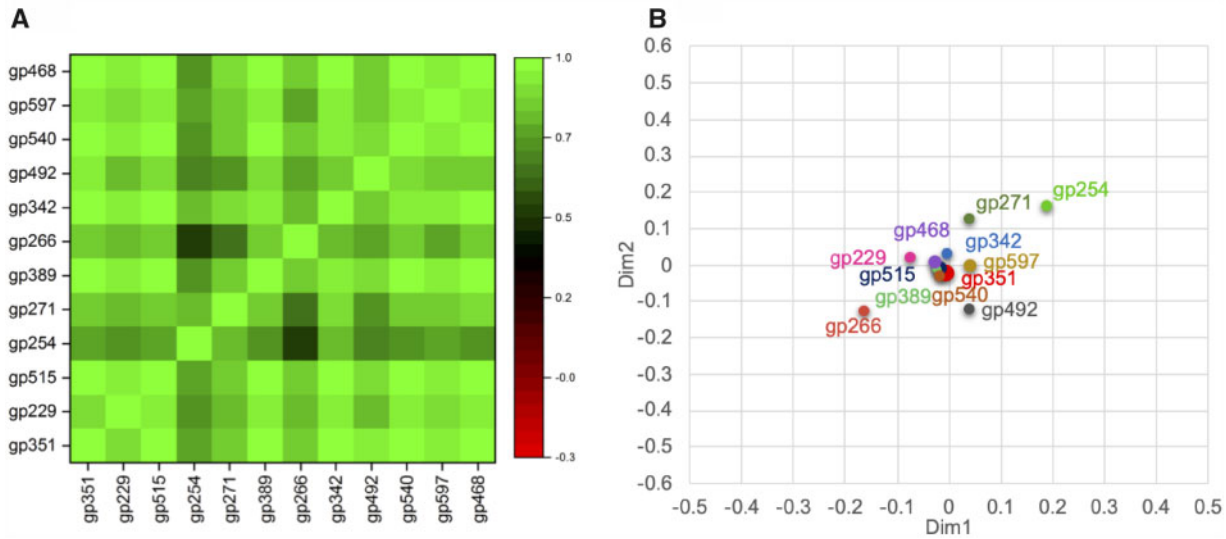

FIG. 5. Representative proteins involved in different processes like DNA replication, transcription, translation, DNA repair, and genome packaging of *Mimiviridae* family viruses show an evidence of coevolution. (*A*) A matrix of correlation coefficients of proteins shown as a heatmap and (*B*) the MDS analysis of the same set of proteins showing the cluster of proteins in 2D space. DNA polymerase (gp351), putative helicase (gp229), DNA topoisomerase 2 (gp515), putative serine/threonine-protein kinase (gp254), probable uracil-DNA glycosylase (gp271), putative DNA mismatch repair protein MutS-like protein (gp389), DNA-directed RNA polymerase subunit 2 (gp266), ribonucleoside-diphosphate reductase large subunit (gp342), putative translation initiation factor 4a (gp492), DNA-directed RNA polymerase subunit 1 (gp540), uncharacterized hydrolase (gp597), and A32 virion packaging ATPase (gp468).

Table 2. Gene Length, GC3s, and CAI of Mimivirus Replication Genes.

Protein	Gene	Gene Length (bp)	GC3s	CAI	dN/dS by M0 (one-ratio model)
gp351	R322	5,262	0.156	0.163	0.018
gp515	R480	3,831	0.185	0.171	0.015
gp132	L115	3,741	0.165	0.149	0.048
gp8	R8	3,108	0.206	0.167	0.031
gp612	R568/R569	3,096	0.173	0.161	0.015
gp229	L207/L206	2,970	0.315	0.17	0.015
gp635	R592	2,641	0.187	0.181	0.011
gp243	L221	2,629	0.13	0.166	0.020
gp1	R1	2,418	0.197	0.203	0.028
gp331	R303	1,920	0.131	0.148	0.006
gp371	R343	1,676	0.331	0.109	0.009
gp441	R411	1,675	0.324	0.098	0.007
gp417	L386	1,570	0.204	0.157	0.004
gp532	R493	1,436	0.33	0.193	0.003
gp544	R505	1,318	0.223	0.205	0.030
gp513	L478	1,190	0.283	0.094	0.007
gp549	R510	1,126	0.286	0.093	0.007
gp538	L499	1,112	0.135	0.15	0.006
gp886	L823	1,039	0.154	0.17	0.005
gp425	R395	1,016	0.322	0.112	0.007
gp326	R298/R299	996	0.433	0.164	0.023

genomes (Lawrence and Ochman 1997). Our analysis also showed the codon biasness of *E. coli* replication genes represented by CAI values ranging from 0.258 to 0.602 with a Pearson skewness of 0.898 and variation in the GC content (49.1–58.2%). Although the GC content of phage T4 genes (28.9–37.4%) showed homogeneity, variation was observed in GC contents of *Homo sapiens* (38.6–65%) and *A. pernix* (50.2–64.1%). The Nc values of phage and cellular genes, with few exceptions, showed low codon usage biasness (supplementary tables 14 and 15, Supplementary Material online).

The GC content analysis of mimiviral replication genes showed a uniform base composition with the overall genome composition of 28% GC (fig. 6*A* and supplementary table 14, Supplementary Material online). Furthermore, the Nc values of all the replication genes also showed limited variance in the codon biasness of all genes (maximum: 54.14, minimum: 35.03, R2: 0.515). This was also reflected in the analysis of CAI variance, 0.001 calculated from CAI, which showed a Pearson skewness of −0.532, indicating a significantly homogenous data set (table 2 and supplementary table 15,

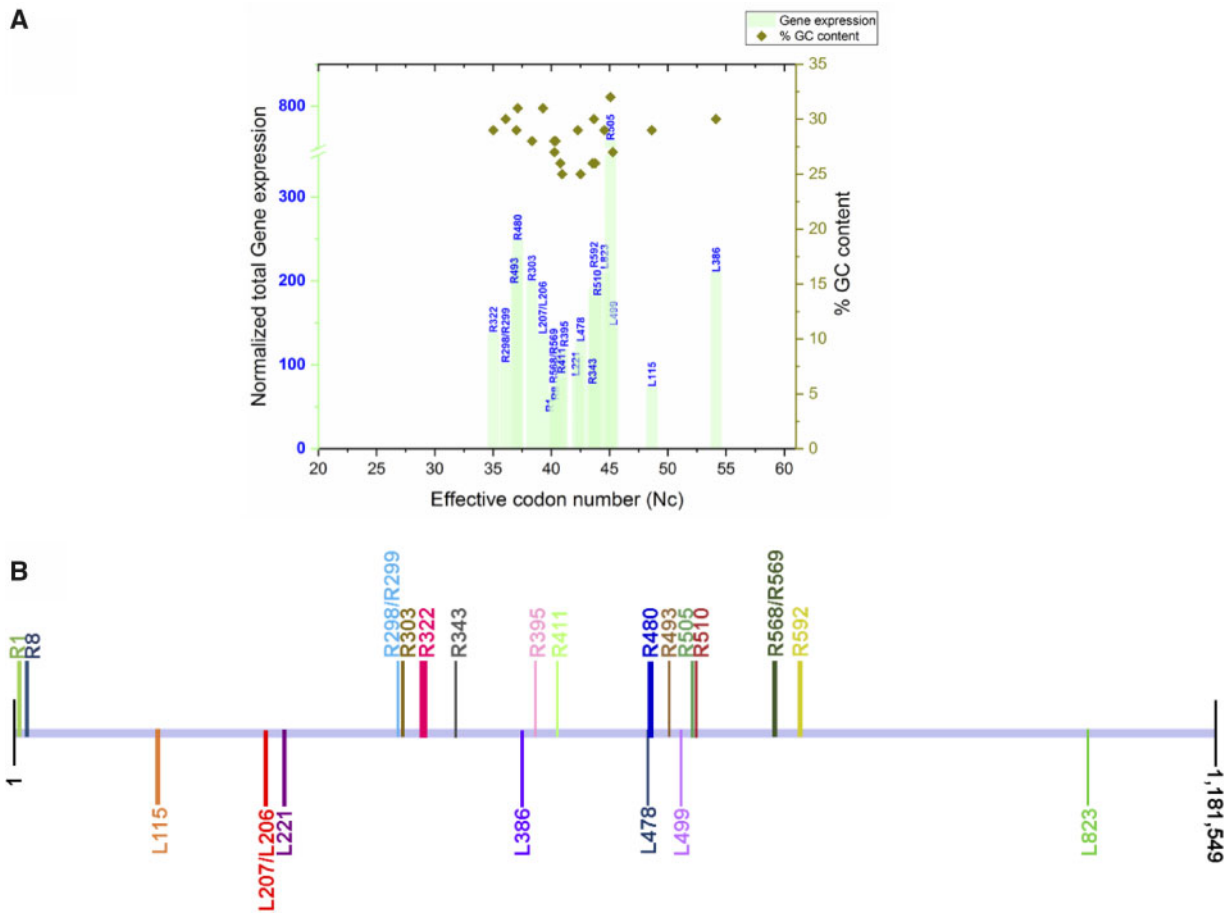

FIG. 6. Mimiviral DNA replication genes are homogeneous and are under purifying selection. The replication genes analyzed (*A*) for their normalized expression level, %GC content with the effective codon number (Nc), and d*N*/d*S* show no evidence of HGT, and (*B*) most of the mimiviral replication genes are located toward the center of the genome.

Supplementary Material online). When compared with CAI values of phage and cellular organisms, Mimivirus genes returned low values (supplementary fig. S4, Supplementary Material online). Although we did not detect any evidence of a recent HGT for mimiviral replication genes, the phylogenetic analysis (fig. 1) showed maximum homology with eukaryotic proteins. We analyzed putative lipocalin of Mimivirus (R877) which is proposed to be horizontally transferred from proteobacteria (Moreira and Brochier-Armanet 2008). Our analysis also showed the gene's codon biasness supported by a low Nc value of 31.11, whereas the CAI and GC content were not discernible (supplementary table S14, Supplementary Material online).

As mimiviral SSBP is of phage, ligase and topoisomerase are of bacterial, and the rest of the proteins exhibit eukaryotic affinities, we investigated the evolutionary pressure on the genes of the coevolving proteins. The evolutionary selection pressure on replication genes was estimated by comparing codon substitution models using the likelihood ratio test (LRTs) (table 2). The CodeML analysis using the site-specific model suggested that all the genes are under purifying selection with a d*N*/d*S* < 1, supported by the *P* value of LRT by comparing four pairs of models (table 2 and supplementary table S16, Supplementary Material online). The result suggests that these genes fall under the nucleotide sequences of coding regions of high functional constraints and the encoded amino acid sequences have been highly conserved.

Further, we recently showed that most of the core genes of amoebal giant viruses are in the central part of the genome and are flanked by repeat domain-containing proteins (RDCP) genes, which could have led to the genome expansion of mimiviruses (Shukla et al. 2018). We traced the location of replication genes in the mimiviral genome and found that most of them are located toward the center of the genome (fig. 6*B*).

Discussion

With the help of coevolutionary analysis and MDS, we show here that the core components of the Mimivirus replication machinery are coevolving. The highly correlated evolutionary network observed suggests that the robust functional complementarities of mimiviral replication proteins have evolved over longer evolutionary time scales and helps in assessing the evolutionary origins and the ancestral state of mimiviruses. Furthermore, evolutionary selection analysis showed that the replication-related genes of mimiviruses are compositionally homogenous and are under purifying selection. Besides, lack of evidence of recent horizontal transfer among replication

machinery components further suggests that the complexity of the Mimivirus replication machinery arose early, supporting the hypothesis of Mimivirus evolution from a complex ancestor.

What would the nature of such a complex ancestor be? Ever since their discovery, the origin and evolution of mimiviruses have been a subject of intense debate. Reconstructing the evolutionary path toward a large genome, as seen in the case of mimiviruses and other giant NCLDVs, is further complicated by varying phylogenetic affinities of their proteins, including the core proteins involved in DNA replication. Many researchers have been trying to shed light on the origin and evolution of giant viruses, and many hypotheses have been put forth (Shackelton and Holmes 2004; Boyer et al. 2010; Koonin and Yutin 2010; Koonin, Dolja, et al. 2015; Koonin and Yutin 2018). A comprehensive analysis of 45 NCLDV genomes from six different families by Yutin et al. (2009), helped identify a core set of 47 conserved genes that were thought to be present in the common ancestor (Yutin et al. 2009; Koonin and Yutin 2010). This included some of the hallmark genes such as the B-family DNA polymerase, DNA primase, and SF2 helicase that were shared by non-NCLDVs such as herpesviruses and baculoviruses (Koonin et al. 2006). However, with the recent discovery of a large number of diverse groups of NCLDVs, the core set has been reduced to only three to five genes (Koonin and Yutin 2018; Guglielmini et al. 2019); furthermore, recent phylogenomic analyses suggested multiple evolutionary origins for giant viruses (Koonin and Yutin 2018). There are mainly two opposing ideas on the origin of giant viruses. Although the first one suggests the emergence of giant viruses from a complex cellular ancestor by shedding genes (Nasir et al. 2012a; Nasir and Caetano-Anollés 2015), the second one proposes a simple ancestor with complexity originating by acquisition and accumulation of genes (Moreira and Brochier-Armanet 2008; Koonin, Krupovic, et al. 2015).

Rather than analyzing only one or two components, in this study, we focused on the entire DNA replication and analyzed the phylogeny, coevolution, and evolutionary selection pressure on the core components of the mimiviral replication machinery. Mimiviruses encode a conserved group of at least 21 proteins with putative functional roles in genome replication. In addition to the core replication proteins, namely, DNA polymerase (gp351), origin binding protein (gp1), primase-helicase bifunctional proteins (gp229), SSBP (gp544), sliding clamps (gp886, gp532, gp124), clamp loaders (gp549, gp538, gp513, gp425, gp441), ligase (gp331), RNase H (gp326, gp371), and topoisomerases (gp243, gp515, gp216), Mimivirus encodes several primases (gp577, gp857), helicases (gp612, gp132, and gp635), and bifunctional primase-helicase (gp8) proteins. Although their role in genome replication is not clear, it was recently shown that some of them, such as gp577, gp857, and gp8, do not appear to be involved in DNA replication (Gupta et al. 2019).

Notwithstanding the diversity of NCLDVs, a phylogenetic analysis of their DNA polymerases supports a strong cladistic relationship suggesting a common ancestor (Claverie and Abergel 2013) (supplementary fig. S2, Supplementary Material online). The DNA polymerases of eukaryotes, archaea, Mimivirus, and other giant viruses belong to the B-family, and studies based on the phylogenetic analysis of the B-family polymerases have suggested the possible giant viral origin of the eukaryotic polymerases (Takemura 2001; Takemura et al. 2015). Furthermore, these cues, when taken together with the completely cytoplasmic lifecycle of NCLDVs, points toward their involvement in the emergence of eukaryotes (Claverie 2006; Takemura et al. 2015). When several proteins are intricately involved in a particular function like DNA replication that requires them to interact physically, it puts a constraint on natural selection, and components involved in the process exhibit correlated pattern of changes to maintain the function (Goh et al. 2000; Tillier and Charlebois 2009).

Sliding clamp and clamp loaders are the essential components of complex DNA replication machinery ensuring a high processivity required for efficient DNA replication (Kelch et al. 2012; Hedglin et al. 2013). Sliding clamps form a ring-shaped structure around template DNA-primer junctions, associate with DNA polymerases tethering, and orienting them for DNA synthesis. They prevent the dissociation of polymerases during DNA synthesis. Clamp loaders are multimeric ATPases that load sliding clamps onto DNA. Despite their structural differences, clamp loaders from all domains of life follow similar mechanisms for loading the sliding clamps onto DNA (Kelch et al. 2012). Of all the replication-related proteins, only clamp and clamp loaders are conserved across all three domains of life, and hence they are thought to be present in LUCA before their divergence (Yao and O'Donnell 2016). Furthermore, the structural studies of the sliding clamp and clamp loaders have revealed that their architecture and composition are homologous in cellular forms and phage T4 (Jeruzalmi et al. 2002; Hedglin et al. 2013). The sliding clamp of bacteria is the β subunit of DNA polymerase III, in phage T4, it is gp45, and that of eukaryotes is PCNA. Archaea generally encode a single PCNA, whereas crenarchaea carry three different types of clamps (PCNA1-3). Bacteria have γ complex of $\delta.\gamma.\tau 2.\delta'.\psi.\chi$ subunits; phage T4 encodes a gp44–gp62 complex, archaea carry an RFCL/RFCS complex whereas eukaryotes have the RFC1-5 complex as clamp loaders. During our study, we found that Mimivirus carries genes coding three different types of PCNAs (gp886, gp532, and gp124) and five RFCs (a large subunit, gp441; and four small subunits gp549, gp513, gp425, and gp538). Our phylogenetic analysis showed that the mimiviral clamp and clamp loaders share a clade with eukaryotic PCNA and RFCs. All NCLDVs do not encode all PCNA and clamp loaders (supplementary table S17, Supplementary Material online). A few *Mimiviridae* family viruses encode all 3 PCNA, whereas other viruses encode 1 to 2 clamps, but most of *Mimiviridae* carry all 5 RFCs.

In summary, phylogenetic analysis showed that out of the nine core replication-related components considered in this study, five are of eukaryotic lineage (DNA polymerase, sliding clamp, clamp loaders 1 and 2, topoisomerase II), two are of bacterial (ligase, and topoisomerase I), one of phage (SSBP) origin, and one distinct protein (D5-like helicase, fig. 1). Although mimiviral replication machinery is a mosaic of

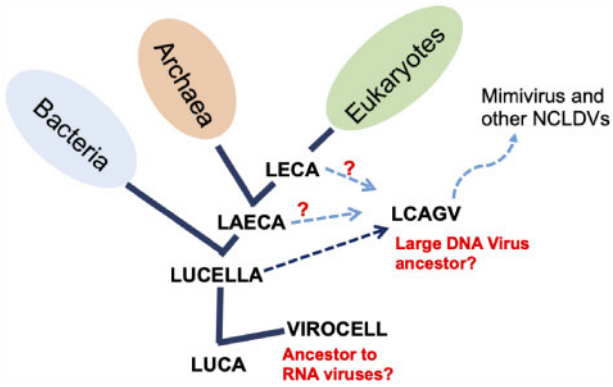

FIG. 7. A speculative hypothesis based on the ancestral nature of replication machinery of Mimivirus suggests the origin and evolution of giant viruses are from a descendent of LUCELLA; Last Common Ancestor of Giant Viruses (LCAGV) with a double-stranded DNA genome carrying complex DNA replication machinery. LUCA, last universal common ancestor; LUCELLA, last universal cellular ancestor; LAECA, last archaeo-eukaryotic common ancestor; LECA, last eukaryotic common ancestor.

proteins with varying phylogenetic affinities, our analysis suggests that most of them are coevolving and are part of the same functional complex. Furthermore, DNA compositional analysis suggested that the data set is homogeneous, and no evidence of a recent HGT in the mimiviral replication machinery was detected (fig. 6 and table 2), whereas the phylogenetic analysis (fig. 1) showed maximum homology with eukaryotic proteins. In addition, all the genes considered were found to be under strong purifying selection (table 2 and supplementary table S16, Supplementary Material online). Taking together, we infer these results as a reflection of the antiquity of the replication machinery of mimiviruses and propose that mimiviruses might have evolved from a complex cellular ancestor primarily by reductive evolution.

The three prevailing hypotheses on the origin of viruses are the virus-first, escape from host cells, and reduction from cells (Nasir et al. 2012b). A thorough analysis of the fold families and fold superfamilies (FFs and FSFs) from several thousand viral and cellular proteomes, including giant viruses, supported the latter hypothesis (Nasir and Caetano-Anollés 2015). Accordingly, modern viruses are the reduced forms that originated from multiple ancient cellular ancestor lineages. This hypothesis identifies two direct descendants of the LUCA, namely, the last universal cellular ancestor (LUCELLA) and the archaic virocell ancestor. Although the LUCA is thought to be the ancestor of cells that evolved protein synthetic machinery, the virocell ancestor took the parasitic route to become the modern-day viruses (Nasir et al. 2012b). This thought process is supported by the virocell concept proposed by Forterre (Forterre 2011) to underscore the importance of the intracellular stage of the virus lifecycle, and the notion is gaining traction among researchers. According to this hypothesis, virocells (or ribovirocells) are a subpopulation of "normal" cells that produce virions (Forterre 2012).

The presence of eukaryotic-like replication components with homologous sliding clamps and clamp loaders suggests that these viruses share common roots with the ancestor of eukaryotes which might be LUCELLA, last archaeo-eukaryotic common ancestor (LAECA), or last eukaryotic common ancestor (LECA) (fig. 7). The genome analysis of Mimivirus has shown that the gene products involved in transcription, translation, protein modifications, amino acid, and lipid metabolism are related to eukaryotic homologs; and, those involved in nucleotide synthesis and polysaccharide metabolism are close to bacterial homologs whereas proteins involved in the DNA repair share homology either with eukaryotes or bacteria (Suzan-Monti et al. 2006). This chimeric nature of the mimiviral genome suggests that the viral ancestor might be a cellular ancestor sharing genes with three domains of life possibly a descendent of LUCELLA. Furthermore, when we extended our study to putative components of DNA repair machinery, genome packaging, signal transduction pathways, biosynthesis of nucleotides, transcription, and the remnants of the translation machinery, we found them to be coevolving (fig. 5A and B; supplementary table S13, Supplementary Material online). In addition, they were found to be located centrally in the genome, which appears to be a characteristic feature of core genes (Shukla et al. 2018) supporting the hypothesis.

As documented, there are several instances of HGT, both "replacing" and "additive" that have most likely played important roles in the evolution of mimiviruses and other giant viruses (Shackelton and Holmes 2004; Koonin, Krupovic, et al. 2015; Koonin and Yutin 2018). However, HGT alone does not explain how the degree of complexity in the DNA replication machinery of mimiviruses arose. The propensity of a gene to undergo horizontal transfer has been greatly debated in the last two decades (Garcia-Vallve et al. 2000; Ochman et al. 2000; Soucy et al. 2015). Although most genes are amenable to HGT, the success of such transfers and their retention in a new environment is highly variable. The complexity hypothesis proposed about two decades ago posited that the frequency of a successful transfer depends on the functional context of the gene (Jain et al. 1999). Thus, it was shown that informational genes such as the ones involved in complex processes like transcription and translation are generally less transferable compared with the operational or metabolic genes. It was also shown that the extent of protein–protein interactions is a major determinant of transferability. This conclusion was drawn from studying over 300 sets of orthologous genes from six prokaryotic genomes. A later study further extended this hypothesis and demonstrated that higher connectivity of proteins is a major deterrent to HGT (Cohen et al. 2011). Consistent with this hypothesis, our studies showed that mimiviral replication proteins are highly connected and no extensive HGT was observed. These insights further imply that the complexity of mimiviral replication machinery is a vertically transferred feature and could be traced back to a common ancestor. It is also consistent with the complete autonomy of mimiviral DNA replication

in the host cytoplasm. To a larger extent, the DNA replication machinery of mimiviruses has retained complexity, which is consistent with the complete autonomy of their DNA replication in the host cytoplasm. Reductive evolution has largely occurred with translation and metabolism-related genes. Although the evidence presented here does not rule out the accretion hypothesis, taking the evidence together, a parsimonious explanation would be that mimiviruses have radiated from a complex cellular ancestor, probably a reasonably evolved descendant of LUCELLA, to adapt to a parasitic lifestyle. Based on our findings, we propose that the mimiviruses and large DNA viruses might have evolved from a Last Common Ancestor of Giant Viruses (LCAGV), a descendant of LUCELLA (fig. 7). Isolation of more giant viruses will hopefully help in filling the many gaps in our understanding of the origins of giant viruses.

Materials and Methods

Sequence Analysis

Proteins involved in DNA replication of Mimivirus were selected from UniProt (Consortium 2006), National Centre for Biotechnology Information (NCBI), and the transcriptomics (Raoult et al. 2004; Legendre et al. 2010), and proteomics data (Fridmann-Sirkis et al. 2016). All DNA replication proteins of Mimivirus were identified for their putative function by sequence similarity with the characterized proteins and by domain prediction tools, namely, InterProScan (Zdobnov and Apweiler 2001), and the NCBI Conserved Domain Database, CDD (Marchler-Bauer et al. 2004). Protein BLAST was performed for individual proteins of Mimivirus against NCBI nonredundant database, and sequences from the *Mimiviridae* family were selected for analysis (table 1).

DNA replication proteins of bacteriophage T4, bacteria *E. coli*, archaeon *Aeropyrum pernix*, and the eukaryote *Homo sapiens* selected are shown in table 1. Most of these proteins have been biochemically characterized for their functions and many of the interactions have also been experimentally established except for the archaeal DNA replication machinery where only a few proteins have been characterized (Thömmes and Hübscher 1990; Nossal 1992; Kelman and O'Donnell 1994). All sequences of the replication systems were retrieved from the UniProt, and the NCBI database and sequences for analyses were retrieved by iterative BLAST.

Phylogenetic Analyses

We selected functionally related homologs of replicative DNA polymerase, replicative helicase, sliding clamp, clamp loader 1 and 2, topoisomerases I and II, single-strand binding proteins (SSBP), and ligase from the respective bacterial, archaeal, eukaryotic, phage, and *Mimiviridae* systems for phylogenetic analysis (supplementary table S1, Supplementary Material online). We constructed phylogenetic trees using MEGA6.0 (Tamura et al. 2013) for NJ, ME and UPGMA methods and FastTree v.2.1.(Price et al. 2010) was used for ther ML method with default parameters.

Protein Level Coevolution Analysis

The correlation analysis was performed by the linear regression method (Goh et al. 2000; Gupta et al. 2017) using XLSTAT. For calculation of the Pearson correlation coefficient (r), the distance matrices of each protein were used from the same set of *Mimiviridae* family viruses, bacteriophages, bacteria, archaea, and eukaryotes (supplementary table S2, Supplementary Material online). The multiple sequence alignment (MSA) was generated by MUSCLE and the distance matrix for the individual protein was computed as pairwise distances from the aligned sequences using default parameters by MEGA6.0 (Goh et al. 2000; Gupta et al. 2017; Tamura et al. 2013). The r between protein pairs were calculated from distance matrices using the Pearson correlation coefficient formula (Efron 1979) for all proteins with each other in XLSTAT. A correlation coefficient of >0.7 between proteins suggests their coevolution. The correlation or similarity matrix of proteins of each replication system was used to generate a heatmap using the Origin 2018b software.

Phylogenetic trees of mimiviral replication proteins were also generated to understand their cophylogenetic mirror patterns. The same sequence data set was used for both correlation and phylogenetic analyses. FastTree v.2.1 software (Price et al. 2010) with default parameters using the generalized time-reversible (GTR) substitution model for 1,000 bootstrap replicates was used to infer phylogeny.

Coevolving Protein Visualization

The correlation coefficients of proteins of different replication systems were represented in the 2D space by multidimensional scaling analysis (MDS). MDS provides a visual representation of the pattern of proximities among a set of correlation coefficients and helps the assessment of similarities and distances. Taken together with the biological information, MDS is useful in assessing the involvement of a component in the replication process. The metric or classical MDS was employed to the similarity matrix calculated by the Pearson correlation coefficient for all protein pairs. The classical MDS, based on the SMACOF (Scaling by MAjoring a COmplicated Function) algorithm (de Leeuw 1988), was performed using the XLSTAT software. The similarity matrix was first transformed into a dissimilarity matrix and the absolute model of classical MDS was used where the disparity is equal to the dissimilarity matrix. For the analysis, we represented the data in two dimensions using ten repetitions and 1,000 iterations in the XLSTAT. The goodness of fit is based on the difference between the actual distance and their predicted values which is measured by Kruskal's stress-1 (Kruskal and Wish 1978). Here, for a given number of dimensions ($p = 2$), the weaker value represents the better quality of the representation. The perfect representation is when stress is 0 and fair when the value is 0.1.

Horizontal Gene Transfer Detection

Here, we have measured the GC content and codon usage of Mimivirus replication genes in the context of the entire genome. Codon bias was studied by measuring the Codon Adaptation Index (CAI), the Effective Codon Number (Nc),

and the third codon position GC content (GC3s) using CodonW (http://codonw.sourceforge.net, last accessed January 23, 2021; Peden 1999) with default parameters. CAI value ranges between 0 and 1, a stronger bias of synonymous codon usage is shown by a higher value (Sharp and Li 1987; Lawrence and Ochman 1997). Nc value varies from 20 to 61, and the extreme biasness is represented by 20 when only one codon for one amino acid is used whereas there is no bias in codon usage when the value is 61 (Wright 1990). The gene expression data of mimiviral replication genes were retrieved from http://www.igs.cnrs-mrs.fr/mimivirus/ (last accessed January 23, 2021) (Legendre et al. 2010). We conducted a similar analysis of phage T4, *E. coli*, *A. pernix*, and *Homo sapiens* replication genes. Statistical analysis was carried out by XLSTAT.

Evolutionary Selection Analysis

The evolutionary selection pressure exerted on the genes of the mimiviral replication system was estimated by measuring the d*N*/d*S* ratio. We estimated d*N*/d*S* ratios for mimiviral replication genes using EasyCodeML software, a wrapper of CodeML (Gao et al. 2019). Nucleotide sequences of mimiviral replication proteins used for correlation and MDS analysis were retrieved from NCBI. The MSA of each gene was carried out by MUSCLE and converted into PAML format whereas the respective phylogenetic tree was generated by ML using MEGA6.0 and saved in Newick format. We used the site-specific codon-based substitution model that allows the selection pressure to vary among sites (Yang and Nielsen 2002). We considered four pairs of codon substitution models, M0 (one ratio) versus M3 (discrete), M1a (Nearly Neutral) versus M2a (Positive Selection), M7 (β) versus M8 (β and ω), and M8 versus M8a (β and $\omega = 1$), and the fit of the models was compared using likelihood-ratio tests (LRTs) (Nielsen and Yang 1998; Yang et al. 2000). LRT was performed by calculating 2ΔLnL by taking twice the difference of log-likelihood (lnL) between two models and testing χ^2 distribution for significance (*P* value) with the degree of freedom (df) by taking the difference of the number of parameters (np) between models.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

K.K. conceptualized the study. K.K. and S.P. designed the analysis. S.P. conducted the analysis. K.K. and S.P. analyzed the data and wrote the article. K.K. acknowledges the support of Department of Science and Technology, DST (EMR/2016/005155) and S.P. acknowledges the support of IIT Bombay Research Fellowship. Authors thank Shailesh Lad for useful comments on the article.

Data Availability

The data underlying this article are available in the article and in its Supplementary Material online.

References

Atanassova N, Grainge I. 2008. Biochemical characterization of the mini-chromosome maintenance (MCM) protein of the crenarchaeote *Aeropyrum pernix* and its interactions with the origin recognition complex (ORC) proteins. *Biochemistry* 47(50):13362–13370.

Benarroch D, Claverie JM, Raoult D, Shuman S. 2006. Characterization of mimivirus DNA topoisomerase IB suggests horizontal gene transfer between eukaryal viruses and bacteria. *J Virol*. 80(1):314–321.

Benarroch D, Shuman S. 2006. Characterization of mimivirus NAD+-dependent DNA ligase. *Virology* 353(1):133–143.

Bhardwaj A, Ghose D, Thakur KG, Dutta D. 2018. *Escherichia coli* β-clamp slows down DNA polymerase I dependent nick translation while accelerating ligation. *PLoS One* 13(6):e0199559.

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* 277(5331):1453–1462.

Boyer M, Madoui MA, 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.

Cann IK, Ishino S, Nomura N, Sako Y, Ishino Y. 1999. Two family B DNA polymerases from Aeropyrum pernix, an aerobic hyperthermophilic crenarchaeote. *J Bacteriol*. 181(19):5984–5992.

Champoux JJ. 2001. DNA topoisomerases: structure, function, and mechanism. *Annu Rev Biochem*. 70(1):369–413.

Claverie JM. 2006. Viruses take center stage in cellular evolution. *Genome Biol*. 7(6):110–115.

Claverie JM, Abergel C. 2013. Chapter two – open questions about giant viruses. *Adv Virus Res*. 85:25–56.

Cohen O, Gophna U, Pupko T. 2011. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol*. 28(4):1481–1489.

Consortium TU. 2006. The Universal Protein Resource (UniProt). *Nucleic Acids Res*. 35:D193–D197.

Daimon K, Kawarabayasi Y, Kikuchi H, Sako Y, Ishino Y. 2002. Three proliferating cell nuclear antigen-like proteins found in the hyperthermophilic archaeon *Aeropyrum pernix*: interactions with the two DNA polymerases. *J Bacteriol*. 184(3):687–694.

de Leeuw J. 1988. Convergence of the majorization method for multidimensional scaling. *J Classif*. 5(2):163–180.

ΘDrake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of spontaneous mutation. *Genetics* 148(4):1667–1686.

Efron B. 1979. Computers and the theory of statistics: thinking the unthinkable. *SIAM Rev*. 21(4):460–480.

Forterre P. 2002. The origin of DNA genomes and DNA replication proteins. *Curr Opin Microbiol*. 5(5):525–532.

Forterre P. 2006. Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. *Proc Natl Acad Sci U S A*. 103(10):3669–3674.

Forterre P. 2011. Manipulation of cellular syntheses and the nature of viruses: the virocell concept. *Comptes Rendus Chim*. 14(4):392–399.

Forterre P. 2012. The Virocell concept. The encyclopedia of life science. Chister: John Wiley & Sons, Ltd.

Forterre P, Gadelle D. 2009. Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. *Nucleic Acids Res*. 37(3):679–692.

Forterre P, Prangishvili D. 2013. The major role of viruses in cellular evolution: facts and hypotheses. *Curr Opin Virol*. 3(5):558–565.

Fridmann-Sirkis Y, Milrot E, Mutsafi Y, Ben-Dor S, Levin Y, Savidor A, Kartvelishvily E, Minsky A. 2016. Efficiency in complexity: composition and dynamic nature of mimivirus replication factories. *J Virol*. 90(21):10039–10047.

Frigola J, He J, Kinkelin K, Pye VE, Renault L, Douglas ME, Remus D, Cherepanov P, Costa A, Diffley JFX. 2017. Cdt1 stabilizes an open MCM ring for helicase loading. *Nat Commun*. 8:15720.

Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW. 2019. EasyCodeML: a visual tool for analysis of selection using CodeML. *Ecol Evol*. 9(7):3891–3898.
Garcia-Vallve S, Romeu A, Palau J. 2000. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res*. 10(11):1719–1725.
Gelderblom HR. 1996. Structure and classification of viruses. In: Baron S, editor. 4th ed. Medical microbiology. Galveston (TX): University of Texas Medical Branch at Galveston.
Goh C, Bogan AA, Joachimiak M, Walther D, Cohen FE. 2000. Co-evolution of proteins with their interaction partners. *J Mol Biol*. 299(2):283–293.
Guglielmini J, Woo AC, Krupovic M, Forterre P, Gaia M. 2019. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc Natl Acad Sci U S A*. 116(39):19585–19592.
Gupta A, Lad SB, Ghodke PP, Pradeepkumar PI, Kondabagil K. 2019. Mimivirus encodes a multifunctional primase with DNA/RNA polymerase, terminal transferase and translesion synthesis activities. *Nucleic Acids Res*. 47(13):6932–6945.
Gupta A, Patil S, Vijayakumar R, Kondabagil K. 2017. The polyphyletic origins of primase – helicase bifunctional proteins. *J Mol Evol*. 85(5–6):188–204.
Hedglin M, Kumar R, Benkovic SJ. 2013. Replication clamps and clamp loaders. *Cold Spring Harb Perspect Biol*. 5:1–19.
Imamura K, Fukunaga K, Kawarabayasi Y, Ishino Y. 2007. Specific interactions of three proliferating cell nuclear antigens with replication-related proteins in *Aeropyrum pernix*. *Mol Microbiol*. 64(2):308–318.
Iyer LM, Koonin EV, Leipe DD, Aravind L. 2005. Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res*. 33(12):3875–3896.
Jain R, Rivera MC, Lake JA. 1999. Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. 96(7):3801–3806.
Jeruzalmi D, O'Donnell M, Kuriyan J. 2002. Clamp loaders and sliding clamps. *Curr Opin Struct Biol*. 12(2):217–224.
Karam JD, Konigsberg WH. 2000. DNA polymerase of the T4-related bacteriophages. *Prog Nucleic Acid Res Mol Biol*. 64:65–96.
Kazlauskas D, Krupovic M, Venclovas C. 2016. The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Res*. 44(10):4551–4564.
Kazlauskas D, Venclovas C. 2012. Two distinct SSB protein families in nucleo-cytoplasmic large DNA viruses. *Bioinformatics* 28(24):3186–3190.
Kelch BA, Makino DL, O'Donnell M, Kuriyan J. 2012. Clamp loader ATPases and the evolution of DNA replication machinery. *BMC Biol*. 10(1):14.
Kelman Z, O'Donnell M. 1994. DNA replication: enzymology and mechanisms. *Curr Opin Genet Dev*. 4(2):185–195.
Koonin EV. 1993. A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res*. 21(11):2541–2547.
Koonin EV. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*. 1(2):127–136.
Koonin EV, Dolja VV, Krupovic M. 2015. Origins and evolution of viruses of eukaryotes: the ultimate modularity. *Virology* 479–480:2–25.
Koonin EV, Krupovic M, Yutin N. 2015. Evolution of double-stranded DNA viruses of eukaryotes: from bacteriophages to transposons to giant viruses. *Ann NY Acad Sci*. 1341(1):10–24.
Koonin EV, Senkevich TG, Dolja VV. 2006. The ancient Virus World and evolution of cells. *Biol Direct*. 1(1):29.
Koonin EV, Yutin N. 2010. Origin and evolution of eukaryotic large nucleo-cytoplasmic DNA viruses. *Intervirology* 53(5):284–292.
Koonin EV, Yutin N. 2018. Multiple evolutionary origins of giant viruses. *F1000Res*. 7:1840–1812.
Kornberg A, Baker TA. 1992. DNA replication. 2nd ed. New York: W. H. Freeman and Company.
Kruskal JB, Wish M. 1978. Multidimensional scaling. Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-011. Newbury Park: Sage Publications.
Labib K, Diffley JF. 2001. Is the MCM2-7 complex the eukaryotic DNA replication fork helicase? *Curr Opin Genet Dev*. 11(1):64–70.
Lang S, Huang L. 2015. The *Sulfolobus solfataricus* GINS complex stimulates DNA binding and processive DNA unwinding by minichromosome maintenance helicase. *J Bacteriol*. 197(21):3409–3420.
Lawrence JG, Ochman H. 1997. Amelioration of bacterial genomes: rates of change and exchange. *J Mol Evol*. 44(4):383–397.
Legendre M, Audic S, Poirot O, Hingamp P, Seltzer V, Byrne D, Lartigue A, Lescot M, Bernadac A, Poulain J, et al. 2010. mRNA deep sequencing reveals 75 new genes and a complex transcriptional landscape in Mimivirus. *Genome Res*. 20(5):664–674.
Leipe DD, Aravind L, Koonin EV. 1999. Did DNA replication evolve twice independently? *Nucleic Acids Res*. 27(17):3389–3401.
Loh E, Salk JJ, Loeb LA. 2010. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc Natl Acad Sci U S A*. 107(3):1154–1159.
Lu YB, Ratnakar PV, Mohanty BK, Bastia D. 1996. Direct physical interaction between DnaG primase and DnaB helicase of *Escherichia coli* is necessary for optimal synthesis of primer RNA. *Proc Natl Acad Sci U S A*. 93(23):12902–12907.
Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet*. 17(11):704–714.
Maki H, Maki S, Kornberg A. 1988. DNA polymerase III holoenzyme of *Escherichia coli*. IV. The holoenzyme is an asymmetric dimer with twin active sites. *J Biol Chem*. 263(14):6570–6578.
Marceau AH, Bahng S, Massoni SC, George NP, Sandler SJ, Marians KJ, Keck JL. 2011. Structure of the SSB-DNA polymerase III interface and its role in DNA replication. *EMBO J*. 30(20):4236–4247.
Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z. 2004. CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*. 33(Database issue):D192–D196.
Miller ES, Kutter E, Mosig G, Arisaka F, Kunisawa T, Rüger W. 2003. Bacteriophage T4 genome. *Microbiol Mol Biol Rev*. 67(1):86–156.
Moreira D, Brochier-Armanet C. 2008. Giant viruses, giant chimeras: the multiple evolutionary histories of Mimivirus genes. *BMC Evol Biol*. 8:12.
Nasir A, Caetano-Anollés G. 2015. A phylogenomic data-driven exploration of viral origins and evolution. *Sci Adv*. 1(8):e1500527.
Nasir A, Kim KM, Caetano-Anolles G. 2012a. Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC Evol Biol*. 12(1):156.
Nasir A, Kim KM, Caetano-Anollés G. 2012b. Viral evolution: primordial cellular origins and late adaptation to parasitism. *Mob Genet Elements*. 2(5):247–252.
ΘNielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148(3):929–936.
Nossal NG. 1992. Protein–protein interactions at a DNA replication fork: bacteriophage T4 as a model. *FASEB J*. 6(3):871–878.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.
Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein Eng Des Sel*. 14(9):609–614.
Peden JF. 1999. Analysis of codon usage [thesis]. Nottingham, United Kingdom: Department of Genetics, University of Nottingham.
Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3):e9490.
Raoult D. 2013. TRUC or the need for a new microbial classification. *Intervirology* 56(6):349–353.
Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, La Scola B, Suzan M, Claverie JM. 2004. The 1.2-megabase genome sequence of Mimivirus. *Science* 306(5700):1344–1350.

Sarmiento F, Long F, Cann I, Whitman WB. 2014. Diversity of the DNA replication system in the archaea domain. *Archaea* 2014:1–15.
Schaeffer PM, Headlam MJ, Dixon NE. 2005. Protein–protein interactions in the eubacterial replisome. *IUBMB Life*. 57(1):5–12.
Shackelton LA, Holmes EC. 2004. The evolution of large DNA viruses: combining genomic information of viruses and their hosts. *Trends Microbiol*. 12(10):458–465.
Sharp PM, Li WH. 1987. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 15(3):1281–1295.
Shukla A, Chatterjee A, Kondabagil K. 2018. The number of genes encoding repeat domain-containing proteins positively correlates with genome size in amoebal giant viruses. *Virus Evol*. 4:1–11.
Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat Rev Genet*. 16(8):472–482.
Suzan-Monti M, La Scola B, Raoult D. 2006. Genomic and evolutionary aspects of Mimivirus. *Virus Res*. 117(1):145–155.
Takemura M. 2001. Poxviruses and the origin of the eukaryotic nucleus. *J Mol Evol*. 52(5):419–425.
Takemura M, Yokobori S, Ogata H. 2015. Evolution of eukaryotic DNA polymerases via interaction between cells and large DNA viruses. *J Mol Evol*. 81(1–2):24–33.
Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol*. 30(12):2725–2729.
Thömmes P, Hübscher U. 1990. Eukaryotic DNA replication: enzymes and proteins acting at the fork. *Eur J Biochem*. 194(3):699–712.
Tillier ERM, Charlebois RL. 2009. The human protein coevolution network. *Genome Res*. 19(10):1861–1871.
Villarreal LP, DeFilippis VR. 2000. A hypothesis for DNA viruses as the origin of eukaryotic replication proteins. *J Virol*. 74(15):7079–7084.
Wickner S, Hurwitz J. 1975. Interaction of *Escherichia coli* DnaB and DnaC(D) gene products in vitro. *Proc Natl Acad Sci U S A*. 72(3):921–925.
Wright F. 1990. The "effective number of codons" used in a gene. *Gene* 87(1):23–29.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19(6):908–917.
Yang Z, Nielsen R, Goldman N, Pedersen AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
Yao NY, O'Donnell ME. 2016. Evolution of replication machines. *Crit Rev Biochem Mol Biol*. 51(3):135–149.
Yin C, Yau SST. 2017. A coevolution analysis for identifying proteinprotein interactions by Fourier transform. *PLoS One* 12(4):e0174862–e0174919.
Yutin N, Wolf YI, Raoult D, Koonin EV. 2009. Eukaryotic large nucleocytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. *Virol J*. 6(1):223.
Zannis-Hadjopoulos M, Sibani S, Price GB. 2004. Eucaryotic replication origin binding proteins. *Front Biosci*. 9(1–3):2133–2143.
Zdobnov EM, Apweiler R. 2001. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848.