Phage Origin of Mitochondrion-Localized Family A DNA Polymerases in Kinetoplastids and Diplonemids

Ryo Harada[1] and Yuji Inagaki[1,2,*]

[1]Graduate School of Life and Environmental Sciences, University of Tsukuba, Japan

[2]Center for Computational Sciences, University of Tsukuba, Japan

*Corresponding author: E-mail: yuji@ccs.tsukuba.ac.jp.

Accepted: 3 January 2021

Abstract

Mitochondria retain their own genomes as other bacterial endosymbiont-derived organelles. Nevertheless, no protein for DNA replication and repair is encoded in any mitochondrial genomes (mtDNAs) assessed to date, suggesting that the nucleus primarily governs the maintenance of mtDNA. As the proteins of diverse evolutionary origins occupy a large proportion of the current mitochondrial proteomes, we anticipate finding the same evolutionary trend in the nucleus-encoded machinery for mtDNA maintenance. Indeed, none of the DNA polymerases (DNAPs) in the mitochondrial endosymbiont, a putative α-proteobacterium, seemingly had been inherited by their descendants (mitochondria), as none of the known types of mitochondrion-localized DNAP showed a specific affinity to the α-proteobacterial DNAPs. Nevertheless, we currently have no concrete idea of how and when the known types of mitochondrion-localized DNAPs emerged. We here explored the origins of mitochondrion-localized DNAPs after the improvement of the samplings of DNAPs from bacteria and phages/viruses. Past studies have revealed that a set of mitochondrion-localized DNAPs in kinetoplastids and diplonemids, namely PolIB, PolIC, PolID, PolI-Perk1/2, and PolI-dipl (henceforth designated collectively as "PolIBCD+") have emerged from a single DNAP. In this study, we recovered an intimate connection between PolIBCD+ and the DNAPs found in a particular group of phages. Thus, the common ancestor of kinetoplastids and diplonemids most likely converted a laterally acquired phage DNAP into a mitochondrion-localized DNAP that was ancestral to PolIBCD+. The phage origin of PolIBCD+ hints at a potentially large contribution of proteins acquired via nonvertical processes to the machinery for mtDNA maintenance in kinetoplastids and diplonemids.

Key words: DNA replication, DNA repair, autographivirus, Euglenozoa, lateral gene transfer, mitochondria.

Significance

Multiple types of DNA polymerases (DNAPs) localized in mitochondria have been identified among eukaryotes, but their precise origins have yet to be elucidated. This study reports that a type of mitochondrion-localized DNAP in kinetoplastids and diplonemids was derived from a DNAP of a particular group of phages. The results presented here suggest that the machinery for DNA maintenance in the mitochondria of kinetoplastid/diplonemid has been remodeled by the proteins acquired via nonvertical genetic processes.

Introduction

Mitochondria in the extant eukaryotes are the descendants of an endosymbiotic α-proteobacterium in the last eukaryotic common ancestor (Roger et al. 2017). The mitochondrial (mt) proteins, which are localized in mitochondria, are almost entirely nucleus-encoded and evolutionarily multifarious (Gabaldón and Huynen 2007; Wang and Wu 2014; Gray 2015). Only 10–20% of mt proteins were predicted to be of the α-proteobacterial origin, suggesting that the original proteome of the mitochondrial endosymbiont has been remodeled largely (Gray 2015). There are three possible evolutionary paths that co-opt non-α-proteobacterial proteins into the molecular machinery in mitochondria. The non-α-proteobacterial mt proteins could emerge 1) de novo, 2) by recycling of the pre-existing eukaryotic proteins, or 3) via lateral gene transfer. Mitochondria, in principle, retain their own genomes that have been descended from the mitochondrial endosymbiont, albeit the entire set of proteins required for mtDNA maintenance (replication and repair) is nucleus-encoded with some exceptions. To our knowledge, the genes encoding proteins for DNA maintenance have been identified rarely in mtDNAs and are regarded as the product of a lineage-specific horizontal gene transfer (Bilewitch and Degnan 2011) or the reminiscent of linear plasmids in mitochondria (Fricova et al. 2010; Swart et al. 2012; Nishimura et al. 2019). Thus, as a part of the mitochondrial proteome, the machinery for mtDNA maintenance may be dominated by non-α-proteobacterial proteins. Indeed, none of the known DNA polymerases (DNAPs) localized in mitochondria is most unlikely the direct descendants of the DNAPs in the α-proteobacterial endosymbiont that gave rise to the ancestral mitochondrion (see below).

Phylogenetically diverse eukaryotes possess family A (famA) DNAPs that are evolutionarily related to DNA polymerase I (PolI) in bacteria (Jung et al. 1987; Moriyama et al. 2011; Moriyama and Sato 2014). Some of famA DNAPs in eukaryotes are known to be localized in mitochondria (Krasich and Copeland 2017). So far, four distinct types of mitochondrion-localized famA DNAP have been identified. First, "plant and protist organellar DNA polymerase (POP)" appeared to be broadly distributed among eukaryotes (Moriyama et al. 2011; Hirakawa and Watanabe 2019). Second, animals and fungi are known to use DNA polymerase gamma (Polγ) for mtDNA maintenance (Graziewicz et al. 2006). The third type of mitochondrion-localized famA DNAP is "PolIA" that is shared among the members of the classes Kinetoplastea, Diplonemea, and Euglenida, which comprise the phylum Euglenozoa (Klingbeil et al. 2002; Harada et al. 2020). The members of Kinetoplastea and Diplonemea possess the fourth type of mitochondrion-localized famA DNAP. "PolIB," "PolIC," and "PolID" were originally reported from a model kinetoplastid *Trypanosoma brucei*, and later identified in the broad members of Kinetoplastea (Klingbeil et al. 2002; Harada et al. 2020). The three DNAPs were shown to be closely related to one another in phylogenetic analyses. A recent study further identified multiple DNAPs, which are closely related to but distinct from PolIB, C, or D, in an early-branching kinetoplastid *Perkinsela* sp. and diverse diplonemids (PolI-Perk1/2 and PolI-dipl; Harada et al. 2020). PolIB, C, D, and their related DNAPs were derived from a single molecule and, thus, can be regarded collectively as the fourth type of mitochondrion-localized famA DNAP (henceforth termed as "PolIBCD+" in this study). Pioneering studies have not considered any of the known mitochondrion-localized famA DNAPs as the direct descendant of PolI in the mitochondrial endosymbiont, but have failed to clarify how and when POP, Polγ, PolIA, and PolIBCD+ were established in the eukaryotic evolution (Moriyama et al. 2011; Hirakawa and Watanabe 2019; Harada et al. 2020).

In this study, we explored the origins of mitochondrion-localized famA DNAPs by analyzing an improved data set wherein the sequence sampling from bacteria and phages was improved drastically. We recovered the intimate affinity between PolIBCD+ and the famA DNAPs of a particular group of phages in the phylogenetic analyses. Furthermore, these DNAPs appeared to share a unique insertion of consecutive eight amino acid (aa) residues. Altogether, we conclude that the extant DNAPs belonging to PolIBCD+ were derived from a single phage famA DNAP acquired by the common ancestor of Kinetoplastea and Diplonemea. We also propose that PolIA in Euglenozoa emerged from a type of cytosolic famA DNAP (Polθ). The emergences of PolIA and PolIBCD+ may represent a tip of remodeling the machinery of mtDNA maintenance undergone in Kinetoplastea and Diplonemea.

Results

Prior to this study, the origin of none of the four types of mitochondrion-localized famA DNAPs (i.e., Polγ, POP, PolIA, and PolIBCD+) has been elucidated in detail. This study successfully clarified the origin of PolIBCD+ by analyzing phylogenetic alignments that are much richer in bacterial and phage famA DNAPs than those analyzed in the past studies. The sampling of the bacterial homologs was insufficient to reflect the diversity of bacteria in the previously published phylogenies of famA DNAPs (Moriyama et al. 2011; Hirakawa and Watanabe 2019; Harada et al. 2020). Furthermore, only a few famA DNAPs of phages have been included in the phylogenetic analyses. In this study, we prepared the "global famA DNAP" alignment by incorporating diverse bacterial and phage sequences (446 in total) deposited in public databases and 27 sequences that represent the four mitochondrion-localized types of DNAPs (Polγ, POP, PolIA, and PolIBCD+), a single cytosolic DNAP (Polθ), and a single plastid-localized DNAP found exclusively in apicomplexans and chrompodellids (PREX).

The global famA DNAP phylogeny reconstructed four clades, with all exclusively comprising the eukaryotic homologs: 1) POP, 2) PolIA plus Polθ, 3) PREX, and 4) Polγ (shaded in blue in fig. 1*A*; see the supplementary materials for the tree with sequence names; Inagaki and Harad [dataset] 2020). The maximum likelihood bootstrap values (MLBPs) for the four clades varied from 69% up to 100%. The POP, PolIA plus Polθ, or Polγ sequences showed no clear affinity to any bacterial or phage famA DNAPs, leaving their origins uncertain. The PREX sequences grouped with bifunctional 3′–5′ exonuclease/DNA polymerases found in phylogenetically limited bacteria as previously reported (Janouškovec et al. 2015; Hirakawa and Watanabe 2019; Harada et al. 2020). Curiously, the PolIBCD+ sequences were paraphyletic but

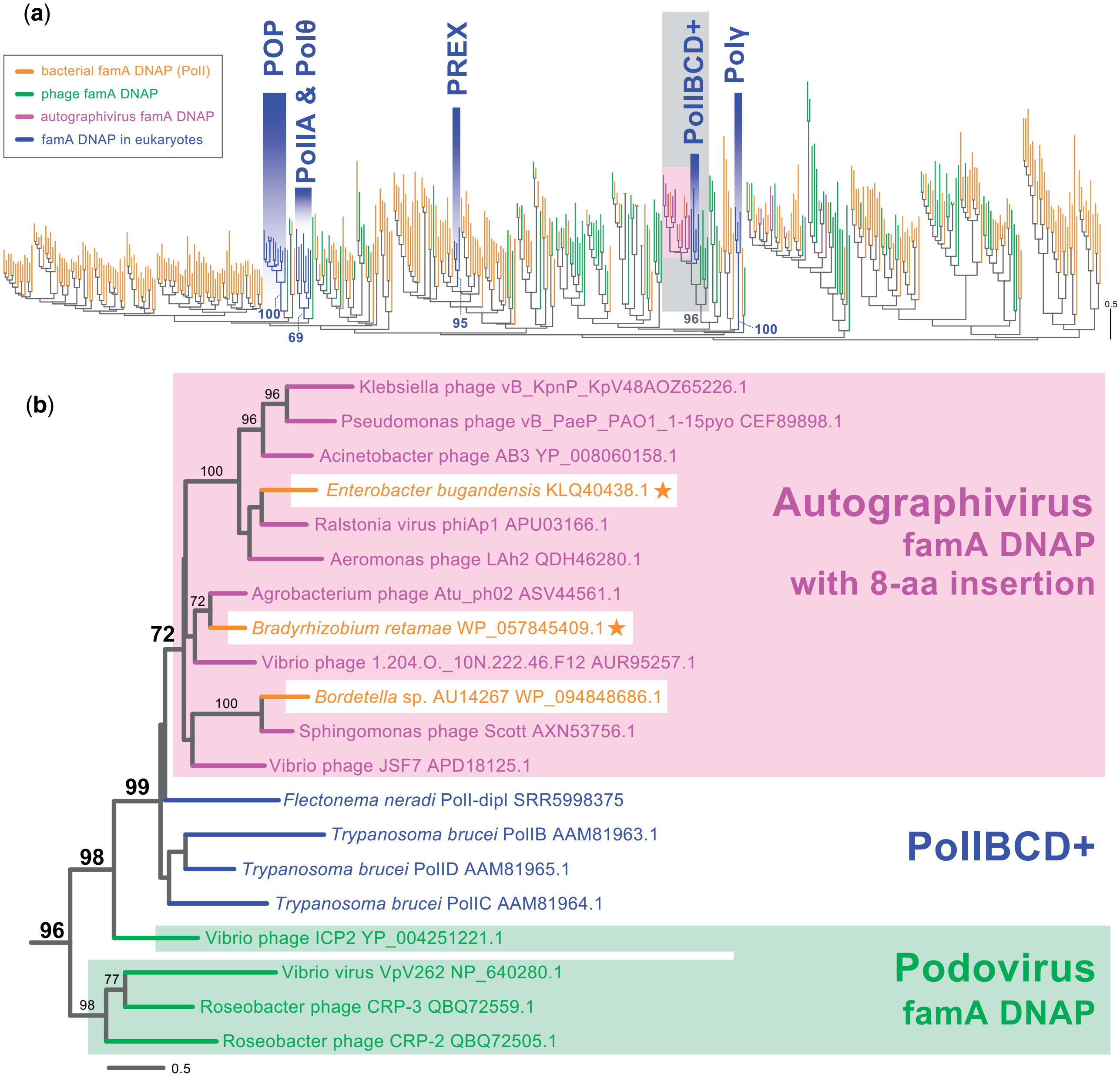
(a)
bacterial famA DNAP (PolI)
phage famA DNAP
autographivirus famA DNAP
famA DNAP in eukaryotes
POP
PolIA & Polθ
PREX
PolIBCD+
Polγ
100
69
95
96
100
0.5
(b)
96
96
100
72
72
100
99
98
96
77
98
Klebsiella phage vB_KpnP_KpV48AOZ65226.1
Pseudomonas phage vB_PaeP_PAO1_1-15pyo CEF89898.1
Acinetobacter phage AB3 YP_008060158.1
Enterobacter bugandensis KLQ40438.1 ★
Ralstonia virus phiAp1 APU03166.1
Aeromonas phage LAh2 QDH46280.1
Agrobacterium phage Atu_ph02 ASV44561.1
Bradyrhizobium retamae WP_057845409.1 ★
Vibrio phage 1.204.O._10N.222.46.F12 AUR95257.1
Bordetella sp. AU14267 WP_094848686.1
Sphingomonas phage Scott AXN53756.1
Vibrio phage JSF7 APD18125.1
Flectonema neradi PolI-dipl SRR5998375
Trypanosoma brucei PolIB AAM81963.1
Trypanosoma brucei PolID AAM81965.1
Trypanosoma brucei PolIC AAM81964.1
Vibrio phage ICP2 YP_004251221.1
Vibrio virus VpV262 NP_640280.1
Roseobacter phage CRP-3 QBQ72559.1
Roseobacter phage CRP-2 QBQ72505.1
Autographivirus famA DNAP with 8-aa insertion
PolIBCD+
Podovirus famA DNAP
0.5

Fig. 1.—Maximum likelihood (ML) phylogenetic tree inferred from an alignment of the famA DNAP sequences of bacteria, phages/viruses, and eukaryotes. (*A*) Overview of the entire ML tree. All of the sequence names are omitted. The bacterial and eukaryotic sequences are shown in orange and blue, respectively. The sequences of autographiviruses are shown in magenta. A subset of autographiviruses possesses famA DNAPs in the pink-shaded clade bears the characteristic insertion of eight amino acid residues (AGV^{+ins} famA DNAPs; see the main text for the details). Other phage/viral sequences are shown in green. Only ML bootstrap values of interest are shown. The subtree containing PolIBCD+ and AGV^{+ins} famA DNAP sequences (shaded in gray) is enlarged and presented as (*B*). ML bootstrap values greater than 70% are shown. AGV^{+ins} famA DNAP sequences marked by stars are of the putative lysogenic phages in bacterial genomes.

nested within a robustly supported clade mainly comprising the famA DNAP homologs of phages belonging to families Autographiviridae and Podoviridae (fig. 1*B*; this figure corresponds to the portion shaded in gray in fig. 1*A*). The famA DNAP homologs of autographiviruses and three bacteria formed a subclade with an MLBP of 72%. The coding regions of two out of the three bacterial famA DNAP homologs in this subclade (marked by stars in fig. 1*B*) are flanked by phage-like open reading frames (ORFs) in the corresponding genome assemblies deposited under the GenBank accession numbers LEDQ01000001.1 and NZ_LLYA01000167.1. Phage-like ORFs including that of famA DNAP encompass more than 40 kb consecutively in the two bacterial genomes. Thus, the two "bacterial famA DNAPs" are most likely of lysogenic autographiviruses in the bacterial genomes. On the other hand, no phage-like ORF was found around that of famA DNAP in the genome of *Bordetella* genomosp. 9 strain AU14267 (NZ_CP021109.1), suggesting that this bacterium horizontally acquired a famA DNAP gene from an autographivirus. The four PolIBCD+ sequences were positioned at the base of the Autographiviridae clade described above and the grouping of PolIBCD+ sequences and autographivirus famA DNAPs as a whole received an MLBP of 99% (fig. 1*B*). The global famA DNAP phylogeny strongly suggests an intimate evolutionary affinity between PolIBCD+ and autographivirus famA DNAPs.

The members of Autographiviridae commonly display head-to-tail capsid structures and possess double-stranded linear DNA genomes of approximately 41 kb in length. This viral family comprises nine subfamilies and 132 genera (Lavigne et al. 2008; Adriaenssens et al. 2020). We searched for autographivirus famA DNAPs in the GenBank nr database and detected 175 homologs of 99 members belonging to 57 genera and 76 unclassified members. Each of the 175 members of Autographiviridae seemingly possesses a single famA DNAP. Intriguingly, the autographivirus famA DNAPs were split into two types based on the presence/absence of "8-aa insertion" in the polymerase domain (supplementary fig. S1 and table S1, Supplementary Material online). In this study, we designate autographivirus famA DNAPs with 8-aa insertion as "AGV^{+ins} famA DNAPs." Each AGV^{+ins} famA DNAPs was predicted to possess only polymerase domain by InterProScan5 with the Pfam database (Jones et al. 2014; El-Gebali et al. 2019) (supplementary table S2, Supplementary Material online). AGV^{+ins} famA DNAPs were found in 40 members belonging to 23 genera, and 51 unclassified members (supplementary fig. S1, Supplementary Material online). Although only a subset of the 175 autographivirus famA DNAPs were included, the global famA DNAP phylogeny demonstrated the distant relationship between AGV^{+ins} famA DNAPs and other autographivirus famA DNAPs lacking 8-aa insertion (fig. 1*A*).

To re-examine the phylogenetic affinity between PolIBCD+ and AGV^{+ins} famA DNAPs, we selected nonredundant sequences from the 91 AGV^{+ins} famA DNAPs and aligned with 24 PolIBCD+ sequences and four famA DNAPs of phages belonging to a family Podoviridae as the outgroup. The second famA DNAP alignment was subjected to both ML and Bayesian methods. In the second phylogenetic analyses, AGV^{+ins} famA DNAPs and PolIBCD+ sequences formed a clade supported by an MLBP of 100% and a Bayesian posterior probability (BPP) of 1.0 (fig. 2). PolIBCD+ sequences appeared to possess eight amino acids that are most likely homologous to 8-aa insertion in AGV^{+ins} famA DNAPs (fig. 2), strengthening the phylogenetic affinity between PolIBCD+ and AGV^{+ins} famA DNAPs. Besides PolIBCD+ and AGV^{+ins} famA DNAPs, 8-aa insertion was found solely in the famA DNAP homolog of Vibrio phage ICP2 placed at the basal position of the clade of PolIBCD+ and AGV^{+ins} famA DNAPs (fig. 2). In the analyses of the second alignment, AGV^{+ins} famA DNAPs grouped together with an MLBP of 92% and a BPP of 0.99, excluding PolIBCD+ sequences that formed a clade with an MLBP of 72% and a BPP of 0.66 (fig. 2). The weak statistical support for the monophyly of PolIBCD+ sequences is not incongruent with their paraphyletic relationship reconstructed in the global famA DNAP analysis (fig. 1*B*).

Discussion

The phylogenetic analyses of the global alignment of famA DNAPs (fig. 1*A* and *B*) and the second alignment rich in AGV^{+ins} famA DNAP homologs (fig. 2) consistently recovered the specific affinity between PolIBCD+ and AGV^{+ins} famA DNAPs. These results strongly suggest that PolIBCD+ in the extant kinetoplastids and diplonemids can be traced back to a single autographivirus famA DNAP, particularly the one with 8-aa insertion. In other words, PolIBCD+ is a typical example of non-α-proteobacterial mt proteins acquired laterally from a phage. Prior to this study, the phage origins of mitochondrion-localized RNA polymerase (RNAP) and Twinkle mtDNA helicase/primase (simply termed Twinkle below) have been known (Shutt and Gray 2006). Unlike PolIBCD+ of which distribution is restricted to Kinetoplastea and Diplonemea, the mitochondrion-localized RNAP and Twinkle are ubiquitous among eukaryotes, implying that the last eukaryotic common ancestor (LECA) had already used the two proteins for transcription and replication in the mitochondrion, respectively. The phage origins of the mitochondrion-localized RNAP, DNAP, and helicase/primase prompt us to propose that individual eukaryotic lineages possess unique mt proteins acquired from phylogenetically diverse phages.

We failed to pinpoint the exact origin of PolIBCD+ even by analyzing the second alignment, wherein the known diversity of AGV^{+ins} famA DNAPs was covered (fig. 2). We might be able to find an AGV^{+ins} famA DNAP homolog that branches PolIBCD+ sequences directly in a future phylogenetic study covering the true diversity of phage famA DNAPs. In particular, we regard that autographivirus famA DNAP genes in bacterial genomes are significant. To our knowledge, no autographivirus has been reported to infect eukaryotes. Thus, the common ancestor of kinetoplastids and diplonemids may have acquired the famA DNAP gene from a lysogenic autographivirus in a bacterial genome. If so, the bacterial genomes harboring AGV^{+ins} famA DNAP genes are critical in investigating the origin of PolIBCD+ at a finer level than that in this study.

Members of classes Kinetoplastea and Diplonemea, together with Euglenida, share another type of mitochondrion-localized famA DNAP, namely PolIA (Harada et al. 2020). It is reasonable to postulate that the common ancestor of the three classes—most likely the ancestral

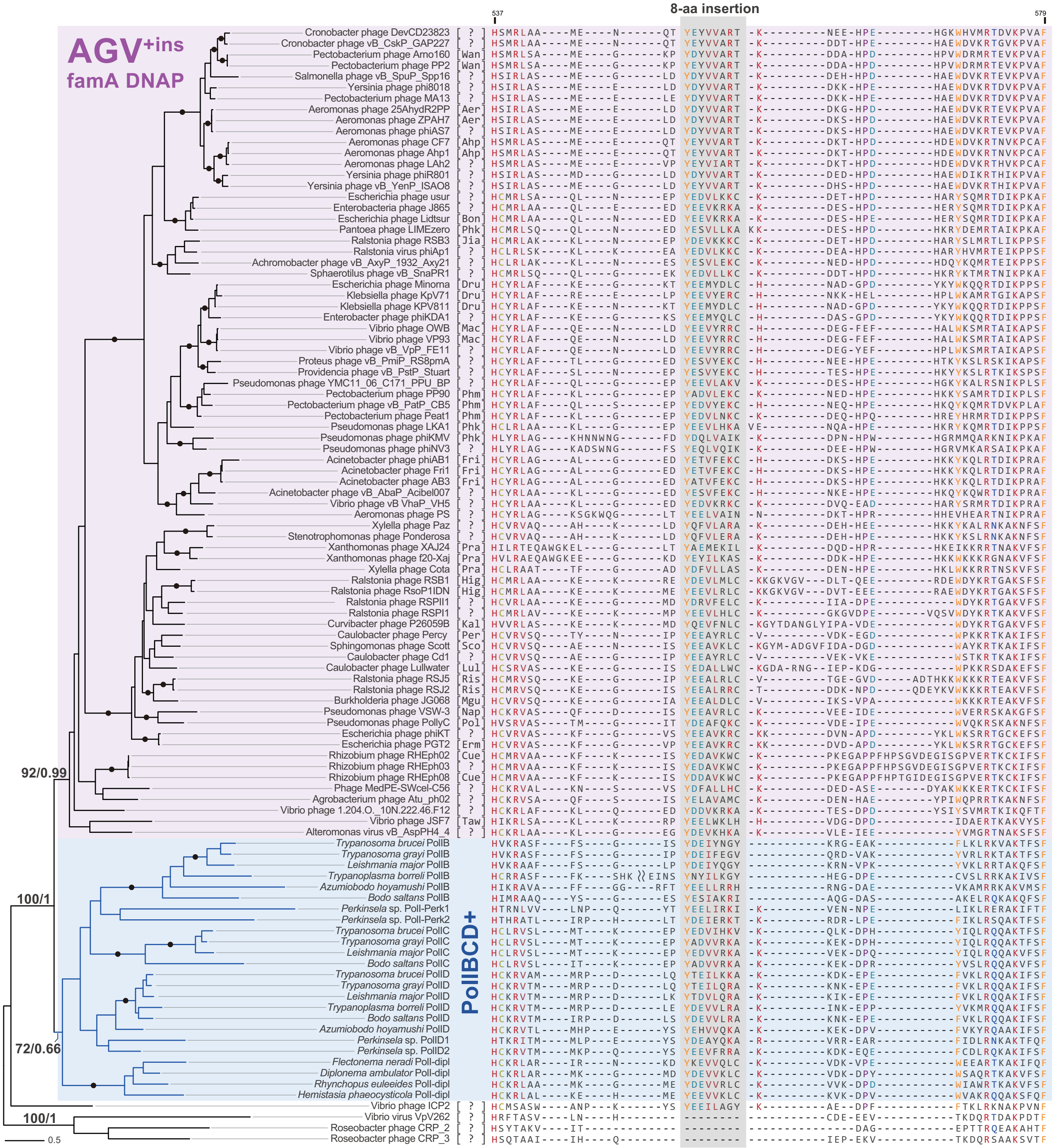

Fig. 2.—Phylogenetic relationship among 74 AGV^{+ins} famA DNAP and 24 PolIBCD+ sequences that share a unique insertion of eight amino acid residues (8-aa insertion). The tree topology and branch lengths inferred by the maximum likelihood (ML) method are shown on the left. ML bootstrap values (MLBPs) and Bayesian posterior probabilities (BPPs) for only the nodes critical to infer the origin of PolIBCD+ are shown. As ML and Bayesian analyses reconstructed the essentially same tree topology, only BPPs for the selected nodes are presented. The nodes supported by an MLBP of 100% and a BPP of 1.0 are marked by dots. The genus names of the autographiviruses (and podoviruses), from which famA DNAPs were sampled, are given in brackets. Abbreviations are follows: Aer, Aerosvirus; Ahp, Ahphunavirus; Bon, Bonnellvirus; Cue, Cuernavacavirus; Dru, Drulisvirus; Erm, Ermolevavirus; Fri, Friunavirus; Hig, Higashivirus; Jia, Jiaoyazivirus; Kal, Kalppathivirus; Lul, Lullwatervirus; Mac, Maculvirus; Mgu, Mguuvirus; Nap, Napahaivirus; Per, Percyvirus; Phk, Phikmvvirus; Phm, Phimunavirus; Pol, Pollyceevirus; Pra, Pradovirus; Ris, Risjevirus; Sco, Scottvirus; Taw, Tawavirus; Wan, Wanjuvirus; ?, unclassified. The amino acid sequences of 8-aa insertions and their flanking regions are shown on the right. 8-aa insertions are shaded in gray. The residues are colored according to their degrees of conservation. The amino acid residue numbers shown on the left and right edges of the alignment are based on the famA DNAPs of Cronobacter phage DevCD23823 (YP_009223394.1).

euglenozoan—had established the ancestral PolIA. Although the origin of PolIA has not been addressed explicitly, past studies recovered the phylogenetic link between PolIA and Polθ, a type of famA DNAP operated in the cytosol of eukaryotic cells. The original study reporting PolIA, B, C, and D in *Trypanosoma brucei* has hinted at the phylogenetic affinity between PolIA and Polθ (Klingbeil et al. 2002). A recent phylogeny including famA DNAPs sampled from eukaryotes and limited bacteria (note that no phage homolog was included) reconstructed a clade of PolIA and Polθ sequences with high statistical support (Harada et al. 2020). The PolIA–Polθ affinity persisted even after the sampling of famA DNAPs from bacteria and phages was improved drastically in this study (fig. 1*A*). We here propose that the ancestral PolIA was derived from a Polθ homolog followed by the change in subcellular localization from the cytosol to the mitochondrion. Noteworthy, the evolutionary processes yielded PolIA and PolIBCD+ are different substantially from each other. The former emerged through the recycling of a pre-existing eukaryotic protein, whereas the latter is of phage origin (see above).

The repertories of mitochondrion-localized DNAPs in euglenozoans appeared to be more complex than those in the majority of other eukaryotes in which a single type of mitochondrion-localized DNAP (i.e., POP or Polγ) seemingly operates. The complexity in the repertory of DNAPs in euglenozoan mitochondria seems to coincide with that in the structure of their mtDNAs (Lukeš et al. 2002; Roy et al. 2007; Spencer and Gray 2011; Dobáková et al. 2015; Yabuki et al. 2016; Burger and Valach 2018). Nevertheless, it is unlikely that the non-α-proteobacterial background is restricted to PolIA and PolIBCD+ among the proteins involved in mtDNA maintenance. Rather, the machinery for mtDNA maintenance in the common ancestor of kinetoplastids and diplonemids (and its descendants) are heavily remodeled by both incorporating exogenous proteins via lateral gene transfer and recycling the pre-existing nucleus-encoded proteins. The above conjecture can be examined only after we identify the major proteins involved in DNA maintenance in kinetoplastid/diplonemid mitochondria and their evolutionary origins.

Finally, we here explore briefly the early evolution of mitochondrion-localized DNAP. None of the known types of mitochondrion-localized famA DNAP showed any affinity to PolI of the extant α-proteobacteria (fig. 1*A*), suggesting that the famA DNAP operated in the mitochondrial endosymbiont was discarded in the early eukaryotic evolution. It is intriguing to point out that POP has been identified in phylogenetically diverse groups, whereas Polγ, PolIA, and PolIBCD+ appeared to be specific to Opisthokonta, Euglenozoa, and a subclade in Euglenozoa (i.e., Kinetoplastea and Diplonemea), respectively (Moriyama et al. 2011; Moriyama and Sato 2014; Hirakawa and Watanabe 2019; Harada et al. 2020). The parsimonious interpretation of the distribution of POP, Polγ, PolIA, and PolIBCD+ in the tree of eukaryotes nominates POP to be the most ancient mitochondrion-localized DNAP. If so, the switch of mitochondrion-localized DNAP from POP to Polγ may have occurred in the ancestral opisthokont species. In contrast, it is not straightforward to hypothesize how the inventory of mitochondrion-localized famA DNAPs in euglenozoans had been shaped. Recent phylogenomic studies unified jakobids, heteroloboseans, *Tsukubamonas globosa*, and euglenozoans into one of the major eukaryotic assemblages Discoba (Rodríguez-Ezpeleta et al. 2007; Yabuki et al. 2011, 2018; Kamikawa et al. 2014). Nevertheless, to our knowledge, none of POP, Polγ, PolIA, or PolIBCD+ has been found in the discobid members except euglenozoans. Thus, even though euglenids use mitochondrion-localized POPs (POP_e1 and POP_Rhabd; Harada et al. 2020), it is uncertain whether the euglenid POPs were the direct descendants of the hypothetical POP established in the early eukaryotic evolution. Unfortunately, there are many lineages for which the information of mitochondrion-localized DNAPs is absent and we are currently not in the position to propose any scenarios for the early evolution of mitochondrion-localized DNAP (including that in the LECA) with confidence. To address the above issue appropriately, we need to understand the true diversity and distribution of mitochondrion-localized famA DNAPs in eukaryotes in the future.

Materials and Methods

Global Phylogeny of famA DNAPs

We searched for the amino acid (aa) sequences of bacterial and phage famA DNAPs in the NCBI nr database as of March 6, 2020, by BlastP using the polymerase domain of *Escherichia coli* PolI (KHH06131.1; the portion corresponding to the 491st–928th aa residues) as a query (Camacho et al. 2009; Sayers et al. 2020). We retrieved the sequences matched to the query with *E* values $\leq 1 \times 10^{-4}$ and covered more than 200 aa in the polymerase domain. Note that the sequences derived from metagenome analyses were excluded from this study. The redundancy within famA DNAP sequences was removed by a cluster analysis using CD-HIT v4.7 with a threshold of 40% (Li and Godzik 2006; Fu et al. 2012). We finally selected 119 and 327 aa sequences of phage and bacterial famA DNAPs, respectively, for the downstream analyses (see below).

The bacterial and phage famA DNAP aa sequences (446 in total) were aligned with those in eukaryotes (27 in total), namely 1) mitochondrion-localized famA DNAPs in Kinetoplastea and Diplonemea (PolIA, B, C, D, and PolI-dipl), 2) mitochondrion-localized famA DNAPs in animals and fungi (Polγ), 3) Polθ localized in the cytosol, 4) mitochondrion and/or plastid-localized famA DNAPs in diverse eukaryotes (POP), and 5) plastid-localized famA DNAPs in apicomplexan parasites and their relatives (PREX). The aa sequences were aligned by MAFFT v7.455 with the L-INS-i model (Katoh and

Standley 2013). Ambiguously aligned positions were discarded manually, and gap-containing positions were trimmed by using trimAl v1.4 with the -gt 0.95 option (Capella-Gutiérrez et al. 2009). The final "global famA DNAP" alignment comprised 473 sequences with 316 unambiguously aligned aa positions. The final global famA alignment is provided as a part of the supplementary materials (Inagaki and Harad[dataset] 2020). We subjected this alignment to the ML phylogenetic analysis by IQ-TREE v1.6.12 using the LG + Γ + F + C60 + PMSF model that was selected by ModelFinder with -madd option (Nguyen et al. 2015; Kalyaanamoorthy et al. 2017; Wang et al. 2018). The guide tree was obtained using the LG + Γ + F model, which was selected by ModelFinder with -m option. The statistical support for each bipartition in the ML tree was calculated by a 100-replicate nonparametric bootstrap analysis.

Phylogenetic Analyses of an Alignment Rich in Autographivirus famA DNAPs

We retrieved 175 famA DNAP aa sequences of autographiviruses from the NCBI nr database. The details of the survey were the same as described above. The 175 famA DNAPs were sampled from 99 members belonging to 57 genera and 76 unclassified members in the family Autographiviridae. The autographivirus famA DNAPs were found to comprise two types based on the presence/absence of an insertion of eight aa residues (8-aa insertion; see above). The famA DNAPs with 8-aa insertion (AGV^{+ins} famA DNAPs) appeared to be closely related to PolIBCD+, mitochondrion-localized famA DNAPs in kinetoplastids (PolIB, C, D, PolI-Perk1/2) and that in diplonemids (PolI-dipl). The redundancy among the AGV^{+ins} famA DNAPs was reduced by a cluster analysis using CD-HIT v4.7 with a threshold of 90%. Finally, we aligned the aa sequences of 74 AGV^{+ins} famA DNAPs, 24 PolIBCD+, and famA DNAPs of four members of Podoviridae by MAFFT v7.455 with the L-INS-i model. Ambiguously aligned positions were discarded manually, and gap-containing positions were trimmed by using trimAl v1.4 with the -gt 0.9 option. The final version of the second alignment is provided as a part of the supplementary materials (Inagaki and Harad[dataset] 2020). The final alignment containing 102 sequences with 581 unambiguously aligned aa positions was subjected to both ML and Bayesian phylogenetic analyses. The ML and ML bootstrap analyses were performed as described above. For Bayesian analysis using Phylobayes v4.1 (Lartillot et al. 2009), we ran four Markov Chain Monte Carlo chains for 100,000 cycles with burn-in of 25,000 (maxdiff = 0.09472) and calculated the consensus tree with branch lengths and BPPs from the remaining trees. The aa substitution model was set to CAT + GTR in Phylobayes analysis described above.

Supplementary Material

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

Acknowledgments

This work was supported by the grants from the Japanese Society for Promotion of Sciences awarded to Y.I. (numbers 18KK0203 and 19H03280). We also thank three reviewers for their constructive suggestions and comments on the manuscript.

Data Availability

The supplementary data are available from the Dryad Digital Repository: https://doi.org/10.5061/dryad.9kd51c5fv.

Literature Cited

Adriaenssens EM, et al. 2020. Taxonomy of prokaryotic viruses: 2018–2019 update from the ICTV bacterial and archaeal viruses subcommittee. Arch Virol. 165(5):1253–1260.

Bilewitch JP, Degnan SM. 2011. A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. BMC Evol Biol. 11(1):228.

Burger G, Valach M. 2018. Perfection of eccentricity: mitochondrial genomes of diplonemids. IUBMB Life. 70(12):1197–1206.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Dobáková E, Flegontov P, Skalický T, Lukeš J. 2015. Unexpectedly streamlined mitochondrial genome of the euglenozoan *Euglena gracilis*. Genome Biol Evol. 7(12):3358–3367.

El-Gebali S, et al. 2019. The Pfam protein families database in 2019. Nucleic Acids Res. 47(D1):D427–D432.

Fricova D, et al. 2010. The mitochondrial genome of the pathogenic yeast *Candida subhashii*: GC-rich linear DNA with a protein covalently attached to the 5′ termini. Microbiology 156(7):2153–2163.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28(23):3150–3152.

Gabaldón T, Huynen MA. 2007. From endosymbiont to host-controlled organelle: the hijacking of mitochondrial protein synthesis and metabolism. PLoS Comput Biol. 3(11):2209–2218.

Gray MW. 2015. Mosaic nature of the mitochondrial proteome: implications for the origin and evolution of mitochondria. Proc Natl Acad Sci U S A. 112(33):10133–10138.

Graziewicz MA, Longley MJ, Copeland WC. 2006. DNA polymerase γ in mitochondrial DNA replication and repair. Chem Rev. 106(2):383–405.

Harada R, et al. 2020. Inventory and evolution of mitochondrion-localized family A DNA polymerases in Euglenozoa. Pathogens 9(4):257.

Hirakawa Y, Watanabe A. 2019. Organellar DNA polymerases in complex plastid-bearing algae. Biomolecules 9(4):140–112.

Inagaki Y, Harada R. [dataset] 2020. Data from: phage origin of mitochondrion-localized family A DNA polymerases in kinetoplastids and diplonemids. Dryad. doi: 10.5061/dryad.9kd51c5fv.

Janouškovec J, et al. 2015. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc Natl Acad Sci U S A. 112(33):10200–10207.

Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240.

Jung GH, Leavitt MC, Hsieh JC, Ito J. 1987. Bacteriophage PRD1 DNA polymerase: evolution of DNA polymerases. Proc Natl Acad Sci U S A. 84(23):8287–8291.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 14(6):587–589.

Kamikawa R, et al. 2014. Gene content evolution in discobid mitochondria deduced from the phylogenetic position and complete mitochondrial genome of *Tsukubamonas globosa*. Genome Biol Evol. 6(2):306–315.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Klingbeil MM, Motyka SA, Englund PT. 2002. Multiple mitochondrial DNA polymerases in *Trypanosoma brucei*. Mol Cell. 10(1):175–186.

Krasich R, Copeland WC. 2017. DNA polymerases in the mitochondria: a critical review of the evidence. Physiol Behav. 22(1):692–709.

Lartillot N, Lepage T, Blanquart S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25(17):2286–2288.

Lavigne R, Seto D, Mahadevan P, Ackermann HW, Kropinski AM. 2008. Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP-based tools. Res Microbiol. 159(5):406–414.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659.

Lukeš J, et al. 2002. Kinetoplast DNA network: evolution of an improbable structure. Eukaryot Cell. 1(4):495–502.

Moriyama T, Sato N. 2014. Enzymes involved in organellar DNA replication in photosynthetic eukaryotes. Front Plant Sci. 5:480.

Moriyama T, Terasawa K, Sato N. 2011. Conservation of POPs, the plant organellar DNA polymerases, in eukaryotes. Protist 162(1):177–187.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Nishimura Y, et al. 2019. Horizontally-acquired genetic elements in the mitochondrial genome of a centrohelid *Marophrys* sp. Sci Rep. 9(1):4850.

Rodríguez-Ezpeleta N, et al. 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of Jakobids and Cercozoans. Curr Biol. 17(16):1420–1425.

Roger AJ, Muñoz-Gómez SA, Kamikawa R. 2017. The origin and diversification of mitochondria. Curr Biol. 27(21):R1177–R1192.

Roy J, Faktorová D, Lukeš J, Burger G. 2007. Unusual mitochondrial genome structures throughout the Euglenozoa. Protist 158(3):385–396.

Sayers EW, et al. 2020. Database resources of the national center for biotechnology information. Nucleic Acids Res. 48(D1):D9–D16.

Shutt TE, Gray MW. 2006. Bacteriophage origins of mitochondrial replication and transcription proteins. Trends Genet. 22(2):90–95.

Spencer DF, Gray MW. 2011. Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. Mol Genet Genomics. 285(1):19–31.

Swart EC, et al. 2012. The *Oxytricha trifallax* mitochondrial genome. Genome Biol Evol. 4(2):136–154.

Wang HC, Minh BQ, Susko E, Roger AJ. 2018. Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. Syst Biol. 67(2):216–235.

Wang Z, Wu M. 2014. Phylogenomic reconstruction indicates mitochondrial ancestor was an energy parasite. PLoS One 9(10):e110685.

Yabuki A, et al. 2011. *Tsukubamonas globosa* n. gen., n. sp., a novel excavate flagellate possibly holding a key for the early evolution in "Discoba". J Eukaryot Microbiol. 58(4):319–331.

Yabuki A, Gyaltshen Y, Heiss AA, Fujikura K, Kim E. 2018. *Ophirina amphinema* n. gen., n. sp., a new deeply branching discobid with phylogenetic affinity to jakobids. Sci Rep. 8(1):1–14.

Yabuki A, Tanifuji G, Kusaka C, Takishita K, Fujikura K. 2016. Hyper-eccentric structural genes in the mitochondrial genome of the algal parasite *Hemistasia phaeocysticola*. Genome Biol Evol. 8(9):2870–2878.

Associate editor: Martin Embley