



**HAL**  
open science

## Viral origin of eukaryotic type IIA DNA topoisomerases

Julien Guglielmini, Morgan Gaia, Violette da Cunha, Mart Krupovic, Patrick Forterre

► **To cite this version:**

Julien Guglielmini, Morgan Gaia, Violette da Cunha, Mart Krupovic, Patrick Forterre. Viral origin of eukaryotic type IIA DNA topoisomerases. *Virus Evolution*, 2022, 8 (2), pp.veac097. 10.1093/ve/veac097. pasteur-03841210

**HAL Id: pasteur-03841210**

**<https://pasteur.hal.science/pasteur-03841210>**

Submitted on 6 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# Viral origin of eukaryotic type IIA DNA topoisomerases

Julien Guglielmini,<sup>1,†</sup> Morgan Gaia,<sup>2</sup> Violette Da Cunha,<sup>3</sup> Alexis Criscuolo,<sup>1</sup> Mart Krupovic,<sup>4,‡</sup> and Patrick Forterre<sup>3,4,§</sup>

<sup>1</sup>Institut Pasteur, Université de Paris, Bioinformatics and Biostatistics Hub, F-75015 Paris, France, <sup>2</sup>Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, 91000 Evry, France, <sup>3</sup>Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198 Gif-sur-Yvette, France and <sup>4</sup>Institut Pasteur, Université Paris Cité, CNRS UMR 6047, Archaeal Virology Unit, 75015 Paris, France

<sup>†</sup><https://orcid.org/0000-0002-8566-1726>

<sup>‡</sup><https://orcid.org/0000-0001-5486-0098>

<sup>§</sup><https://orcid.org/0000-0002-1843-7214>

\*Corresponding author: E-mail: [patrick.forterre@pasteur.fr](mailto:patrick.forterre@pasteur.fr)

## Abstract

Type II DNA topoisomerases of the family A (Topo IIAs) are present in all Bacteria (DNA gyrase) and eukaryotes. In eukaryotes, they play a major role in transcription, DNA replication, chromosome segregation, and modulation of chromosome architecture. The origin of eukaryotic Topo IIA remains mysterious since they are very divergent from their bacterial homologs and have no orthologs in Archaea. Interestingly, eukaryotic Topo IIAs have close homologs in viruses of the phylum *Nucleocyotiviricota*, an expansive assemblage of large and giant viruses formerly known as the nucleocytoplasmic large DNA viruses. Topo IIAs are also encoded by some bacteriophages of the class *Caudoviricetes* (tailed bacteriophages). To elucidate the origin of the eukaryotic Topo IIA, we performed in-depth phylogenetic analyses on a dataset combining viral and cellular Topo IIA homologs. Topo IIAs encoded by Bacteria and eukaryotes form two monophyletic groups nested within Topo IIA encoded by *Caudoviricetes* and *Nucleocyotiviricota*, respectively. Importantly, *Nucleocyotiviricota* remained well separated from eukaryotes after removing both Bacteria and *Caudoviricetes* from the data set, indicating that the separation of *Nucleocyotiviricota* and eukaryotes is probably not due to long-branch attraction artifact. The topologies of our trees suggest that the eukaryotic Topo IIA was probably acquired from an ancestral member of the *Nucleocyotiviricota* of the class *Megaviricetes*, before the emergence of the last eukaryotic common ancestor (LECA). This result further highlights a key role of these viruses in eukaryogenesis and suggests that early proto-eukaryotes used a Topo IIB instead of a Topo IIA for solving their DNA topological problems.

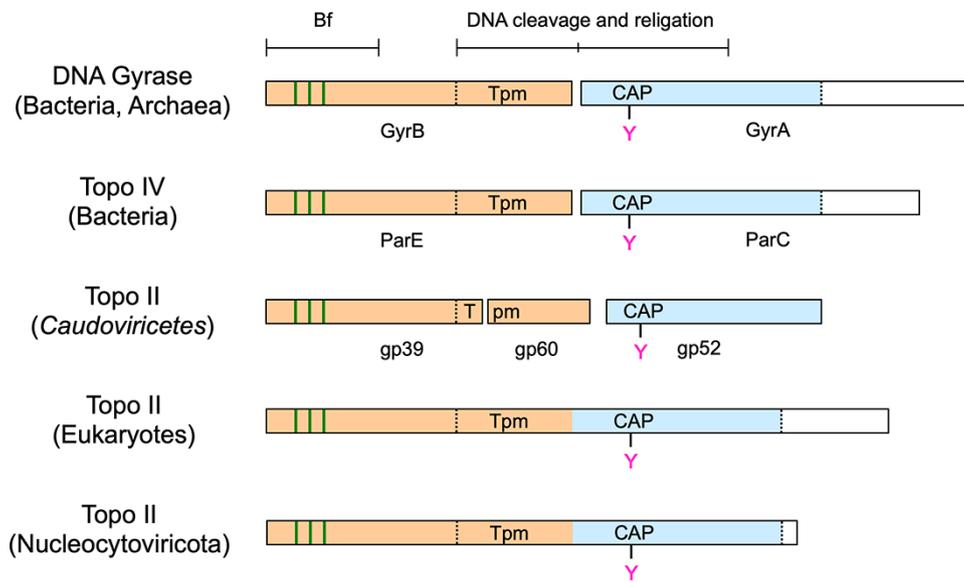
**Key words:** DNA topoisomerases; virus evolution; NCLDV; *Nucleocyotiviricota*; *Duplodnavira*; eukaryogenesis.

## Introduction

DNA topoisomerases are ubiquitous enzymes that are essential for solving topological problems inherent to the helical structure of DNA (Champoux 2001; Forterre and Gabelle 2009; Wang 2009; Vos et al. 2011; Forterre 2012; Pommier et al. 2016). Based on mechanistic properties, they have been classified into types I and II. Type I DNA topoisomerases (Topo I) produce transient single-strand breaks in double-stranded DNA (dsDNA) and catalyze the transfer of one DNA strand through this break. In contrast, type II DNA topoisomerases (Topo II) produce transient double-strand breaks and catalyze the transfer of a dsDNA segment (from either the same or different dsDNA molecule) through this break. Five different families of DNA topoisomerases have been defined based on amino-acid sequences and structural similarities: three families of Topo I (Topo IA, Topo IB, and Topo IC) (Champoux 2001; Forterre 2006) and two families of Topo II (Topo IIA and Topo IIB) (Bergerat et al. 1997; Gabelle et al. 2003). All Topo II and some Topo I (IB and IC) can relax positive and negative superturns that otherwise would accumulate in front and behind the replication forks and transcription bubbles, respectively. In addition, Topo II can eliminate the catenanes that can accumulate at the

end of the chromosome replication. In eukaryotes, Topo IIAs are also intrinsic structural components of the chromosomal scaffold (Hizume et al. 2007) and play a major role in modulating chromosome architecture and long-range chromatin structure (Nitiss 2009; Nielsen et al. 2020).

DNA topoisomerases have been extensively studied because they are the targets of important antibiotics and antitumor drugs (Pommier 2013). These drugs interfere with the breakage-reunion mechanisms of the enzyme and transform the transient intermediate topoisomerase–DNA covalent complexes into stable poisons, interfering with replication and transcription. However, these enzymes are also very interesting (and intriguing) in terms of the history of life on our planet. Indeed, the distribution patterns of the different DNA topoisomerase families within the three domains of life, Archaea, Bacteria, and Eukarya (eukaryotes), do not fit the usual distribution of informational proteins, such as ribosomal proteins or DNA-dependent RNA polymerases (Da Cunha et al. 2017), raising challenging questions and prompting unorthodox hypotheses (Forterre and Gabelle 2009). Whereas informational proteins from eukaryotes usually much more closely resemble their archaeal homologs than their



**Figure 1.** Schematic representation of the domain composition of Topo IIAs. The different domains correspond to the Bergerat fold/GHKL (Bf), TOPRIM (Tpm), and the 5Y-CAP or winged helix (WHD) domain containing the catalytic Tyrosine.

bacterial ones, the universal eukaryotic Topo II (member of the Topo IIA family) has no obvious ortholog in Archaea. All Archaea (except for certain Thermoplasmatales) contain an enzyme of the Topo IIB family, dubbed DNA topoisomerase VI, suggesting that the last archaeal common ancestor contained no Topo IIA but a Topo IIB for relaxation of positive superturns and chromosome decatenation (Forterre and Gabelle 2009). All bacteria encode a unique Topo IIA, DNA gyrase, which is a distant homolog of the eukaryotic enzyme. DNA gyrases are heterotetramers composed of two subunits (GyrA and GyrB) that are homologous to the C- and N-terminal moieties of the homodimeric Topo IIA of eukaryotes, respectively (Fig. 1). Some Archaea encode a two-subunit DNA gyrase very similar to the bacterial enzyme and highly divergent from the eukaryotic Topo IIA. Phylogenetic analysis has indicated that these DNA gyrases were recruited from Bacteria by lateral gene transfer (Forterre et al. 2014). Similarly, some eukaryotes, such as *Archaeplastida*, encode a bacterial-like DNA gyrase (Topo IIA) present in chloroplasts and mitochondria that was most likely acquired from *Cyanobacteria* during the endosymbiotic event that led to the emergence of the chloroplasts (Wall, Mitchenall, and Maxwell 2004). These eukaryotic Topo II are very similar to their bacterial counterparts and in phylogenetic analyses are nested within the clade of bacterial gyrases (Forterre et al. 2007). It seems unlikely that the very divergent eukaryotic Topo IIA originated through a similar endosymbiotic pathway. The origin of the Topo IIA in eukaryotes thus remains enigmatic.

A possible answer to this enigma could reside in the virosphere. The first viral Topo IIA was discovered in 1980 in the T4 bacteriophage (Liu, Liu, and Alberts 1979), the iconic virus of the class *Caudoviricetes* from the recently proposed family *Straboviridae*. Surprisingly, the T4 Topo IIA, a heterotrimer (Fig. 1), is not specifically related to bacterial Topo IIA but branched between DNA gyrases and eukaryotic Topo IIA in phylogenetic trees (Gabelle et al. 2003). Later on, Topo IIA genes were discovered in several members of the phylum *Nucleocytoviricota* (Fig. 1), formerly known as nucleocytoplasmic large DNA viruses (NCLDV) (Lavrukhin et al. 2000; Gabelle et al. 2003; Coelho et al. 2015, 2016; Erives 2017). Inhibition of this Topo IIA disrupts replication of the African swine fever virus (ASFV; family *Asfarviridae*) *in vitro* (Freitas et al. 2016),

indicating that compounds active against the ASFV-Topo IIA, such as fluoroquinolones, are promising drugs against the highly contagious and fatal disease caused in pigs by ASFV.

The Topo IIAs encoded by *Nucleocytoviricota* are very similar to the ubiquitous eukaryotic Topo IIA at the sequence level, and in that they are homodimers devoid of gyrase activity (Fig. 1). In the traditional view that considers viruses as pickpockets of cellular proteins, this suggests that Topo IIAs were acquired by *Nucleocytoviricota* from their eukaryotic hosts. However, in the framework of the 'out of viruses' hypothesis for the origin of DNA (Forterre 2002), it is tempting to suggest that this gene transfer took place the other way around and that eukaryotic Topo IIA was acquired from the *Nucleocytoviricota* (Forterre and Gabelle 2009). A preliminary phylogenetic analysis provided ambiguous results: some Topo IIAs from *Nucleocytoviricota* branched between T4 and Eukarya, suggesting that Topo IIA was indeed transferred from viruses to cells, whereas other viral enzymes branched within eukaryotes in agreement with transfers from cells to viruses (Gabelle et al. 2003; Forterre et al. 2007).

At the time of these analyses, only six Topo IIAs from four families (*Asfarviridae*, *Mimiviridae*, *Iridoviridae*, and *Phycodnaviridae*) within *Nucleocytoviricota* were known (Forterre et al. 2007). During the last decade, a great number of new *Nucleocytoviricota* genomes became available, including those of giant viruses from the families *Mimiviridae*, *Marseilleviridae*, and *Pandoraviridae*, which encode Topo IIA (Abergel, Legendre, and Claverie 2015; Colson et al. 2017). Notably, it was shown that the Topo IIA encoded by *Marseilleviridae* branched as a sister clade to Eukarya (Erives 2017). We thus decided to update the Topo IIA phylogenetic classification, focusing on viral and eukaryotic Topo IIAs. Our results strongly suggest that eukaryotic Topo IIA originated from a Topo IIA ancestor encoded by a virus closely related to modern *Megaviricetes*, a class of *Nucleocytoviricota* that includes many giant viruses, such as *Mimiviridae*. We have previously reported phylogenetic analyses, suggesting that eukaryotic RNA polymerase II was probably recruited by eukaryotes from a virus related to *Imitervirales* in a tree including *Nucleocytoviricota* and the three nuclear RNA polymerases present in all eukaryotes (Guglielmini et al. 2019). One can speculate that both RNA polymerase II and Topo IIA were

possibly acquired simultaneously by a proto-eukaryote in the lineage leading to the LECA, in agreement with the fact that these two enzymes interact functionally and physically in modern eukaryotes. Regardless, our results support the hypothesis that interactions between proto-eukaryotes and *Nucleocytoviricota* have played an important role in shaping the physiology of modern eukaryotic cells.

## Material and methods

### Data collection

For Bacteria, we downloaded the full proteomes of a set of 112 bacterial strains spanning 10 representative groups (Aquificae, Dictyoglomi, Elusimicrobia, FCB group Plactomycetales, Verrucomicrobiales, Chlamydiae, Nitrospirae, PVC group Fibrobacterota, Chlorobiota, and Bacteroidota, Proteobacteria, Spirochaetes, Terrabacteria group, and Thermotogae). We used BLASTP v2.9.0+ (Ramsay et al. 2000; Camacho et al. 2009) recursively to collect the homologs of *Escherichia coli* K12 GyrB and GyrA proteins (WP\_000072067.1, NP\_416734.1) in those 112 proteomes. Finally, we concatenated each related pair of GyrB and GyrA hits. We also added the sequence of *E. coli* Topo IV (Supplementary Table S1) that branched at the base of DNA gyrase in previous analyses.

For *Caudoviricetes* (tailed phages), we downloaded all the 1,131,926 *Caudoviricetes* proteins and next used BLASTP to search for the homologs of the three subunits of the T4 phage topoisomerase II (E-value lower than  $1e-10$ ). We kept only *Caudoviricetes* lineages for which we obtained a hit for the three subunits and concatenated the corresponding sequences. Interestingly, besides the group of previously known Topo IIA closely related to T4 Topo IIA infecting Enterobacteriaceae, we detected several new related Topo IIA sequences in distantly related members of the *Myoviridae* infecting Rhizobiaceae and Firmicutes, but also in three members of the *Ackermannviridae* infecting Enterobacter and Rhizobiaceae as well as three unclassified siphoviruses infecting Firmicutes (Supplementary Table S1).

For *Nucleocytoviricota* (NCLDV), we searched for Topo IIA sequences in *Nucleocytoviricota* genomes that we previously used to determine the list of core genes of *Nucleocytoviricota* and the phylogeny based on the concatenation of eight core genes (Guglielmini et al. 2019) (Supplementary Table S1). To improve this data set, we added one sequence from *Acanthocystis turfacea* Chlorella virus and three sequences from Pithoviruses (Pithovirus LCPAC104, Pithovirus LCPAC201 and PithovirusLCPAC202). Finally, in order to improve our sampling in the *Asfuvirales* and *Pandoravirales* orders, we used our TopoIIA HMM profile against the Giant Virus Database (Aylward et al. 2021). We obtained seven hits in *Pandoravirales* and twenty in *Asfuvirales* that were not yet in our data set (three stars in Supplementary Table S1). A preliminary phylogenetic analysis led to an unusual placement for some TopoIIA proteins from metagenomic assemblies. This could represent true phylogenetic positions which could be explained by horizontal gene transfers or artifactual positions. The latter could, for example, be caused by contamination occurring during the binning. Because finding an explanation for these unorthodox phylogenetic placements is out of the scope of this work, we decided to remove these sequences (four from *Asfuvirales* and three from *Pandoravirales*).

Topo IIA turned out to be present in all *Imitervirales* (formerly *Megavirales* in Guglielmini et al. 2019), an order that includes *Mimiviridae* and several more recently described large DNA viruses (Catovirus, Hokovirus, Indivirus, Klosneuvirus, and Tupanvirus) (Guglielmini et al. 2019). Topo IIA is also present in some

members of the *Pimascovirales*, in particular, in all members of *Marseilleviridae* and viruses of the Pitho-like group (Orpheovirus, Cedratvirus, and Pithovirus). Finally, Topo IIA is present in all members of the order *Asfuvirales*, which includes *Asfarviridae* and related viruses (Kaumovirus, Faustovirus, and Pacmanvirus).

It turned out to be more challenging to assemble the data set of eukaryotic Topo IIA sequences because we found many fragments of Topo IIA in eukaryotic proteomes. Thus, we produced position-specific scoring matrices (PSSMs) for GyrA and GyrB proteins using alignments from the PFAM database (files provided as Supplementary material) and searched for coding sequences matching both profiles. We defined a list of fifty-two eukaryotic organisms, representative of the known eukaryotic diversity. When possible, we downloaded the corresponding proteomes and used PSSM as PSI-BLAST queries to obtain Topo IIA homologs. For those organisms where no proteome was available, we looked for transcriptomic data and performed *de novo* assembly using Trimmomatic v0.36 (Bolger, Lohse, and Usadel 2014) for the read pre-processing step, SortMeRNA v2.1b (Kopylova, Noé, and Touzet 2012) to filter out rRNA sequences, and Trinity v2.2.0 (Grabherr et al. 2011) for the assembly (two stars in Supplementary Table S1). We then used the GyrA and GyrB PSSMs as queries for a TBLASTN search against the assemblies and kept hits matching both profiles. One star in Supplementary Table S1 means that the accession number corresponds to the eukaprot db.id. (Richter et al. 2020).

### Phylogenetic analyses

All multiple amino-acid sequence alignments were performed using MAFFT v7.407 (Katoh and Standley 2013) and the E-INSI algorithm. Sites containing more than 50 per cent of gaps were filtered out. Of note, for the tree with the largest taxonomic sampling, we used Divvier v1.0 (Ali, Bogusz, and Whelan 2019) to reduce alignment errors with the MAFFT output.

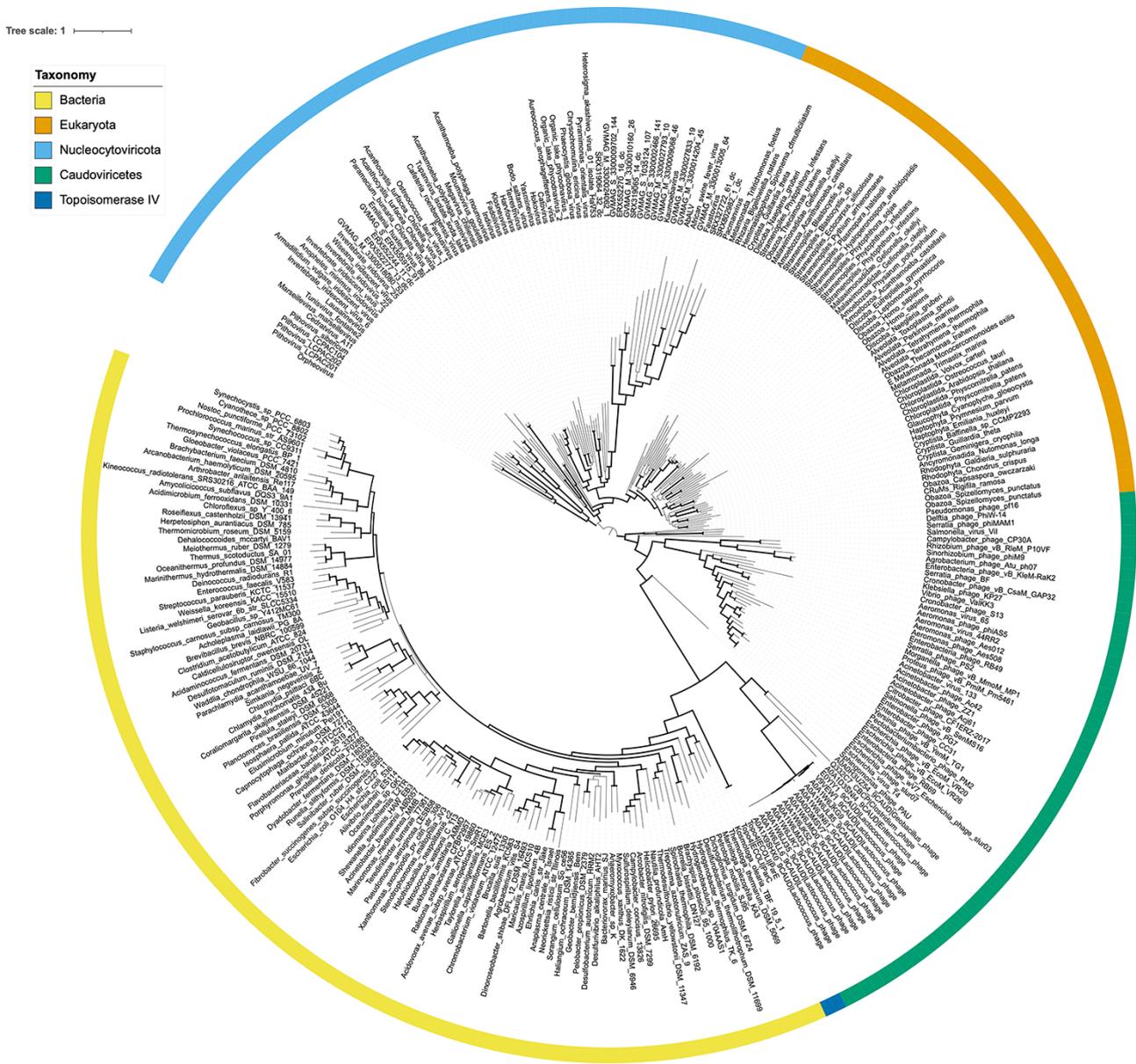
All phylogenetic analyses were performed using IQ-TREE v1.6.7.2 (Nguyen et al. 2015). We selected the best-fit model using the IQ-TREE's model finder (Wong et al. 2017) according to the Bayesian information criterion. We made the search for the best tree more thorough by using the 'allnni' option as well as setting the 'pers' parameter to 0.2 and the 'nstop' parameter 500. We always used ten independent runs (-runs option of IQ-TREE) and selected the best one. Confidence branch supports were assessed using the transfer bootstrap expectation (1,000 replicates except for the tree including all sequences, where 100 replicates were used) (Lemoine et al. 2018). We used iTOL v4.4.2 (Letunic and Bork 2016) to generate the figures. All trees and alignments are available 10.5281/zenodo.7081315.

## Results

### Viral Topo IIA branch between bacterial and eukaryotic Topo IIA in a global phylogeny

We first built a tree spanning the whole diversity of Topo IIA by including sequences from Bacteria, *Caudoviricetes*, *Nucleocytoviricota*, and eukaryotic Topo IIA (Fig. 2) (Supplementary Table S1). We did not include archaeal DNA gyrases because they branch within bacterial DNA gyrases in previous phylogenetic analyses (Forreter et al. 2007; Raymann et al. 2014). Importantly, we did not detect orthologs of eukaryotic-like Topo IIA in the metagenome-assembled genomes of different lineages of Asgard archaea, but only bacterial-like DNA gyrases.

The four groups of sequences (Bacteria, *Caudoviricetes*, *Nucleocytoviricota*, and eukaryotes) were clearly separated in the

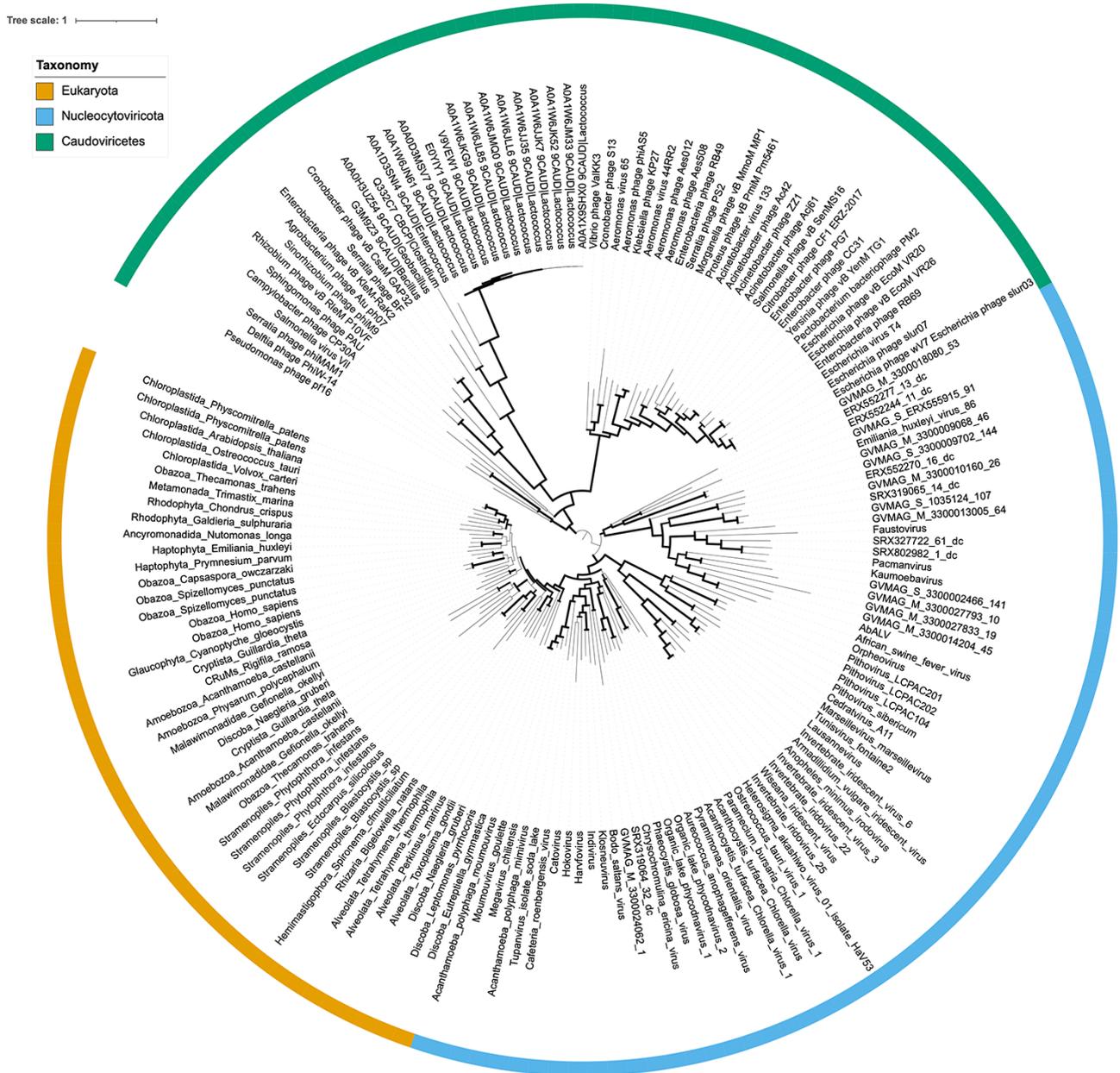


**Figure 2.** Phylogenetic tree of the Topo IIA maximum-likelihood tree for 269 Topo IIA proteins for Bacteria (113 sequences, including the two Topoisomerase IV proteins from *E. coli*, ParC and ParE), eukaryotes (53 sequences), Nucleocytoviricota (82 sequences), and Caudoviricetes (56 sequences). The outer circle colors represent the group to which the sequences belong. The selected model was LG + R9. Thick branches have a branch support (TBE) greater than 70 per cent.

tree (Fig. 2). The tree was arbitrarily rooted between Nucleocytoviricota and Caudoviricetes for convenience, dividing the tree into two clusters, one grouping eukaryotes and their viruses (Nucleocytoviricota) and the other grouping Bacteria and their viruses (Caudoviricetes). Both Bacteria and eukaryotes were monophyletic. In contrast, it was not possible to obtain the monophyly of either Caudoviricetes or Nucleocytoviricota. Importantly, in contrast to our previous analysis (Forterre et al. 2007), Nucleocytoviricota and eukaryotic Topo IIAs were not intermixed. This was probably due to the higher number of sequences used in the present analysis and the improved model for phylogenetic analyses that allowed the correct positioning of some eukaryotic sequences with long branches.

Although DNA viruses encode many viral-specific DNA replication proteins, they can sometimes recruit cellular replisome components (Krupović et al. 2010). We thus wondered if the

grouping of Caudoviricetes and Nucleocytoviricota between Bacteria and eukaryotes was due to the long-branch attraction artifact, with Caudoviricetes branching within Bacteria and Nucleocytoviricota within eukaryotes. This seemed unlikely considering the great divergence between viral Topo IIAs and their cellular counterparts. However, to test this hypothesis, we built several subtrees, both to remove groups with long branches and to enhance the signal by increasing the number of meaningfully aligned amino acids. After removing bacterial sequences, the most distant outgroup, we obtained a tree topology largely reproducing the relationships between Caudoviricetes, Nucleocytoviricota and eukaryotes observed in the global phylogeny (Fig. 3). More importantly, after removing both Bacteria and Caudoviricetes, Nucleocytoviricota remained well separated from eukaryotes (Fig. 4A, B). This indicates that the separation of Nucleocytoviricota and eukaryotes in the tree was not due to an attraction of Nucleocytoviricota by Caudoviricetes and/or



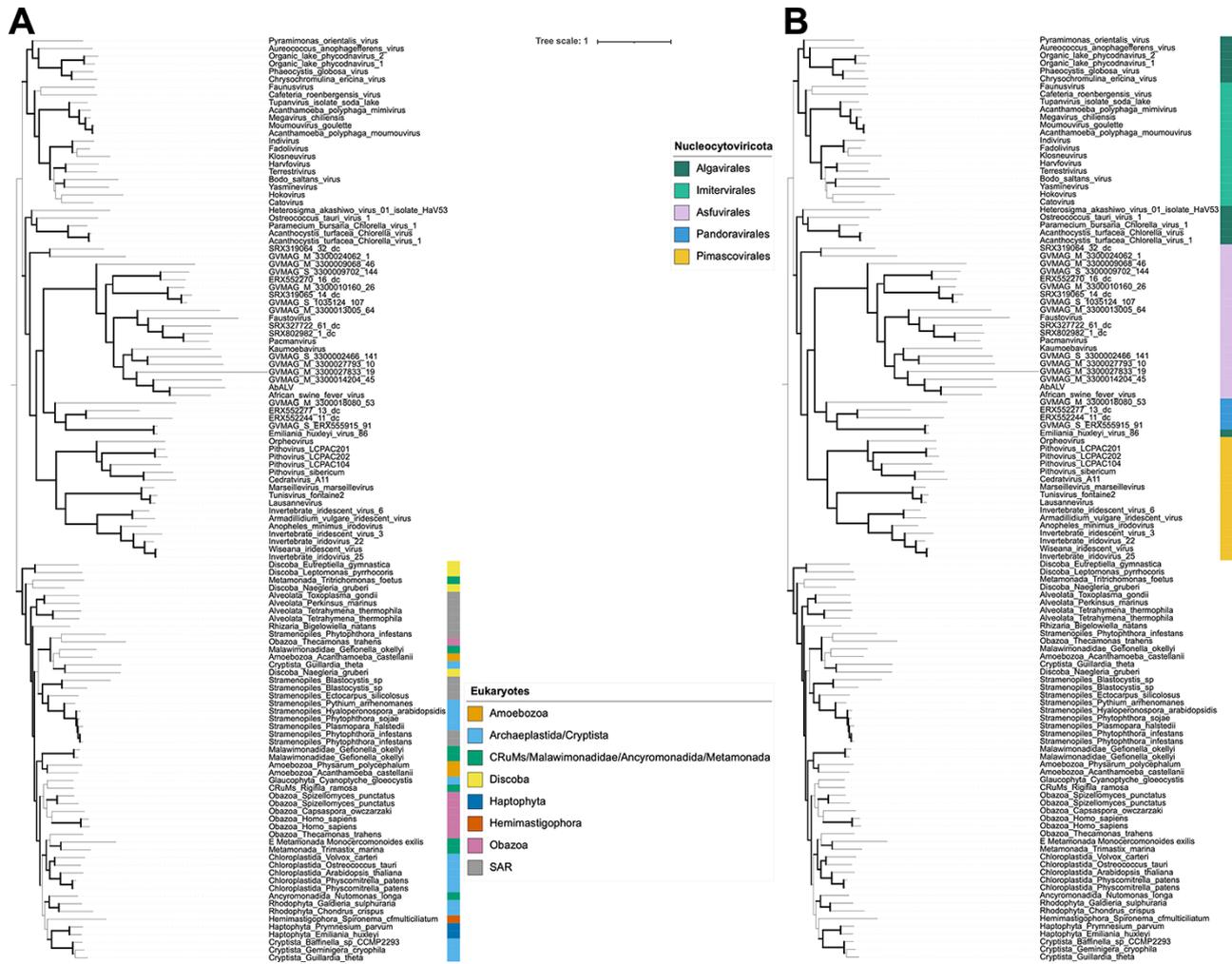
**Figure 3.** Phylogenetic tree of the Topo IIA for the eukaryotes, *Nucleocytoviricota* and *Caudoviricetes* Maximum-likelihood tree for 156 Topo IIA proteins for eukaryotes (53 sequences), *Nucleocytoviricota* (82 sequences) and *Caudoviricetes* (56 sequences). The outer circle colors represent the group to which the sequences belong. The selected model was LG+F+R8. Thick branches have a branch support (TBE) greater than 70 per cent.

Bacteria. Similarly, *Caudoviricetes* remained well separated from Bacteria after removing both *Nucleocytoviricota* and eukaryotes (Fig. 5).

The distribution and phylogeny of Topo IIA can provide information about their presence, or not, in the ancestors of each group of organisms. The ubiquity of DNA gyrase in bacteria leaves little doubt that this enzyme was present in the last bacterial common ancestor (LBCA). Similarly, the ubiquity of the single-subunit Topo IIA in Eukarya testifies to the presence of at least one Topo IIA in the LECA. However, we did not recover the monophyly of all major eukaryotic divisions in our phylogeny (Fig. 4A). Members of certain divisions were present in different parts of the tree, suggesting a complex history of Topo IIA during the diversification of eukaryotes, including gene duplication and gene loss.

Several eukaryotes indeed contain more than one Topo IIA gene (Forster et al. 2007). Some correspond to recent duplications (such as the Topo II $\alpha$  and Topo II $\beta$  in vertebrates), but others probably correspond to more ancient gene duplications or possibly gene transfers between eukaryotic lineages. With the root of the eukaryotic tree being still debated (Burki et al. 2019), it is difficult to propose a scenario for the evolution of Topo IIAs in eukaryotes. From our phylogenetic analyses, one cannot exclude that LECA already contained more than one Topo IIA.

The broad representation of Topo IIA in *Nucleocytoviricota* suggests that this enzyme was also present in the last *Nucleocytoviricota* common ancestor (LNCA) and was subsequently lost in a few lineages. This hypothesis is supported by the congruence between the phylogenetic tree of *Nucleocytoviricota* Topo IIA (Fig. 4B) and



**Figure 4.** Phylogenetic tree of the Topo IIA for the eukaryotes and Nucleocytoivricota maximum-likelihood tree for 100 Topo IIA proteins for eukaryotes (fifty-three sequences) and Nucleocytoivricota (eighty-two sequences). The selected model was LG+F+R8. Thick branches have a branch support (TBE) greater than 70 per cent. (A) The colored bar represents the eukaryotes classification. (B) The colored bar represents the Nucleocytoivricota classification.

the global phylogenetic classification of Nucleocytoivricota based on the concatenation of eight (core) genes present in most families of this phylum (Guglielmini et al. 2019). In the eight-core-gene phylogeny, Nucleocytoivricota were divided into two clusters that we named PAM (*Phycodnaviridae*, *Asfarviridae*, and *Mimiviridae*) and MAPI (*Marseilleviridae*, *Ascoviridae*, *Pitho*-like viruses, and *Iridoviridae*), respectively. The PAM cluster included viruses corresponding to the recently proposed orders *Imitervirales*, *Algavirales*, and *Asfurvirales*, whereas the MAPI cluster corresponded to the recently proposed order *Pimascovirales* (Koonin et al. 2020). In the Topo IIA NCLDV phylogenetic tree rooted with eukaryotes as the outgroup (Fig. 4B), we recovered the monophyly of *Imitervirales*, *Asfurvirales*, and *Pimascovirales*, the root of the tree being located within *Algavirales*.

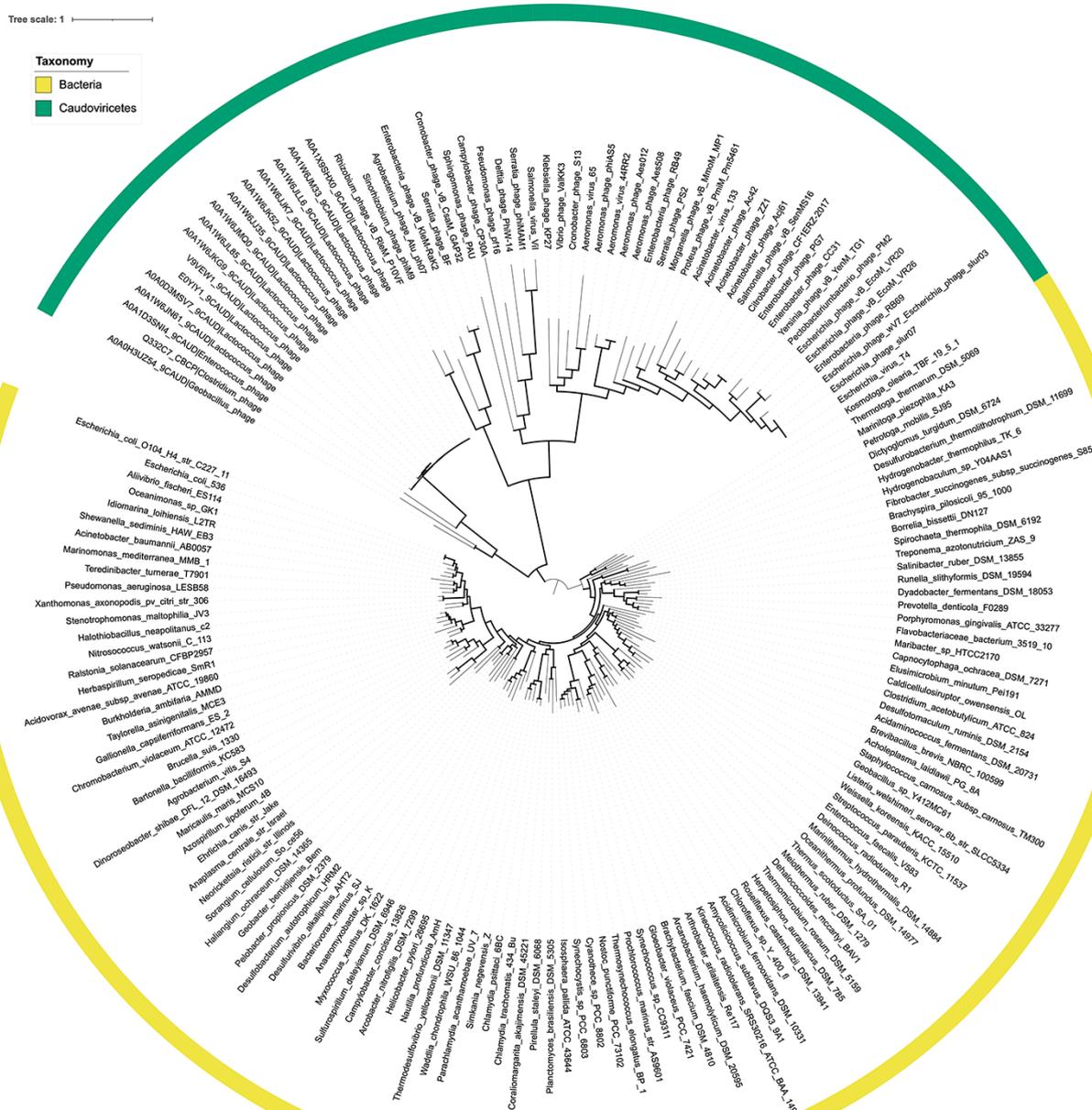
In contrast to the situation with Bacteria, eukaryotes, and Nucleocytoivricota, Topo IIAs are only present in a few subgroups of *Caudoviricetes*. Most Topo IIAs are encoded by T4-like myoviruses (i.e. viruses with contractile tails recently reclassified into the family *Straboviridae*) with larger genomes, suggesting that Topo IIA was present in the last common ancestor of this phage group. Topo IIA encoded by *Ackermannviridae* (another family of phages with contractile tails) branched within *Straboviridae*, suggesting lateral gene transfer between these viral families (Fig. 5B). Three of the

four Topo IIAs encoded by viruses infecting Firmicutes have been tentatively assigned to the family *Siphoviridae* (phages with long non-contractile tails). They were grouped with Topo IIA of T4-like viruses, as a sister clade of bacterial homologs if the tree is rooted between Nucleocytoivricota and *Caudoviricetes*.

### Discussion

To discuss possible evolutionary scenarios, we arbitrarily rooted the Topo IIA phylogenetic tree (Fig. 2) at the three possible positions between the four clusters (Fig. 6A–C).

Rooting the tree between Nucleocytoivricota and eukaryotes (Fig. 6A) would suggest that Nucleocytoivricota and eukaryotic Topo IIA originated from a common viral or cellular ancestor. This scenario appears unlikely since it also implies that *Caudoviricetes* Topo IIA originated from Nucleocytoivricota Topo IIAs and *in fine* that bacterial DNA gyrases themselves originated from Nucleocytoivricota via *Caudoviricetes*. In that case, one should imagine that the LBCA originated after the diversification of Nucleocytoivricota. Since this diversification took place before LECA, at the time when ancestral Nucleocytoivricota infected proto-eukaryotic hosts, this scenario would suggest that proto-eukaryotes evolved before Bacteria.

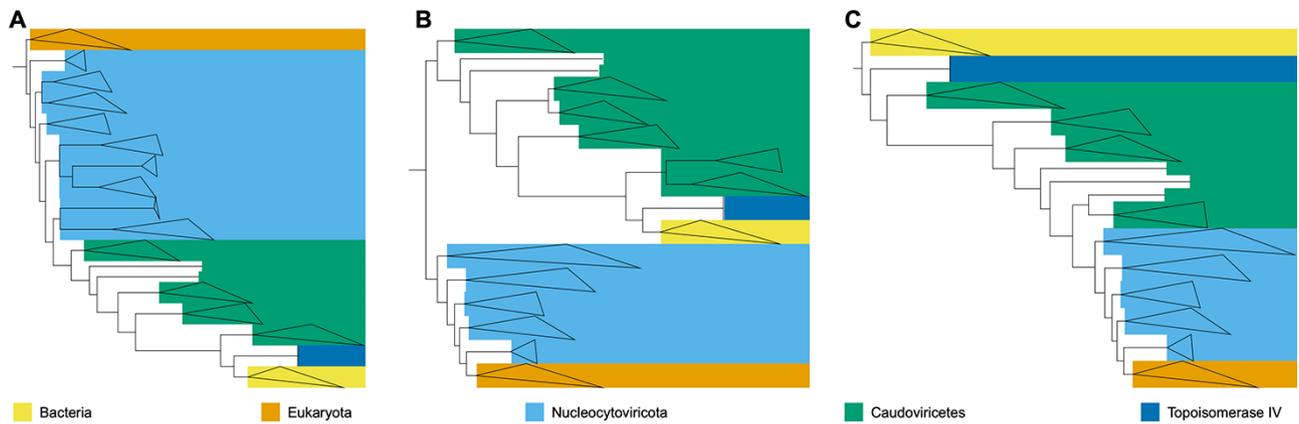


**Figure 5.** Phylogenetic tree of the Topo IIA for the Bacteria and Caudoviricetes. Maximum-likelihood tree for 167 Topo IIA proteins for Bacteria (111 sequences) and Caudoviricetes (56 sequences). The outer circle colors represent the group to which the sequences belong. The selected model was LG+R11. Thick branches have a branch support (TBE) greater than 70 per cent.

Rooting the tree between Caudoviricetes and Nucleocytoviricota (Fig. 6B) produced two clusters corresponding to Bacteria/Caudoviricetes and Nucleocytoviricota/eukaryotes. This rooting suggests that bacterial DNA gyrase originated from Topo IIAs of Caudoviricetes, whereas eukaryotic Topo IIA originated from those of Nucleocytoviricota. Considering the universal conservation of Topo IIA in Bacteria and eukaryotes, this scenario suggests that the transfer from viruses to cells took place before the emergence of the LBCA and LECA, respectively. Hence, both Caudoviricetes and Nucleocytoviricota should have originated and diversified before the emergence of the LBCA and LECA, infecting proto-bacterial and proto-eukaryotic hosts, respectively. Such ancient origin would explain the great divergence between the different versions of Topo IIA. The diversification of Nucleocytoviricota before LECA

is indeed supported by the RNA polymerase phylogenetic tree, including both viral and cellular enzymes (Guglielmini et al. 2019). Moreover, it has been suggested that Caudoviricetes, which also infect archaeal hosts (Liu et al. 2021), have diverged even prior to the emergence of the last universal cellular ancestor (Krupovic, Dolja, and Koonin 2020). In that scenario, the restricted distribution of Topo IIA to a few subgroups of Caudoviricetes seems surprising, but it resembles the restricted distribution of a recently described new version of RNA polymerase in a subgroup of these viruses (Weinheimer and Aylward 2020). In a variant of this hypothesis, one can even consider a possibility that the root of the Topo IIA tree is located within Caudoviricetes.

Rooting the tree between Bacteria and Caudoviricetes (Fig. 6C) produced a tree in which Topo IIA of Bacteria and Caudoviricetes



**Figure 6.** Schematic representation of the different possible rooting for the Topo IIA phylogenetic tree. The tree is the same as in Fig. 2. Monophyletic clades encompassing sequences from the same group have been collapsed and colored following Fig. 2. (A) Rooting using the eukaryotes as an outgroup. (B) Rooting using the Bacteria and Caudoviricetes as an outgroup. (C) Rooting using the Bacteria as an outgroup.

diverged from a common ancestor that predated the LBCA. In that case, the Caudoviricetes Topo IIA would have diverged from their bacterial counterparts before LBCA and continued during the diversification of Bacteria. The tree of Fig. 6C is consistent with the scenario in which eukaryotic viruses originated from a melting pot of bacterial viruses that infected the bacterium at the origin of mitochondria (Koonin, Dolja, and Krupovic 2015; Koonin, Krupovic, and Yutin 2015) or another ancient bacterial endosymbiont present in a proto-eukaryotic ancestor of modern eukaryotes. In that case, the Topo IIA from a Caudoviricetes present in this putative bacterial endosymbiont would have been transferred to an ancestor of Nucleocytoviricota, potentially with other components of the DNA replication machinery shared between Caudoviricetes and Nucleocytoviricota, including NAD-dependent DNA ligase (Yutin and Koonin 2009). Notably, a comparison of DNA replication machinery of all dsDNA viruses revealed a strong evolutionary and likely functional coupling between DNA topoisomerases and DNA ligases, with 96 per cent of viruses encoding DNA topoisomerases also carrying a gene for a ligase (Kazlauskas, Krupovic, and Venclovas 2016). To explain the great divergence between the Topo IIAs encoded by Caudoviricetes and those encoded by Nucleocytoviricota in terms of sequences and structure, this scenario entails that the rate of Topo IIA evolution increased dramatically following the transfer of the Caudoviricetes version into the lineage leading to the LNCA, with the fusion of the three Topo IIA subunits of Caudoviricetes Topo IIA into a single polypeptide.

These trees can be also interpreted in the framework of the ‘out of virus hypothesis’ for the origin of DNA topoisomerases (Forterre 2002; Forterre and Gabelle 2009). Conceivably, the different versions of Topo IIA originated in an ancient viral world. The scenario illustrated in Fig. 6C explicitly posits that proto-bacteria acquired their Topo IIA from an ancient Caudoviricetes, whereas in Fig. 6A,B, the cellular and viral Topo IIA evolved from a common ancestor, which may or may not have been a virus.

Finally, one should not forget that the results presented here reflect the current sampling of the biosphere and may change if/once new cellular and viral lineages are discovered, as has been the case with the discovery of giant viruses. Such discoveries would help to refine the scenarios proposed here or lead to new ones.

Importantly, if we exclude the unlikely conjecture in which the Topo IIA phylogenetic tree is rooted between Nucleocytoviricota

and eukaryotes (Fig. 6A), the branching of all eukaryotes within *Algavirales* in all other configurations suggests that a Topo IIA was introduced into eukaryotes from a member of this viral order. If the node at the base of the eukaryotic monophyletic clade corresponds to the position of LECA, as expected from the ubiquity of this enzyme in eukaryotes, the transfer of Topo IIA should have occurred before the emergence of LECA, i.e. from a member of *Algavirales* to a proto-eukaryote. Alternatively, Topo IIA could have been introduced from *Algavirales* to a particular eukaryotic lineage and later transferred from this lineage to all other lineages by horizontal gene transfer. This last scenario seems unlikely considering that Topo IIAs are present in all contemporary lineages of eukaryotes, without exception, and the enzyme is essential for several key functions conserved in all eukaryotic lineages. The fact that eukaryotes emerge in our analysis within *Algavirales* is consistent with the previous conclusion that the diversification of Nucleocytoviricota into several major families has predated the emergence of LECA (Guglielmini et al. 2019).

Our result raises an interesting question: which Topo II did proto-eukaryotes use before they captured the viral Topo IIA? A likely answer is that they relied on Topo IIB, since this enzyme is ubiquitous in Archaea, but also present in many eukaryotes. A Topo IIB-like protein with a very divergent V-B subunit is present in all eukaryotes and is part of the complex responsible for the initiation of meiotic recombination (Vrielynck et al. 2016; Robert, de Massy, and Grelon 2017), whereas several eukaryotic lineages, e.g. Viridiplantae, contain a *bona fide* archaeal-like Topo IIB (Forterre et al. 2007; Malik et al. 2007; Forterre and Gabelle 2009).

By comparing the Nucleocytoviricota core genes’ phylogeny with the phylogeny of the three eukaryotic nuclear RNA polymerases and those of Nucleocytoviricota, we have previously shown that two of the eukaryotic RNA polymerases, Pol I and Pol II, were probably introduced into the proto-eukaryotic lineage from Nucleocytoviricota (Guglielmini et al. 2019). This possibility was strongly supported in the case of the RNA polymerase II. It is worth noting that, like the position of Topo IIA in the present study, the RNA Pol II branched as a sister group of a clade including *Imitervirales* and some *Algavirales* in the RNA polymerase tree. One can speculate that these two proteins (that play a major role in the eukaryotic transcription machinery) were recruited together from the same virus. This would make sense from the viewpoint of cell physiology since the two enzymes interact both functionally and structurally.

Indeed, it has been shown that Topo IIA is a structural component of the holo-Pol II complex and is essential for efficient RNA synthesis of nucleosomal DNA by this complex (Mondal and Parvin 2001). Topo IIA is required to produce long RNA Pol II transcripts (>3 kb) in *Saccharomyces cerevisiae* (Joshi, Piña, and Roca 2012) and enhances the recruitment of RNA Pol II to promoters in budding yeast (Sperling et al. 2011). It is possible that both Topo IIA and RNA Pol II were domesticated by a proto-eukaryote following the integration of a *Nucleocytoviricota* encoding these genes into the host chromosome. Integration of entire or large portions of the genomes of some *Nucleocytoviricota* into the chromosome of modern eukaryotes has been well documented (Delaroque and Boland 2008; Cock et al. 2010; Filée 2014; Moniruzzaman et al. 2020).

The viral origin of eukaryotic Topo IIA, in addition to those of RNA Pol II and possibly RNA Pol I, strengthens the idea that giant viruses of the phylum *Nucleocytoviricota* (especially members of the class *Megaviricetes*) played a major role in shaping the identity of modern eukaryotes (Forterre and Gaïa 2016). It is likely that other important proteins involved in eukaryotic physiology originated from *Nucleocytoviricota*. This has been proposed for eukaryotic histones since the four histones from *Medusavirus* and *Marseillevirus* branch at the base of the eukaryotic clades of their respective homologs (Erives 2017; Yoshikawa et al. 2019) and for enzymes involved in messenger RNA capping (Bell 2020). However, in those cases, robust phylogenetic analyses remain to be carried out since the published papers are based on limited sampling of the eukaryotic and *Nucleocytoviricota* diversity. The viral origin of some of the major players in eukaryotic cell biology was probably not limited to nuclear components since we have recently reported that the eukaryotic cytoplasmic actin might have been recruited by proto-eukaryotes from an actin-like protein (viractin) encoded by some *Imitervirales* (Da Cunha et al. 2020).

The eukaryotic molecular fabric appears to be a melting pot of proteins that originated in *Nucleocytoviricota* (mainly *Megaviricetes*), those that emerged *de novo* in the eukaryotic stem branch, proteins inherited from the bacterial ancestor of mitochondria and chloroplasts, and proteins that had ancestors in Archaea (in two-domain scenarios) or in the common ancestor of Archaea and eukaryotes (in three-domain scenarios). Sorting out the viral component of our eukaryotic ancestors is now a major task for understanding eukaryogenesis.

## Data availability

All data are available in the main text or the [Supplementary material](#).

## Supplementary data

[Supplementary data](#) are available at *Virus Evolution* online.

## Acknowledgments

This work was partly supported by the Emergence(s) grant from Ville de Paris (to MK).

**Conflict of interest:** None declared.

## References

- Abergel, C., Legendre, M., and Claverie, J.-M. (2015) 'The Rapidly Expanding Universe of Giant Viruses: Mimivirus, Pandoravirus, Pithovirus and Mollivirus', *FEMS Microbiology Reviews*, 39: 779–96.
- Ali, R. H., Bogusz, M., and Whelan, S. (2019) 'Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments', *Molecular Biology and Evolution*, 36: 2340–51.
- Aylward, F. O. et al. (2021) 'A phylogenomic framework for charting the diversity and evolution of giant viruses', *PLoS Biology*, 19: e3001430.
- Bell, P. J. L. (2020) 'Evidence Supporting a Viral Origin of the Eukaryotic Nucleus', *Virus Research* [Internet], 289: 198168.
- Bergerat, A. et al. (1997) 'An Atypical Topoisomerase II from Archaea with Implications for Meiotic Recombination', *Nature*, 386: 414–7.
- Bolger, A. M., Lohse, M., and Usadel, B. (2014) 'Trimmomatic: A Flexible Trimmer for Illumina Sequence Data', *Bioinformatics*, 30: 2114–20.
- Burki, F. et al. (2019) 'The New Tree of Eukaryotes', *Trends in Ecology and Evolution* [Internet], xx: 43–55.
- Camacho, C. et al. (2009) 'BLAST+: Architecture and Applications', *BMC Bioinformatics*, 10: 1–9.
- Champoux, J. J. (2001) 'DNA Topoisomerases: Structure, Function, and Mechanism', *Annual Review of Biochemistry* [Internet], 70: 369–413.
- Cock, J. M. et al. (2010) 'The Ectocarpus Genome and the Independent Evolution of Multicellularity in Brown Algae', *Nature*, 465: 617–21.
- Coelho, J. et al. (2016) 'Functional Characterization and Inhibition of the Type II DNA Topoisomerase Coded by African Swine Fever Virus', *Virology*, 493: 209–16.
- . (2015) 'African Swine Fever Virus ORF P1192R Codes for a Functional Type II DNA Topoisomerase', *Virology*, 474: 82–93.
- Colson, P. et al. (2017) 'Mimivirus: Leading the Way in the Discovery of Giant Viruses of Amoebae', *Nature Reviews. Microbiology*, 15: 243–54.
- Da Cunha, V. et al. (2017) 'Lokiarchaea are Close Relatives of Euryarchaeota, Not Bridging the Gap between Prokaryotes and Eukaryotes', *PLoS Genetics*, 13: e1006810.
- . (2020) 'Giant Viruses Encode Actin-Related Proteins', *Molecular Biology and Evolution*, 39: msac022.
- Delaroque, N., and Boland, W. (2008) 'The Genome of the Brown Alga *Ectocarpus Siliculosus* Contains a Series of Viral DNA Pieces, Suggesting an Ancient Association with Large dsDNA Viruses', *BMC Evolutionary Biology*, 8: 110.
- Erives, A. J. (2017) 'Phylogenetic Analysis of the Core Histone Doublet and DNA Topo II Genes of Marseilleviridae: Evidence of Proto-eukaryotic Provenance', *Epigenetics & Chromatin* [Internet], 10: 1–11.
- Filée, J. (2014) 'Multiple Occurrences of Giant Virus Core Genes Acquired by Eukaryotic Genomes: The Visible Part of the Iceberg?', *Virology*, 466–467: 53–9.
- Forterre, P. (2002) 'The Origin of DNA Genomes and DNA Replication Proteins', *Current Opinion in Microbiology*, 5: 525–32.
- . (2006) 'DNA Topoisomerase V: A New Fold of Mysterious Origin', *Trends in Biotechnology*, 24: 245–7.
- . (2012) 'Introduction and Historical Perspective', In: Pommier, Y. (ed.) *DNA Topoisomerases and Cancer*, pp. 1–52. Springer: New York.
- Forterre, P., and Gadelle, D. (2009) 'Phylogenomics of DNA Topoisomerases: Their Origin and Putative Roles in the Emergence of Modern Organisms', *Nucleic Acids Research*, 37: 679–92.
- Forterre, P., and Gaïa, M. (2016) 'Giant Viruses and the Origin of Modern Eukaryotes', *Current Opinion in Microbiology*, 31: 44–9.
- Forterre, P. et al. (2007) 'Origin and Evolution of DNA Topoisomerases', *Biochimie*, 89: 427–46.
- . (2014) 'Plasmids from Euryarchaeota', *Microbiology Spectrum*, 2: PLAS-0027-2014.

- Freitas, F. B. et al. (2016) 'In Vitro Inhibition of African Swine Fever Virus-topoisomerase II Disrupts Viral Replication', *Antiviral Research*, 134: 34–41.
- Gadelle, D. et al. (2003) 'Phylogenomics of Type II DNA Topoisomerases', *BioEssays*, 25: 232–42.
- Grabherr, M. G. et al. (2011) 'Full-length Transcriptome Assembly from RNA-Seq Data without a Reference Genome', *Nature Biotechnology*, 29: 644–52.
- Guglielmini, J. et al. (2019) 'Diversification of Giant and Large Eukaryotic dsDNA Viruses Predated the Origin of Modern Eukaryotes', *Proceedings of the National Academy of Sciences of the United States of America*, 116: 19585–92.
- Hizume, K. et al. (2007) 'Topoisomerase II, Scaffold Component, Promotes Chromatin Compaction In Vitro in a Linker-histone H1-dependent Manner', *Nucleic Acids Research*, 35: 2787–99.
- Joshi, R. S., Piña, B., and Roca, J. (2012) 'Topoisomerase II Is Required for the Production of Long Pol II Gene Transcripts in Yeast', *Nucleic Acids Research*, 40: 7907–15.
- Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30: 772–80.
- Kazlauskas, D., Krupovic, M., and Venclovas, C. (2016) 'The Logic of DNA Replication in Double-stranded DNA Viruses: Insights from Global Analysis of Viral Genomes', *Nucleic Acids Research*, 44: 4551–64.
- Koonin, E. V., Dolja, V. V., and Krupovic, M. (2015) 'Origins and Evolution of Viruses of Eukaryotes: The Ultimate Modularity', *Virology*, 479–480: 2–25.
- Koonin, E. V. et al. (2020) 'Global Organization and Proposed Megataxonomy of the Virus World', *Microbiology and Molecular Biology Reviews*: MMBR, 84: e00061–19.
- Koonin, E. V., Krupovic, M., and Yutin, N. (2015) 'Evolution of Double-stranded DNA Viruses of Eukaryotes: From Bacteriophages to Transposons to Giant Viruses', *Annals of the New York Academy of Sciences*, 1341: 10–24.
- Kopylova, E., Noé, L., and Touzet, H. (2012) 'SortMeRNA: Fast and Accurate Filtering of Ribosomal RNAs in Metatranscriptomic Data', *Bioinformatics*, 28: 3211–7.
- Krupovic, M., Dolja, V. V., and Koonin, E. V. (2020) 'The LUCA and Its Complex Virome', *Nature Reviews. Microbiology*, 18: 661–70.
- Krupovič, M. et al. (2010) 'The Evolutionary History of Archaeal MCM Helicases: A Case Study of Vertical Evolution Combined with Hitchhiking of Mobile Genetic Elements', *Molecular Biology and Evolution*, 27: 2716–32.
- Lavrukhin, O. V. et al. (2000) 'Topoisomerase II from Chlorella Virus PBCV-1. Characterization of the Smallest Known Type II Topoisomerase', *The Journal of Biological Chemistry*, 275: 6915–21.
- Lemoine, F. et al. (2018) 'Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data', *Nature*, 556: 452–6.
- Letunic, I., and Bork, P. (2016) 'Interactive Tree of Life (iTOL) V3: An Online Tool for the Display and Annotation of Phylogenetic and Other Trees', *Nucleic Acids Research*, 44: 242–5.
- Liu, L. F., Liu, C. C., and Alberts, B. M. (1979) 'T4 DNA Topoisomerase: A New ATP-dependent Enzyme Essential for Initiation of T4 Bacteriophage DNA Replication', *Nature*, 281: 456–61.
- Liu, Y. et al. (2021) 'Diversity, Taxonomy and Evolution of Archaeal Viruses of the Class Caudoviricetes', *PLoS Biology*, 19: e3001442.
- Malik, S.-B. et al. (2007) 'Protist Homologs of the Meiotic Spo11 Gene and Topoisomerase VI Reveal an Evolutionary History of Gene Duplication and Lineage-specific Loss', *Molecular Biology and Evolution*, 24: 2827–41.
- Mondal, N., and Parvin, J. D. (2001) 'DNA Topoisomerase IIalpha Is Required for RNA Polymerase II Transcription on Chromatin Templates', *Nature*, 413: 435–8.
- Moniruzzaman, M. et al. (2020) 'Widespread Endogenization of Giant Viruses Shapes Genomes of Green Algae', *Nature* [Internet], 588: 141–5.
- Nguyen, L. T. et al. (2015) 'IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-likelihood Phylogenies', *Molecular Biology and Evolution*, 32: 268–74.
- Nielsen, C. F. et al. (2020) 'Topoisomerase II $\alpha$  Is Essential for Maintenance of Mitotic Chromosome Structure', *Proceedings of the National Academy of Sciences of the United States of America*, 117: 12131–42.
- Nitiss, J. L. (2009) 'DNA Topoisomerase II and Its Growing Repertoire of Biological Functions', *Nature Reviews. Cancer*, 9: 327–37.
- Pommier, Y. (2013) 'Drugging Topoisomerases: Lessons and Challenges', *ACS Chemical Biology*, 8: 82–95.
- Pommier, Y. et al. (2016) 'Roles of Eukaryotic Topoisomerases in Transcription, Replication and Genomic Stability', *Nature Reviews. Molecular Cell Biology* [Internet], 17: 703–21.
- Ramsay, L. et al. (2000) 'A Simple Sequence Repeat-based Linkage Map of Barley', *Genetics*, 156: 1997–2005.
- Raymann, K. et al. (2014) 'Global Phylogenomic Analysis Disentangles the Complex Evolutionary History of DNA Replication in Archaea', *Genome Biology and Evolution*, 6: 192–212.
- Richter, D. et al. (2020) *EukProt: A Database of Genome-scale Predicted Proteins across the Diversity of Eukaryotes. Figshare. Dataset.*
- Robert, T., de Massy, B., and Grelon, M. (2017) 'Topovil: A Molecular Scissor Essential for Reproduction', *Medecine Science (Paris)*, 33: 512–8.
- Sperling, A. S. et al. (2011) 'Topoisomerase II Binds Nucleosome-free DNA and Acts Redundantly with Topoisomerase I to Enhance Recruitment of RNA Pol II in Budding Yeast', *Proceedings of the National Academy of Sciences of the United States of America*, 108: 12693–8.
- Vos, S. M. et al. (2011) 'All Tangled Up: How Cells Direct, Manage and Exploit Topoisomerase Function', *Nature Reviews. Molecular Cell Biology* [Internet], 12: 827–41.
- Vrielynck, N. et al. (2016) 'A DNA Topoisomerase VI-like Complex Initiates Meiotic Recombination', *Science*, 351: 939–43.
- Wall, M. K., Mitchenall, L. A., and Maxwell, A. (2004) 'Arabidopsis thaliana DNA Gyrase Is Targeted to Chloroplasts and Mitochondria', *Proceedings of the National Academy of Sciences of the United States of America*, 101: 7821–6.
- Wang, J. C. (2009) *Untangling the Double Helix: DNA Entanglement and the Action of the DNA Topoisomerases.* Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.
- Weinheimer, A. R., and Aylward, F. O. (2020) 'A Distinct Lineage of Caudovirales that Encodes a Deeply Branching Multi-subunit RNA Polymerase', *Nature Communications* [Internet], 11: 1–9.
- Wong, T. K. F. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods* [Internet], 14: 587–9.
- Yoshikawa, G. et al. (2019) 'Medusavirus, a Novel Large DNA Virus Discovered from Hot Spring Water', *Journal of Virology*, 93: e02130–18.
- Yutin, N., and Koonin, E. V. (2009) 'Evolution of DNA Ligases of Nucleocytoplasmic Large DNA Viruses of Eukaryotes: A Case of Hidden Complexity', *Biology Direct*, 4: 1–14.