Journal of Molecular Evolution (2023) 91:669–686
https://doi.org/10.1007/s00239-023-10128-x

ORIGINAL ARTICLE

Systematic Analysis of Diverse Polynucleotide Kinase Clp1 Family Proteins in Eukaryotes: Three Unique Clp1 Proteins of *Trypanosoma brucei*

Motofumi Saito[1,2] · Rerina Inose[1] · Asako Sato[1] · Masaru Tomita[1,2,3] · Haruo Suzuki[1,3] · Akio Kanai[1,2,3]

Received: 18 March 2023 / Accepted: 1 August 2023 / Published online: 22 August 2023

Abstract

The Clp1 family proteins, consisting of the Clp1 and Nol9/Grc3 groups, have polynucleotide kinase (PNK) activity at the 5′ end of RNA strands and are important enzymes in the processing of some precursor RNAs. However, it remains unclear how this enzyme family diversified in the eukaryotes. We performed a large-scale molecular evolutionary analysis of the full-length genomes of 358 eukaryotic species to classify the diverse Clp1 family proteins. The average number of Clp1 family proteins in eukaryotes was 2.3 ± 1.0, and most representative species had both Clp1 and Nol9/Grc3 proteins, suggesting that the Clp1 and Nol9/Grc3 groups were already formed in the eukaryotic ancestor by gene duplication. We also detected an average of 4.1 ± 0.4 Clp1 family proteins in members of the protist phylum Euglenozoa. For example, in *Trypanosoma brucei*, there are three genes of the Clp1 group and one gene of the Nol9/Grc3 group. In the Clp1 group proteins encoded by these three genes, the C-terminal domains have been replaced by unique characteristics domains, so we designated these proteins *Tb*-Clp1-t1, *Tb*-Clp1-t2, and *Tb*-Clp1-t3. Experimental validation showed that only *Tb*-Clp1-t2 has PNK activity against RNA strands. As in this example, N-terminal and C-terminal domain replacement also contributed to the diversification of the Clp1 family proteins in other eukaryotic species. Our analysis also revealed that the Clp1 family proteins in humans and plants diversified through isoforms created by alternative splicing.

Keywords Polynucleotide kinase Clp1 · Gene duplication · Family protein · RNA processing · Protein domain structure · Molecular evolution

Introduction

Cleavage factor polyribonucleotide kinase subunit 1 (Clp1) is an enzyme with polynucleotide kinase (PNK) activity, which transfers the phosphate group of ATP to the 5′ end of RNA (Weitzer and Martinez 2007). This enzyme may be involved in splicing precursor transfer RNA (pre-tRNA) (Ramirez et al. 2008; Weitzer and Martinez 2007) and the formation of the 3′-end of messenger RNA (mRNA) (de Vries et al. 2000; Minvielle-Sebastia et al. 1997). Many eukaryotic and archaeal tRNA genes contain intron(s), and precise pre-tRNA splicing reactions are required to produce functional mature tRNAs (Abelson et al. 1998; Sugahara et al. 2009, 2008). For example, consider a pre-tRNA gene with one intron in its anticodon loop region. The removal of the intron is catalyzed by the tRNA splicing endonuclease, and the ligation of the resulting two exons is either via the phosphate group at the 3′ end of the 5′ tRNA exon (the 3′-phosphate ligation pathway) or via the addition of a new phosphate group to the 5′ end of the 3′ tRNA exon (the 5′-phosphate ligation pathway) (Popow et al. 2012). In vitro experiments have shown that, in principle, Clp1 could function in a 5′-phosphate ligation pathway, adding a phosphate group to the 5′ end of the tRNA 3′ exon (Weitzer and Martinez 2007). Mice deficient in Clp1 PNK activity develop motor neuron disease, resulting from a motor neuron cell, which is attributed to the accumulation of tRNA fragments

Handling editor: **Anthony Poole**.

✉ Akio Kanai
akio@sfc.keio.ac.jp

[1] Institute for Advanced Biosciences, Keio University, Tsuruoka, Yamagata 997-0017, Japan

[2] Systems Biology Program, Graduate School of Media and Governance, Keio University, Fujisawa 252-0882, Japan

[3] Faculty of Environment and Information Studies, Keio University, Fujisawa 252-0882, Japan

caused by impaired pre-tRNA splicing (Hanada et al. 2013). A human genetic disease caused by a mutation in the *Clp1* gene has also been reported and presents as a neurological disease (Karaca et al. 2014; Schaffer et al. 2014). In contrast, the amount of mature tRNA formed was almost unchanged in mice deficient in Clp1 PNK activity (Hanada et al. 2013), so the extent to which Clp1 contributes to pre-tRNA splicing remains controversial.

Trl1/Rnl functions in multiple catalytic activities in the reactions of the 5′-phosphate ligation pathway in yeast, plants, and protists. For example, the fungus *Saccharomyces cerevisiae* protein Trl1 (*Sc*-Trl1) (Apostol et al. 1991), the plant *Arabidopsis thaliana* protein Rnl (*At*-Rnl) (Englert and Beier 2005), and the protist *Trypanosoma brucei* protein Trl1 (*Tb*-Trl1) (Lopes et al. 2016) are multidomain enzymes with an adenylyltransferase/ligase domain in the N-terminal region, a PNK domain in the central region, and a cyclic phosphodiesterase (CPDase) domain in the C-terminal region. Interestingly, *S. cerevisiae* Clp1 (*Sc*-Clp1) has no PNK activity (Ramirez et al. 2008), and *Sc*-Trl1 functions in the pre-tRNA splicing reaction (Apostol et al. 1991). The amphioxus *Branchiostoma floridae* has two separate enzymes, adenylyltransferase/ligase and PNK/CPDase, which are thought to co-operate in the 5′-phosphate ligation pathway (Englert et al. 2010). In contrast, the tRNA ligase RtcB/HSPC117 enzyme has been identified in the 3′-phosphate ligation pathway, which is widely conserved among bacteria (Tanaka et al. 2011), archaea (Englert et al. 2011), and eukaryotes (Popow et al. 2011). In terms of the other function of Clp1, in mRNA 3′-end formation, the Clp1s of *S. cerevisiae*, *Homo sapiens*, and *A. thaliana* are involved in mRNA 3′-end cleavage and as components of the polyadenylation factor complex (de Vries et al. 2000; Minvielle-Sebastia et al. 1997; Xing et al. 2008). Notably, mRNA 3′-end formation does not require the PNK activity of Clp1 and occurs even when PNK is inactivated, as in the case of *Sc*-Clp1 (Gross and Moore 2001; Ramirez et al. 2008). Therefore, Clp1 is an enzyme that mainly targets tRNAs and mRNAs.

Eukaryotic Clp1 family proteins consist of two major groups with different biological functions: the Clp1 group mentioned above and the Nol9/Grc3 group (Braglia et al. 2010). Here, the enzymes in the latter group are designated Nol9 in Metazoa and Plantae and as Grc3 in Fungi, and both the Nol9 and Grc3 proteins play important roles in precursor ribosomal RNA (pre-rRNA) processing. For example, *S. cerevisiae* Grc3 (*Sc*-Grc3) and *H. sapiens* Nol9 (*Hs*-Nol9) interact with endoribonuclease LAS1 (LAS1L in mammals) and are involved in rRNA biogenesis through the cleavage of the internal transcribed spacer 2 (ITS2) and have PNK activity against 26 rRNA in *S. cerevisiae* (Castle et al. 2013) and against 28S rRNA in *H. sapiens* (Gordon et al. 2019; Heindl and Martinez 2010). The PNK activity of *Sc*-Grc3 is also involved in transcription termination by RNA polymerase I (Braglia et al. 2010). Therefore, the Clp1 family proteins are some of the important factors that process tRNA, mRNA, and rRNA, the three typical RNAs that control genetic information.

Clp1 proteins and/or enzymes with Clp1-like PNK domains (Clp1_P) have been previously reported in eukaryotes and archaea (Jain and Shuman 2009; Noble et al. 2007). We have analyzed full-length genomes in the three domains of life and reported that the Clp1 family proteins are present in a wide range of phylogenetically diverse but restricted species of bacteria and exert PNK activity against RNA (Saito et al. 2019). Bacterial Clp1 has an essentially simple domain structure, in which the Clp1_P domain is important for its PNK activity. In archaeal Clp1, there is a conserved domain in the C-terminal region. By comparison, eukaryotic Clp1 is essentially composed of three domains, the Clp1_P domain in the central region, and N-terminal and C-terminal domains on either side of it (Noble et al. 2007). Although several conserved domains have been suggested to occur on both sides of the central Clp1_P domain in Nol9 and Grc3 (Gordon et al. 2019; Pillon et al. 2017; Saito et al. 2019), no detailed analysis of the domain structures across a wide range of lineages has been undertaken.

In light of the studies discussed above, in this study, we precisely analyzed the molecular evolution of the Clp1 enzymes in eukaryotes as a whole, to clarify how they evolved from prokaryotic Clp1, which has a simple structure, into diverse molecules in eukaryotes and how PNK-inactive forms of Clp1, such as *Sc*-Clp1, have evolved. Here, our goal was not to determine the eukaryotic phylogenetic relationships using the Clp1 family proteins, but to classify the diverse Clp1 family proteins. Therefore, we performed a large-scale molecular evolutionary analysis of the Clp1 molecular species and their protein domains using 358 full-length genomes of eukaryotic species. An analysis of PNK activity of recombinant enzymes encoded by the three *Clp1* genes of *T. brucei*, a representative species in the phylum Euglenozoa, suggested that the Clp1 group proteins, which are thought to have arisen by gene duplication, include both PNK-active and PNK-inactive forms. In the fish infraclass Teleostei and the plant phylum Magnoliophyta, which are presumed to have undergone multiple *Clp1* gene duplications associated with whole-genome duplications, the amplified *Clp1* gene has been partially modified and diversified by domain gain or deletion during evolution. In these species and some other vertebrates, including humans, the creation of protein isoforms has occurred via alternative splicing. In this study, we systematize the diversity of the eukaryotic Clp1 family proteins and, together with previous findings, provide an

overview of the evolution of the Clp1 family proteins in the three domains of life.

Results and Discussion

Comprehensive Analysis of Clp1 Family Proteins in Complete Eukaryotic Genomes

In our previous paper, we examined the distribution of the Clp1 family proteins in the complete genomes of 288 eukaryotes (Saito et al. 2019). However, no detailed analysis of the abundance of the Clp1 family proteins in each eukaryotic species or the classification of the Clp1 family protein types was performed. In this study, using known Clp1 sequences and those of its fellow family member proteins Nol9/Grc3 as query sequences (Supplementary Table S1), we performed a comprehensive sequence similarity search (E-value $\leq 1e-4$) for each coding sequence on the 358 complete genomes registered in the RefSeq database (as of September 2020). Clp1 family proteins were detected in 1264 of the 12,333,502 sequences searched, and we confirmed that they are conserved in most species of eukaryotes (350/358; 97.8%) (Supplementary Table S2a), except the protists Cryptophyta and Rhizaria. These 1264 Clp1 sequences included sequences annotated as splicing isoforms, including in Metazoa and Viridiplantae, demonstrating that the Clp1 family proteins have diversified through alternative pre-mRNA splicing in these organisms. We then calculated the number of Clp1 family proteins, excluding splicing variants, in each species as the number of independent Clp1 family genes per species at this stage. The results showed that 798 proteins were encoded in the 350 species that had at least one Clp1 family protein (2.3 ± 1.0 proteins per species on average; Supplementary Table S2a). Approximately 10% of Metazoa, approximately 40% of Viridiplantae, and 100% of Euglenozoa contained three or more Clp1 family proteins in each species (Fig. 1a). Approximately 90% of Euglenozoa species encoded four Clp1 family proteins. In contrast, in some taxa outside the Euglenozoa, Metazoa, and Amoebozoa, only one Clp1 family protein was present in each species, probably as the result of gene loss (Fig. 1a).

To classify the Clp1 family proteins present in multiple species in detail according to their sequence similarities, 254 Clp1 family proteins, including the representative sequences in Table 1, were selected from the 1264 sequences to represent the main phylogenetic groups and a molecular evolutionary phylogenetic tree was constructed from their full-length sequences. The resulting phylogenetic tree was largely divided into two groups: the Clp1 and Nol9/Grc3 group proteins (Supplementary Fig. S1). In 22 representative species (Table 1), the classification of the Clp1 family proteins showed that species such as *H. sapiens* (Metazoa), *S. cerevisiae* (Fungi), and *Babesia microti* (Alveolata) had one each of the Clp1 and Nol9/Grc3 group proteins. When a species had two or more Clp1 group proteins or Nol9/Grc3 group proteins, and each protein with unique characteristics domain formed a clade, regardless of the species, it was defined as a "type." The "types" will be described later in detail, together with the domain structures. In other cases, a, b, or c was added to the end of the protein name for convenience to distinguish multiple proteins. For example, *Asparagus officinalis* (Viridiplantae) encodes three Clp1 group proteins and three Nol9/Grc3 group proteins, so we designated the former *Ao*-Clp1-a, *Ao*-Clp1-b, and *Ao*-Clp1-c, and the latter *Ao*-Nol9-a, *Ao*-Nol9-b, and *Ao*-Nol9-c (Table 1). Among the representative species, *A. officinalis* had the largest total number of Clp1 family proteins (six). In contrast, only one Clp1 was encoded in *Cyanidioschyzon merolae* (Rhodophyta) and one Nol9 in *Thalassiosira pseudonana* (Stramenopiles). The lengths of the protein sequences sometimes differed in each species. For example, the two Clp1 group proteins of *Takifugu rubripes* (Metazoa), *Tr*-Clp1-a and *Tr*-Clp1-b, have exactly the same length [436 amino acids (aa)], whereas the two Clp1 group proteins of *Glycine max* (Viridiplantae), *Gm*-Clp1-a and *Gm*-Clp1-b, are 432 and 871 aa, respectively, an approximately two-fold difference. The regions of the protein that reflect these length differences are discussed below, together with the analysis of the protein domain structures.

Because differences in the numbers of Clp1 family proteins were observed among taxonomic groups, we mapped the number of Clp1 family proteins against the phylogenetic tree of the eukaryotes (Adl et al. 2019) to clarify when the Clp1 family proteins were acquired or lost during evolution. In Fig. 1b, the number of Clp1 family proteins is indicated as the mean ($\pm$ standard deviation) for each taxonomic group. The average number of Clp1 family proteins varied from 0 to approximately 4, depending on the eukaryotic taxon, and no clear phylogenetic increase or decrease was detected. It is generally accepted that in eukaryotes, important genes have been genetically duplicated during evolution and that the proteins they encode have also diversified (Marques et al. 2008; Rivera and Swanson 2022). However, variations in the number of Clp1 family proteins among these taxa suggest that the numbers of these proteins have increased or decreased nonphylogenetically. However, we have previously reported that a limited number of prokaryotic species (approximately 1% of bacteria and 40% of archaea) encode Clp1 family proteins (Saito et al. 2019). In that study, most species with a Clp1 family protein had only one protein (average in Bacteria 1.0, and average in Archaea 1.1). In contrast, the

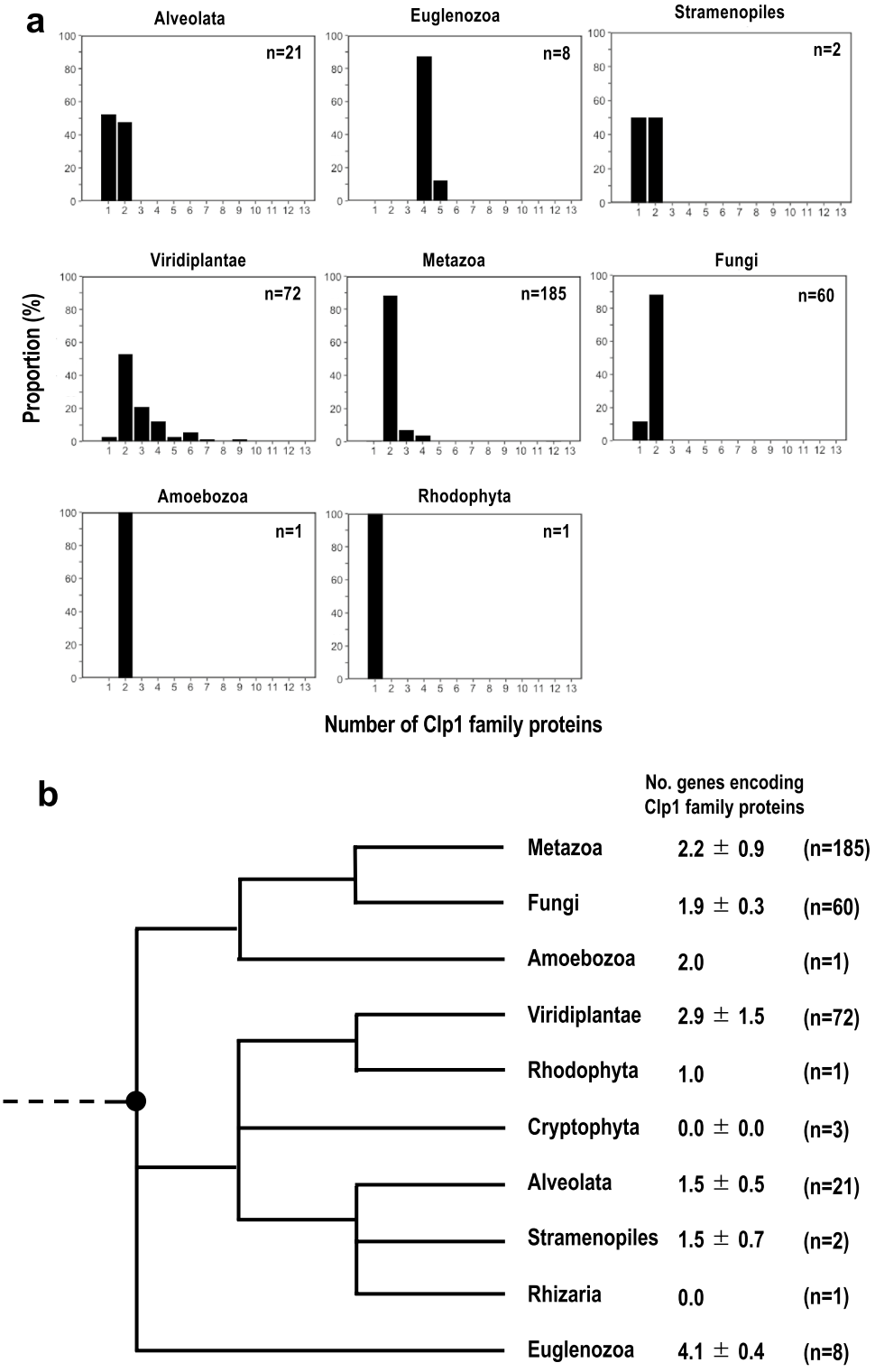


Fig. 1 Numbers of Clp1 family proteins differ among eukaryotic taxa. **a** The distribution of the number of Clp1 family proteins in each taxon is shown. Taxon names are shown above each figure (see also Table 1). *n* indicates the number of species used in the analysis. **b** Average numbers (± standard deviations) of *Clp1* genes are mapped against the evolutionary phylogenetic tree of eukaryotes of (Adl et al. 2019) (modified). Because the origin of the eukaryotes is unclear on this phylogenetic tree, it is indicated by a dotted line. The black circle indicates the possible time point at which gene duplication is presumed to have occurred

average number of Clp1 family proteins in eukaryotes was 2.3 ± 1.0, and most representative species had both Clp1 and Nol9/Grc3 proteins (Table 1). These results suggest that the gene encoding this protein was duplicated at least once in the eukaryotic ancestor (corresponding to the dotted line in Fig. 1b).

Diversification of Clp1 Group Proteins and Their Domain Structures in Eukaryotes

The evolution of the Clp1 family proteins was then analyzed at the level of their functional domains. We used two methods to estimate the conserved domain regions in the 254

Table 1 Classification of Clp1 family proteins of representative eukaryotic species

Taxon	Taxid	Species	Protein name	RefSeq ID	Gene ID	Amino acid length	Number of protein isoforms
Metazoa	9606	*Homo sapiens*	*Hs*-Clp1	NP_006822.1	10,978	425	1
			Hs-Nol9	NP_078930.4	79,707	702	5
	7227	*Drosophila melanogaster*	*Dm*-Clp1	NP_610876.1	36,494	423	0
			Dm-Nol9	NP_611084.2	36,774	995	0
	31,033	*Takifugu rubripes*	*Tr*-Clp1-a	XP_003970380.1	101,064,782	436	0
			Tr-Clp1-b	XP_029703512.1	115,246,464	436	1
			Tr-Nol9	XP_011601118.2	105,416,304	701	0
	6239	*Caenorhabditis elegans*	*Ce*-Clp1	NP_001040858.1	175,442	428	0
			Ce-Nol9	NP_001255744.1	187,190	549	0
	7957	*Carassius auratus*	*Ca*-Clp1-a	XP_026135225.1	113,113,321	289	0
			Ca-Clp1-b	XP_026080793.1	113,057,594	445	0
			Ca-Nol9-a	XP_026092660.1	113,065,544	707	0
			Ca-Nol9-b	XP_026097673.1	113,068,934	708	0
Fungi	559,292	*Saccharomyces cerevisiae* S288C	*Sc*-Clp1	NP_014893.1	854,424	445	0
			Sc-Grc3	NP_013065.1	850,624	632	0
	4896	*Schizosaccharomyces pombe*	*Sp*-Clp1	NP_593741.1	2,541,907	456	0
			Sp-Grc3	NP_588473.1	2,539,329	736	0
	1,380,566	*Pochonia chlamydosporia* 170	*Pc*-Clp1	XP_018143075.1	28,850,432	451	0
			Pc-Grc3	XP_018143592.1	28,850,812	732	0
	796,027	*Sugiyamaella lignohabitans*	*Sl*-Clp1	XP_018734659.1	30,036,306	335	0
			Sl-Grc3	XP_018734907.1	30,037,742	727	0
	367,110	*Neurospora crassa* OR74A	*Nc*-Clp1	XP_962576.2	3,878,710	379	0
			Nc-Grc3	XP_959496.1	3,875,645	804	0
	214,684	*Cryptococcus neoformans* var. *neoformans* JEC21	*Cn*-Clp1	XP_571080.1	3,257,560	548	0
			Cn-Grc3	XP_024512769.1	3,256,885	761	0
Viridiplantae	41,875	*Bathycoccus prasinos*	*Bp*-Clp1	XP_007509223.1	19,011,735	606	0
			Bp-Nol9	XP_007515247.1	19,018,065	717	0
	3702	*Arabidopsis thaliana*	*At*-Clp1-a	NP_198809.2	833,990	424	0
			At-Clp1-b	NP_187119.1	819,626	444	2
			At-Nol9	NP_001330301.1	830,968	368	3
	3847	*Glycine max*	*Gm*-Clp1-a	XP_003536785.1	100,780,061	432	0
			Gm-Clp1-b	XP_003556550.2	100,805,248	871	0
			Gm-Nol9-a	XP_003522357.1	100,818,392	370	1
			Gm-Nol9-b	XP_003525672.1	100,775,314	372	1
			Gm-Nol9-c	XP_003528167.1	100,793,271	354	0
	4686	*Asparagus officinalis*	*Ao*-Clp1-a	XP_020272637.1	109,847,808	321	0
			Ao-Clp1-b	XP_020270004.1	109,845,195	433	0
			Ao-Clp1-c	XP_020250643.1	109,828,019	494	0
			Ao-Nol9-a	XP_020246959.1	109,824,721	289	0
			Ao-Nol9-b	XP_020242603.1	109,820,820	377	0
			Ao-Nol9-c	XP_020250364.1	109,827,752	380	1
Alveolata	1,133,968	*Babesia microti* strain RI	*Bm*-Clp1	XP_012650399.1	24,426,445	494	0
			Bm-Grc3	XP_012648352.1	24,424,371	534	0
	484,906	*Babesia bovis* T2Bo	*Bb*-Clp1	XP_001610803.1	5,479,037	538	0
			Bb-Grc3	XP_001609650.1	5,477,875	607	0
Euglenozoa	435,258	*Leishmania infantum* JPCM5	*Li*-Clp1-t1	XP_001466257.1	5,069,693	445	0
			Li-Clp1-t2	XP_001467088.1	5,071,133	425	0

Table 1 (continued)

Taxon	Taxid	Species	Protein name	RefSeq ID	Gene ID	Amino acid length	Number of protein isoforms
			Li-Clp1-t3	XP_001467282.1	5,071,330	521	0
			Li-Nol9/Grc3	XP_001463261.2	5,066,689	1,368	0
	185,431	*Trypanosoma brucei* TREU927	*Tb*-Clp1-t1	XP_843821.1	3,656,162	441	0
			Tb-Clp1-t2	XP_845487.1	3,658,005	423	0
			Tb-Clp1-t3	XP_844561.1	3,656,946	512	0
			Tb-Nol9/Grc3	XP_846962.1	3,659,102	1,034	0
Rhodophyta	280,699	*Cyanidioschyzon merolae* strain 10D	*Cm*-Clp1	XP_005537677.1	16,995,802	520	0
Stramenopiles	296,543	*Thalassiosira pseudonana*	*Tp*-Nol9	XP_002290763.1	7,448,145	141	0
Amoebozoa	352,472	*Dictyostelium discoideum* AX4	*Dd*-Clp1	XP_638095.1	8,625,143	459	0
			Dd-Nol9	XP_642025.1	8,621,236	683	0

Annotation information for Clp1 family proteins in representative organisms (22 species, 57 entries) are shown: the taxonomic group of the organism (Taxon), NCBI Taxonomy ID (Taxid), NCBI Reference Sequence Database ID (RefSeq ID), NCBI gene identification ID (Gene ID), and amino acid (aa) length. For Clp1 family proteins with splicing isoforms, the one with the longest aa sequence was selected as the representative sequence. If a species had multiple Clp1 or Nol9/Grc3 group proteins, and these proteins were basically independent of the species lineage but dependent on the protein lineage and were further distinguishable by their unique domain structures, they were defined as types (e.g., t1, t2, and t3). When there were multiple Clp1 or Nol9/Grc3 group proteins that were basically dependent on the species lineage, the terms a, b, and c were used at the end of the protein name for convenience to distinguish them

representative protein sequences (Supplementary Table S2b) used in the phylogenetic analysis described above: a search against the Pfam-A domain database and a sequence similarity search. The results are shown in Fig. 2 (domain structures of 110 Clp1 group proteins) and Fig. 3 (domain structures of 144 Nol9/Grc3 group proteins), which present the overall Clp1 phylogenetic tree in two parts. In our previous study, we reported that the Clp1 group proteins are composed of three domains, the Clp1_P domain in the middle of the protein, which is essential for PNK activity, and the Clp1_eN and Clp1_eC domains in N-terminal and C-terminal regions, respectively (Saito et al. 2019). Here, we report that the N-terminal domain of *Caenorhabditis elegans* Clp1 (*Ce*-Clp1) is required for ATP binding, and the C-terminal domain of *Ce*-Clp1 may be involved in the stability of the PNK domain (Dikfidan et al. 2014). A search against the domain database showed that, except for the protists, the Clp1 proteins in the taxa Metazoa, Fungi, and Viridiplantae have three domains (Clp1_eN, Clp1_P, and Clp1_eC). However, in the protists (Euglenozoa and Alveolata), which have not been analyzed in previous studies, although the Clp1_eN and Clp1_P domains were detected, no domain was detected in the C-terminal region. Therefore, we performed a sequence similarity search (E-value of $\leq 1e-4$ and query coverage of $\geq 30\%$) with BLASTP using protein regions that were not identified in the domain search. We then confirmed the conservation of the hit sequences with an aa alignment and defined them as new domains. For example, the average length of each newly defined domain was 144.3 ± 9.7 aa for Clp1_euC1, 205.7 ± 30.4 aa for Clp1_alC, and 77.2 ± 36.2 aa for Nol9_eN3 (Figs. 2 and 3). In this way, we found that different C-terminal domains were conserved in the Euglenozoa and Alveolata Clp1 group proteins: the C-terminal domain was replaced by the Clp1_euC1, Clp1_euC2, or Clp1_euC3 domain in Euglenozoa and by the Clp1_alC domain in Alveolata (Fig. 2, see also Supplementary Tables S3–S4 and Supplementary Figs. S2–S6). Here, the three Clp1 group proteins with different C-terminal regions in Euglenozoa each formed a clade across species and were defined as types 1–3. An aa sequence alignment of all 25 Euglenozoa Clp1 group proteins showed that there were basically three proteins of each type in each of the eight Euglenozoa species (Supplementary Fig. S2 and see also Supplementary Table S3). Among these eight species, *Leishmania mexicana* contained two instances of type 3 (Supplementary Table S3). In summary, 71 of the 107 proteins (66%) used in Fig. 2 are derived from Metazoa, Fungi, and Viridiplantae, and are composed of three domains, as reported previously, indicating that these domains play important roles in a wide range of eukaryotic species. Exceptionally, a small number of proteins were found to lack some of the three domains. For instance, 27 of the 107 proteins (25%) derived from protists (Euglenozoa and Alveolata) had a structure in which only the C-terminal region was replaced by another domain.

Next, we wanted to explain the domain structure of *A. officinalis* Clp1 proteins such as *Ao*-Clp1-a, *Ao*-Clp1-b, and *Ao*-Clp1-c presented in Table 1. As shown in Fig. 2, the Clp1 group proteins are mainly composed of three domains (Clp1_eN, Clp1_P, and Clp1_eC). Although two proteins of *A. officinalis*, *Ao*-Clp1-b (433 aa) and *Ao*-Clp1-c (494 aa), have the same three-domain structure, *Ao*-Clp1-a (321

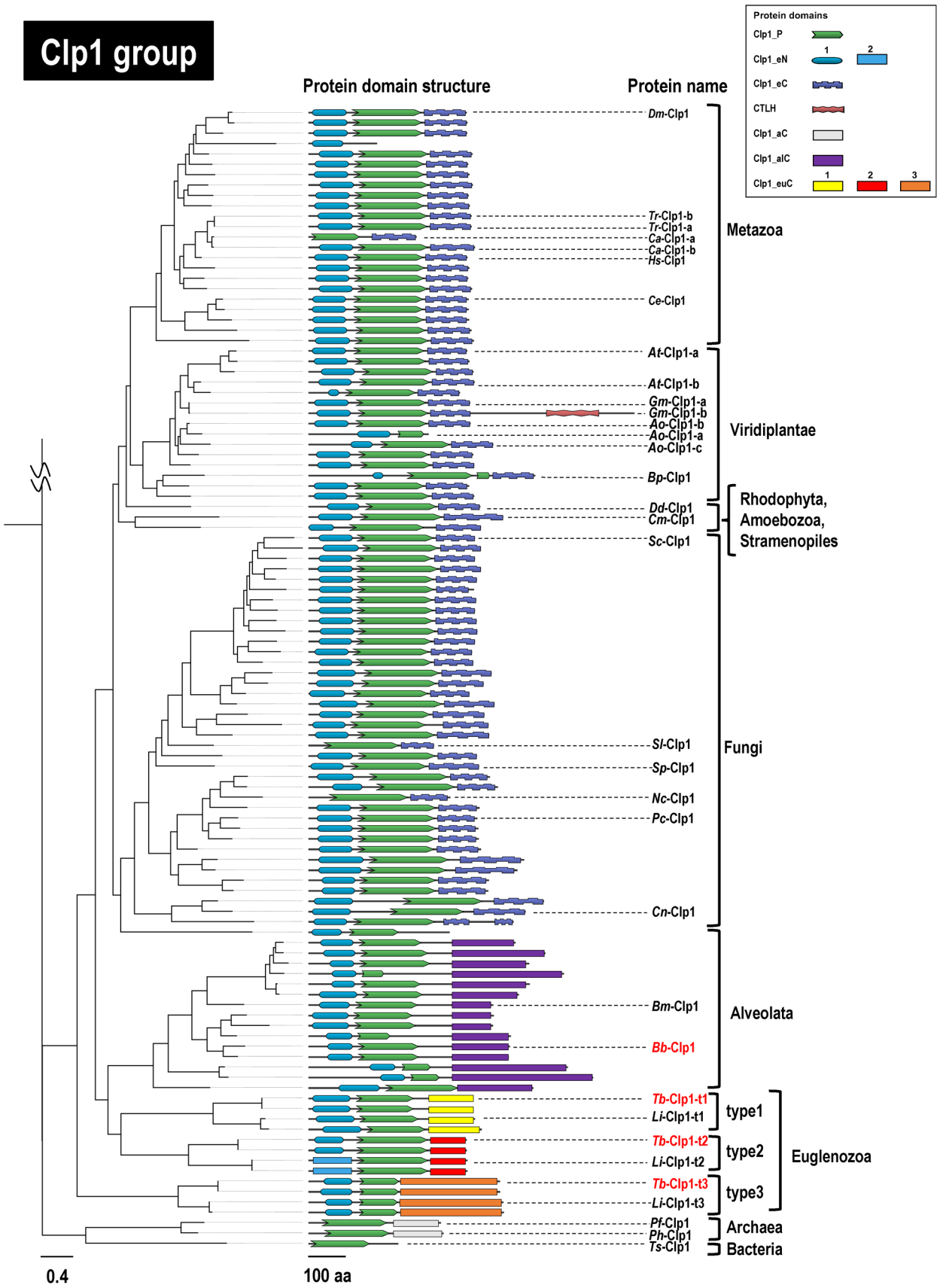
aa), which is the shortest of the three proteins, has only two domains, Clp1_eN and Clp1_P, showing that multiple proteins within a species have diversified by changing their domain structures. Similar examples, although not numerous, were also found in in *Carassius auratus* of Metazoa, *Ca*-Clp1-a (289 aa) and *Ca*-Clp1-b (445 aa) (Table 1 and Fig. 2), suggesting that multiple proteins are responsible for each specific certain function required only for the restricted species. In *G. max*, *Gm*-Clp1-a (432 aa) and *Gm*-Clp1-b (871 aa) differ approximately two-fold in their aa sequence lengths. The two proteins share the same three domains, Clp1_eN, Clp1_P, and Clp1_eC, but *Gm*-Clp1-b has a longer total length due to a sequence of approximately 400 aa containing a C-terminal LisH (CTLH) domain in the C-terminal region. The CTLH domain is found in proteins involved in functions such as microtubule dynamics, cell migration, nucleokinesis, and chromosome segregation, and is reported to play a role in protein–protein interactions (Emes and Ponting 2001).

Nol9/Grc3 Group Proteins are More Diverse in Their Domain Structures than Clp1 Group Proteins

The Nol9/Grc3 group is nominally two groups of proteins: Nol9 found in mammals and plants and Grc3 found in fungi. The *H. sapiens* Nol9 protein (*Hs*-Nol9) is reported to have N- and C-terminal domains in addition to the Clp1_P domain that is essential for PNK activity, as mentioned above (Gordon et al. 2019). In our previous study, we designated these three domains, Nol9_eN, Clp1_P, and Nol9_eC, because they are conserved among the Mammalia (Saito et al. 2019), but because several new domains were detected in the current analysis, we have numbered them systematically and designated them, for instance, Nol9_eN1 and Nol9_eC1 in this paper (Fig. 3). In considering the other protein, Grc3, we subjected the fungal *Sc*-Grc3 protein to a functional analysis of its domains. It has been reported that *Sc*-Grc3 is also composed of a Clp1_P domain (Braglia et al. 2010), an N-terminal domain involved in the survival of the yeast, and a C-terminal domain involved in the stability of endonuclease binding (Pillon et al. 2017). In this study, we again attempted to systematically detect novel domains in the N- and C-terminal regions of the Nol9/Grc3 protein group by combining the aforementioned domain database search and a similarity search. We thus identified the Clp1_P domain in all the proteins of the Nol9/Grc3 group and several proteins with unique N-terminal and C-terminal domains in each taxon (Fig. 3). However, compared with the Clp1 group, in which the proteins predominantly have three-domain structures (Clp1_eN, Clp1_P, and Clp1_eC), the Nol9/Grc3 group is more diverse, with many proteins lacking either the N- or C-terminal domain (Fig. 3 and see also Supplementary Figs. S7–S12). Fungal Grc3 proteins are basically composed of three domains, with an N- and C-terminal domain on either side of the Clp1_P domain, and there are two main groups of proteins with this domain structure, depending on the taxonomic group. One Grc3 group consists of three domains, similar to *Sc*-Grc3 of *S. cerevisiae* described above, and 17 (41%) of the 41 fungal species in Fig. 3 are classified in Saccharomycotina (Fig. 3, see also Supplementary Table S2b and Supplementary Fig. S11). The other group, in which the Grc3 proteins have N- and C-terminal domains that differ from those in the previous Grc3 group, accounted for 11 (27%) of the 41 fungal species, all of which were classified as Pezizomycotina (Fig. 3 and see also Supplementary Fig. S12). An example of this group is the *Nc*-Grc3 protein of *Neurospora crassa*. In addition to these proteins, other lineage-specific proteins in Fungi consisted solely of the Clp1_P domain or lacked either the N- or C-terminal domain, although the proportions of these proteins were low. We speculated that in these proteins, the N- and C-terminal domains were optimized for each taxon; e.g., they were specific for the substrate specificity of the enzyme or some other function. Figure 3 shows that the lengths of the proteins in the Nol9/Grc3 group varied, so we analyzed this variation in detail. First, the mean sequence length (± standard deviation) of the 144 Nol9/Grc3 group proteins used in Fig. 3 was 688 ± 296 aa. The longest was the Nol9 protein (2440 aa) of *Toxoplasma gondii* (Alveolata), which has extremely long N- and C-terminal regions, in which no domain that is conserved among other species was detected. The shortest was the Nol9 protein (141 aa) of *Thalassiosira pseudonana* (Stramenopiles), which completely lacked the N- and C-terminal domains. In 31 Nol9 proteins of Viridiplantae, the mean length of 409 aa was more than 200 aa shorter than the overall mean value due to a missing N-terminal region. In species in which this domain of Nol9/Grc3 is deleted, the function of Nol9/Grc3 may not necessarily require three domains (Gordon et al. 2019). Finally, multiple Nol9/Grc3 group proteins occurred within the same species, such as -a, -b, and -c, and the same argument can be made as for the Clp1 group proteins.

Evolution of the PNK Domain (Clp1_P Domain)

We next analyzed how the Clp1 PNK domain (Clp1_P domain) has evolved over time during eukaryotic evolution. The Clp1_eN domain was detected in the Clp1 sequence of *Nasonia vitripennis* (RefSeqID: XP_008207576.1) in the Metazoa, whereas the Clp1_P domain was not, so the sequence was excluded from the analysis. Apart from those proteins, *N. vitripennis* contains normal Clp1 and Nol9/Grc3 proteins with Clp1_P domains. First, the sequence length of the Clp1_P domain was examined. The mean length (± standard deviation) of 253 Clp1 family proteins was calculated to be 168 ± 34 aa. The largest was the 220-aa Clp1_P


Clp1 group
Protein domain structure
Protein name
Protein domains
Clp1_P
Clp1_eN
Clp1_eC
CTLH
Clp1_aC
Clp1_aIC
Clp1_euC
Dm-Clp1
Tr-Clp1-b
Tr-Clp1-a
Ca-Clp1-a
Ca-Clp1-b
Hs-Clp1
Ce-Clp1
Metazoa
At-Clp1-a
At-Clp1-b
Gm-Clp1-a
Gm-Clp1-b
Ao-Clp1-b
Ao-Clp1-a
Ao-Clp1-c
Viridiplantae
Bp-Clp1
Dd-Clp1
Cm-Clp1
Rhodophyta,
Amoebozoa,
Stramenopiles
Sc-Clp1
Sl-Clp1
Sp-Clp1
Nc-Clp1
Pc-Clp1
Cn-Clp1
Fungi
Bm-Clp1
Bb-Clp1
Alveolata
Tb-Clp1-t1
Li-Clp1-t1
type1
Tb-Clp1-t2
Li-Clp1-t2
type2
Tb-Clp1-t3
Li-Clp1-t3
type3
Euglenozoa
Pf-Clp1
Ph-Clp1
Archaea
Ts-Clp1
Bacteria
0.4
100 aa

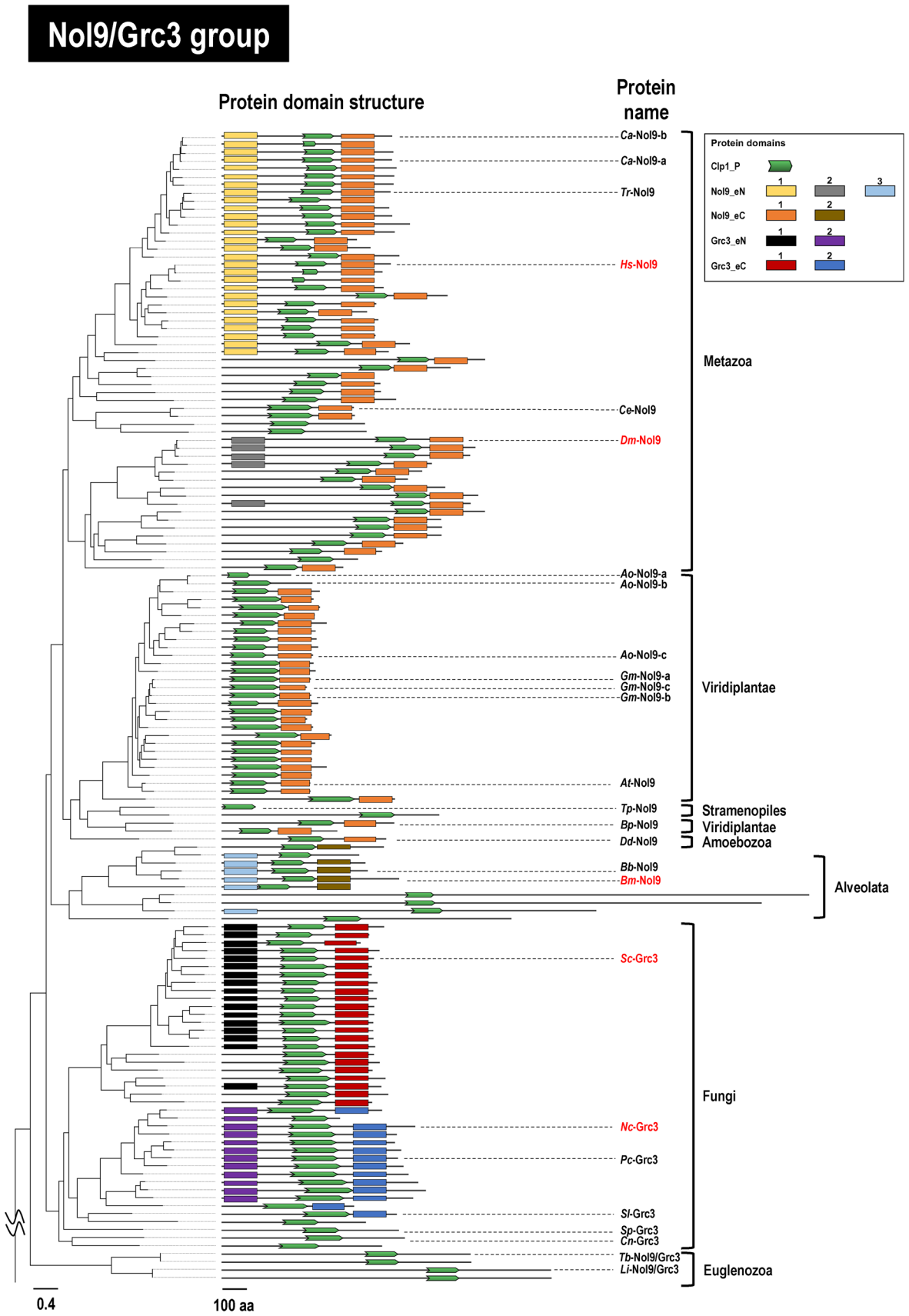
◂**Fig. 2** Phylogenetic relationships among Clp1 family proteins (Clp1 group) and their protein domain structures. A phylogenetic tree constructed from a total of 110 Clp1 group protein sequences is shown (Supplementary Table S2b). The 110 Clp1 group sequences consist of 107 protein sequences from 94 eukaryotes, two protein sequences from two archaea, and one protein sequence from a bacterium. Each full-length amino acid sequence of the Clp1 group proteins was used for the phylogenetic analysis, and midpoint rooting was applied during tree visualization. The LG + F + R8 model was used for the phylogenetic tree. The entire molecular evolutionary phylogenetic tree of the Clp1 family proteins is shown in Supplementary Fig. S1. The scale bar under the tree indicates the number of amino acid substitutions per site. Symbol for protein domain structures are summarized in the box. Protein domains were identified using two methods: (i) searches against the Pfam database and (ii) manual amino acid sequence alignments (all symbols are rectangles). Protein names of the query sequences used to detect the protein domain structures are indicated in red letters (see Supplementary Figs. S3–S6). See also Supplementary Table S4 for details of domain structures

domain of the Clp1 protein (XP_571080.1, 548 aa) of *Cryptococcus neoformans*, which is classified in Fungi, and the smallest was the 55 aa Clp1_P domain of the Nol9 protein (XP_018921794.1, 601 aa) of *Cyprinus carpio*, which is classified in the Metazoa. When we examined the four motifs (Walker A, Walker B, Clasp, and Lid) (Dikfidan et al. 2014) in the Clp1_P domain of *C. carpio* that are important for PNK activity, we found that the sequences corresponding to Clasp and Lid were completely absent. This was also true of 16 proteins in which the Clp1_P domain was shorter than 130 aa (approximately 6% of the total proteins), suggesting that species with Clp1_P domains significantly smaller than the mean length lack PNK activity.

Surprisingly, a phylogenetic tree constructed from the aa sequences of the Clp1_P domain allowed the Clp1 group proteins and the Nol9/Grc3 group proteins to be distinguished simply by the sequence of the Clp1_P domain alone (Supplementary Fig. S13). Because a previous study reported that *Hs*-Clp1 of the Clp1 group has PNK activity during pre-tRNA splicing (Ramirez et al. 2008; Weitzer and Martinez 2007) and *Hs*-Nol9 of the Nol9/Grc3 group has PNK activity during pre-rRNA processing (Gordon et al. 2019; Heindl and Martinez 2010), it is possible that the aa sequence of the Clp1_P domain and its substrate specificity for pre-tRNA or pre-rRNA have co-evolved. Moreover, on the phylogenetic tree based on the full-length sequences of the Clp1 family proteins (Supplementary Fig. S1), the taxa in both the Clp1 group and the Nol9/Grc3 group predominantly clustered together, but on the phylogenetic tree based on the Clp1_P domain (Supplementary Fig. S13), some of the same taxa were divided into a few phylogenetically distant positions. For example, two Fungi and three Euglenozoa species independently localized to the Clp1 group on the tree. Similarly, in the Nol9/Grc3 group, three Fungi and three Metazoa species also localized independently on the tree. In particular, the phylogenetic position of Euglenozoa type 1 Clp1 (*Tb*-Clp1-tl) was close to the fungal enzyme (*Sc*-Clp1), whereas the other two Euglenozoa proteins, *Tb*-Clp1-t2 and *Tb*-Clp1-t3, were not. Therefore, to test branch reliability, we calculated the ultrafast bootstrap approximation (UFBoot2) values for the phylogenetic tree in Supplementary Fig. S13. The results showed that the reliability of the Euglenozoa type 1 Clp1 (*Tb*-Clp1-tl) branches, which are closely related to the fungal enzyme (*Sc*-Clp1), was as low as 44%. In contrast, when we compared the amino acid sequences of the active Clp1 proteins (e.g., *Hs*-Clp1, *Ce*-Clp1, and *Tb*-Clp1-t2) and inactive Clp1 proteins (e.g., *Sc*-Clp1 and *Tb*-Clp1-t1), which were separated from each other on the phylogenetic tree based on the Clp1_P domain region, we found that amino acid mutations occurred in the motifs (especially the Clasp motif) that are important for PNK activity in the inactive Clp1 proteins. However, multiple amino acid residues were also conserved in the inactive Clp1 proteins (Supplementary Fig. S14). These results suggest that a phylogenetic tree based on the Clp1_P domain is unsuitable for classifying species but is useful for classifying similar structural enzymes, and that although the phylogenetic relationship between *Tb*-Clp1-tl and *Sc*-Clp1 is unclear, there are some amino acid sequence similarities between the two. Moreover, using the full-length regions of *T. brucei* Clp1 types 1–3 from Euglenozoa, we analyzed their aa sequence similarities in a round-robin fashion and found only low values of aa identity (24–34%; Table 2). These results suggest that the Clp1_P domain of each type of Euglenozoa Clp1 has evolved independently and that conserved regions outside the Clp1_P domain are more similar to each other than to conserved regions outside the Clp1_P domain in other Clp1 proteins. Furthermore, the prokaryotic Clp1 protein localized at the base of the Nol9/Grc3 group, suggesting that prokaryotic Clp1 probably originated as a regulator of pre-rRNA. Therefore, the function of prokaryotic Clp1 in pre-rRNA processing must be investigated in future studies.

Biochemical Characterization of the PNK Activities of Three Types of Clp1 Proteins in *T. brucei*

We have described Clp1 types 1–3 in Euglenozoa using a sequence analysis, but their PNK activities must be experimentally verified. To detect and compare the PNK activities of the three Clp1 group proteins of Euglenozoa *T. brucei*, three His-tagged recombinant proteins (*Tb*-Clp1-t1, *Tb*-Clp1-t2, and *Tb*-Clp1-t3) were expressed in *Escherichia coli* and partially purified using TALON® Metal (Cobalt) Affinity Chromatography. The calculated molecular weights of these His-tagged recombinant proteins were 48.6 kDa for *Tb*-Clp1-t1, 47.1 kDa for *Tb*-Clp1-t2, and 56.7 kDa for *Tb*-Clp1-t3, and each size of protein detected by the anti-His-tag antibody was consistent with the size predicted from


Nol9/Grc3 group
Protein domain structure
Protein name
Protein domains
Clp1_P
Nol9_eN
Nol9_eC
Grc3_eN
Grc3_eC
Ca-Nol9-b
Ca-Nol9-a
Tr-Nol9
Hs-Nol9
Ce-Nol9
Dm-Nol9
Ao-Nol9-a
Ao-Nol9-b
Ao-Nol9-c
Gm-Nol9-a
Gm-Nol9-c
Gm-Nol9-b
At-Nol9
Tp-Nol9
Bp-Nol9
Dd-Nol9
Bb-Nol9
Bm-Nol9
Sc-Grc3
Nc-Grc3
Pc-Grc3
Sl-Grc3
Sp-Grc3
Cn-Grc3
Tb-Nol9/Grc3
Li-Nol9/Grc3
Metazoa
Viridiplantae
Stramenopiles
Viridiplantae
Amoebozoa
Alveolata
Fungi
Euglenozoa
0.4
100 aa

◂**Fig. 3** Phylogenetic relationships among Clp1 family proteins (Nol9/Grc3 group) and their protein domain structures. A phylogenetic tree constructed from a total of 144 Nol9/Grc3 group protein sequences from 133 eukaryotes is shown (Supplementary Table S2b). Each full-length amino acid sequence of the Nol9/Grc3 group proteins was used for the phylogenetic analysis. Protein names of the query sequences used to detect protein domain structures are indicated in red letters (see Supplementary Figs. S7–S12). The location of the root of the phylogenetic tree is shown in Fig. 2. See also the legend to Fig. 2 for other details

aa sequence deduced from the corresponding genes (Fig. 4 and see also Supplementary Fig. S15a–S15c). For example, for *Tb*-Clp1-t2 (47.1 kDa), a partially purified main band was observed slightly below the 50 kDa molecular weight marker (Fig. 4a). In a western blot analysis with an anti-His-tag antibody, this band also has cross-reactivity and peaks in elution fractions 9–12. These observations were also true for recombinant proteins *Tb*-Clp1-t1 and *Tb*-Clp1-t3 (Figs. 4b–c and see also Supplementary Fig. S15b–S15c). These results shows that each of the three recombinant Clp1 group proteins of *T. brucei* was partly purified to a major component on sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS–PAGE) using this method. The elution fractions of Clp1 types 1–3 were then examined for PNK activity against single-stranded RNA (ssRNA). PNK activity was only detected in *Tb*-Clp1-t2, and the peak of activity was perfectly consistent with the peak of the *Tb*-Clp1-t2 protein on SDS–PAGE (Fig. 4a). In contrast, no clear activity was detected for the remaining two proteins, *Tb*-Clp1-t1 and *Tb*-Clp1-t3 (Figs. 4b & c). These results clearly indicate that among types 1–3 of the Clp1 group proteins of *T. brucei*, only the recombinant *Tb*-Clp1-t2 protein had PNK activity. A similar experiment was performed with recombinant *Leishmania infantum* protein *Li*-Clp1-t2, which also belongs to type 2; however, no PNK activity was detected in this recombinant protein (Supplementary Fig. S15d). The consensus sequences of four motifs (Walker A, Walker B, Clasp, and Lid) important for PNK activity (Dikfidan et al. 2014) were examined according to the Clp1 type, and all four motifs were completely conserved in *Tb*-Clp1-t2 (Supplementary Fig. S2). All four motifs were also completely conserved in *Tbr*-Clp1-t2 of *T. brucei gambiense*, which is classified in the closely related order Trypanosomatida. In contrast, at least part of the motif sequence in *Li*-Clp1-t2 of *L. infantum* and *Lp*-Clp1-t2 of *Leishmania panamensis*, which are also type 2 Clp1, did not match the consensus sequence (Supplementary Fig. S2).

We next examined the substrate specificity of *Tb*-Clp1-t2 and compared it with those of prokaryotic archaea *Pyrococcus furiosus* Clp1 (*Pf*-Clp1) and bacteria *Thermus scotoductus* Clp1 (*Ts*-Clp1) (Figs. 4d–f). In a previous study, we reported that *Pf*-Clp1 and *Ts*-Clp1 have a Clp1_P domain, and that the four motifs cited above are highly conserved within this domain. Bacterial *Ts*-Clp1 was used for comparison in this study because the recombinant protein expressed in *E. coli* was found to have PNK activity against DNA and RNA oligonucleotides (Saito et al. 2019). Among the Archaea, *Ph*-Clp1 of *Pyrococcus horikoshii*, in the same genus as *P. furiosus*, has been characterized (Jain and Shuman 2009), and *Pf*-Clp1 and *Ph*-Clp1 are homologous proteins. To compare the reaction conditions at temperatures similar to those in vivo for each species, the experiments were conducted at 75 °C for archaeal *Pf*-Clp1, 60 °C for bacterial *Ts*-Clp1, and 27 °C for protist *Tb*-Clp1-t2 (Fig. 4f), and reaction time was 15 min for prokaryote Clp1 and 60 min for protist Clp1, so that each change in enzyme activity with the amount of protein used was clear at each temperature. First, prokaryotic *Ts*-Clp1 and *Pf*-Clp1 showed PNK activity against ssRNA and double-stranded RNA (dsRNA) at 5–20 ng per reaction (Figs. 4d and e). In comparison, PNK activity against ssDNA and dsDNA required about 30 times the amount of enzyme (about 150–300 ng/reaction) of that against ssRNA and dsRNA (Figs. 4d and e). In a previous study, archaeal *Ph*-Clp1 also showed PNK activity against DNA and RNA and was reported to have stronger activity against RNA than DNA (Jain and Shuman 2009), and we confirmed that the experimental results for *Pf*-Clp1 were consistent with the previous report of *Ph*-Clp1. Here, protist *Tb*-Clp1-t2 showed PNK activity against ssRNA and dsRNA at about 50–300 ng/reaction, so it required about ten times more enzyme and a longer reaction time than prokaryotic *Ts*-Clp1 or *Pf*-Clp1 (Fig. 4f). Prokaryotic *Ts*-Clp1 and *Pf*-Clp1 also showed PNK activity against ssDNA and dsDNA when a larger amount of enzyme (approximately 150–300 ng/reaction) was used than against RNA (Figs. 4d and e), but no activity was detected for *Tb*-Clp1-t2, even when a larger amount was used (300 ng/reaction; Fig. 4f). In summary, purified prokaryotic *Ts*-Clp1 and *Pf*-Clp1 showed PNK activity against ssRNA, dsRNA, ssDNA, and dsDNA, although more enzyme was required for the ssDNA and dsDNA substrates. In contrast, purified protist *Tb*-Clp1-t2 was only active against RNA substrates; i.e., ssRNA and dsRNA. These results suggest that the Clp1 group acquired substrate specificity for ssRNA and dsRNA or lost substrate specificity for ssDNA and dsDNA during the early formation of the Clp1 group during the evolution of the protists. Furthermore, compared with prokaryotic Clp1s with their simple domain structures, protist *Tb*-Clp1-t2 and other eukaryotic Clp1s acquired specific domains at their N- and C-termini, respectively (Fig. 2). These domains may be involved in the preferential recognition of ssRNA and dsRNA. In previous studies, *Hs*-Clp1 and *Ce*-Clp1 in the Clp1 group showed PNK activity against ssRNA and dsRNA, but not against ssDNA. However,

Table 2 Summary of Clp1 family proteins in *T. brucei*

Protein				Identity (similarity) among family proteins				Chromosomal location
Name	RefSeq ID	Number of amino acid	Molecular weight (kDa)	*Tb*-Clp1-t1	*Tb*-Clp1-t2	*Tb*-Clp1-t3	*Tb*-Nol9/Grc3	
Tb-Clp1-t1	XP_843821.1	441	47.5	–	27 (46)	24 (41)	N/D	#3
Tb-Clp1-t2	XP_845487.1	423	46.0	34 (55)	–	N/D	30 (45)	#6
Tb-Clp1-t3	XP_844561.1	512	55.6	24 (42)	N/D	–	N/D	#4
Tb-Nol9/Grc3	XP_846962.1	1,034	112.0	N/D	30 (45)	N/D	–	#8

– Indicates a 100% match between the amino acid sequence of the query sequence and the amino acid sequence of the target; N/D, not detected

unlike protist *Tb*-Clp1-t2, these enzymes displayed PNK activity against dsDNA (Dikfidan et al. 2014; Weitzer and Martinez 2007). At present, the weaker PNK activities of prokaryotes Clp1s against dsDNA are presumed to have been lost and reacquired in the process of eukaryotic evolution. Therefore, the biological implications of the PNK activity of Clp1 against DNA in nematodes and humans, as well as in prokaryotes, must be verified in future studies.

Finally, we would like to discuss the function of *Tb*-Clp1-t2, the active and early form of eukaryotic Clp1, in vivo. *Trypanosoma brucei* has one tyrosine tRNA containing an intron and requires a precise pre-tRNA splicing reaction, catalyzed by either an RtcB-dependent pathway that uses the 3′-terminal phosphates of tRNA exons or a pathway that uses a 5′-terminal phosphate (possibly provided by Clp1) of the tRNA exon (Yoshihisa 2014). In the 5′-phosphate pathway, a tRNA ligase called Trl1 is present in yeast in which the PNK of Clp1 is inactivated (Sawaya et al. 2003; Wang and Shuman 2005). Furthermore, a tRNA ligase called Rnl is present in plants together with a Clp1-similar protein (CLPS3), whose PNK activity has not yet been clarified (Englert and Beier 2005; Xing et al. 2008). Interestingly, a previous study reported that a protein homologous to yeast Trl1 exists in *T. brucei* and functions in joining tRNA exons (Lopes et al. 2016). Therefore, *Tb*-Clp1-t2, whose PNK activity was detected in this study, may function as a backup for Trl1/Rnl or may provide the PNK activity required for RNA repair (Zhang et al. 2012). It should be noted that a previous study suggested that *Tb*-Clp1-t2 is negligibly involved in polyadenylation (Koch et al. 2016).

Remarkably, the protist *T. brucei* is the only representative eukaryote that encodes all of the multiple factors directly involved in joining tRNA exons, such as RtcB, Trl1/Rnl, and Clp1, although the number of its tRNAs that contain introns is very small (Supplementary Table S5). On the contrary, Metazoa, Fungi, and Viridiplantae, which have more intron-containing tRNAs than protists, lack either RtcB or Trl1/Rnl among the RNA ligase family proteins, although Clp1 is present in each taxon. We presume that an evolutionary event selected the optimal enzyme from Trl1/Rnl and RtcB, depending on the situation of each taxon, which may have included an increased number of intron-containing tRNAs or the inactivation of Clp1.

Summary of Diverse Clp1 Family Proteins in the Three Domains of Life

Figure 5 summarizes the diversity and relationships of the Clp1 family proteins in the three domains of life, using the main representative species used in this study. When Clp1 originally emerged in prokaryotes, there was basically one Clp1 family protein in each species, with a simple domain structure consisting mainly of the Clp1_P domain and loose substrate specificity. This protein played a role via its PNK activity in the reactions of various RNA and DNA substrates (Figs. 2 and 4). In addition to the Clp1_P domain, a C-terminal domain was present in archaeal Clp1 (Saito et al. 2019). In contrast, in the eukaryotic ancestor, duplication of the *Clp1* gene formed the Clp1 and Nol9/Grc3 groups, increasing the number of Clp1 family proteins. In the present study, we found that the Euglenozoa Clp1 family proteins, diversified by gene duplication, giving rise to approximately three Clp1 group proteins (e.g., *Tb*-Clp1-t1, *Tb*-Clp1-t2, and *Tb*-Clp1-t3) and one Nol9/Grc3 group protein (e.g., *Tb*-Nol9/Grc3) in each species. We detected no PNK activity in *Tb*-Clp1-t1 and *Tb*-Clp1-t3, at least under the conditions used (Figs. 4b and c). *Sc*-Clp1 has no PNK activity (Ramirez et al. 2008), and its position on the phylogenetic tree based on the Clp1_P domain (Supplementary Fig. 13) and conserved relationships of amino acid sequences in inactive enzymes (Supplementary Fig. 14) suggest that it is structurally close to PNK-inactive *Tb*-Clp1-t1. Because *Sc*-Clp1 is involved in pre-mRNA 3′-end formation, which does not require PNK activity (Ghazy et al. 2012; Noble et al. 2007), *Tb*-Clp1-t1 is thought to have a similar function (Clayton and Michaeli 2011). No protein phylogenetically close to *Tb*-Clp1-t3 was found, suggesting that the *Tb*-Clp1-t3 lineage was not acquired by any other organism during the course of evolution or was lost after its acquisition. *Tb*-Clp1-t2, the only PNK active against RNA substrates, was possibly the

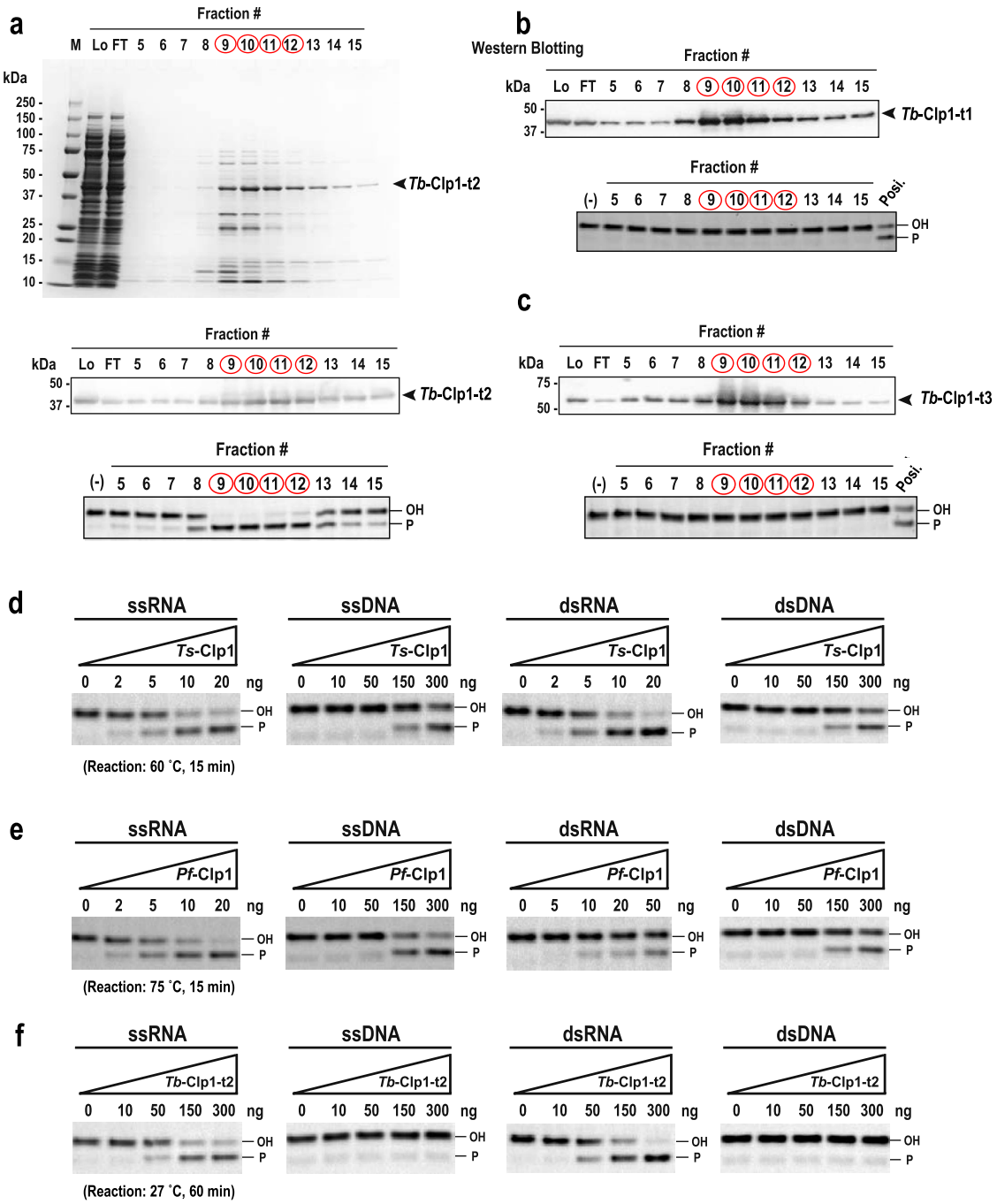


Fig. 4 Biochemical characterization of the recombinant Clp1 group proteins of *T. brucei*. **a** Purification of recombinant *Tb*-Clp1-t2 protein and its PNK activity. SDS–PAGE (10–20%) analysis of the purification of recombinant *Tb*-Clp1-t2 protein using TALON Metal (Cobalt) Affinity Chromatography stained with Coomassie Brilliant Blue (top). Western blot analysis with an anti-His-tag antibody (middle). PNK activity against single-stranded RNA (ssRNA) as the substrate (bottom). Fractions of the protein peak are indicated by red circles. **b**, **c** Purification of the recombinant *Tb*-Clp1-t1 and *Tb*-Clp1-t3 proteins (western blotting analysis, top) and their PNK activities (bottom). Arrows indicate the positions of each recombinant protein. Recombinant *Tb*-Clp1-t2 protein was used as the positive control. **d**–**f** Comparison of PNK activities of three Clp1 proteins: **d** bacterial *Thermus scotoductus* Clp1 (*Ts*-Clp1), **e** archaeal *Pyrococcus furiosus* Clp1 (*Pf*-Clp1), and **f** eukaryotic *Tb*-Clp1-t2 on four substrates (ssRNA, ssDNA, double-stranded RNA [dsRNA], and dsDNA) (Color figure online)

earliest member of the eukaryotic Clp1 group. As mentioned above, the prokaryotic Clp1 family proteins, which appeared early in the evolution of life are active against both RNA and DNA substrates, whereas *Hs*-Clp1 and *Ce*-Clp1 in Metaozoa

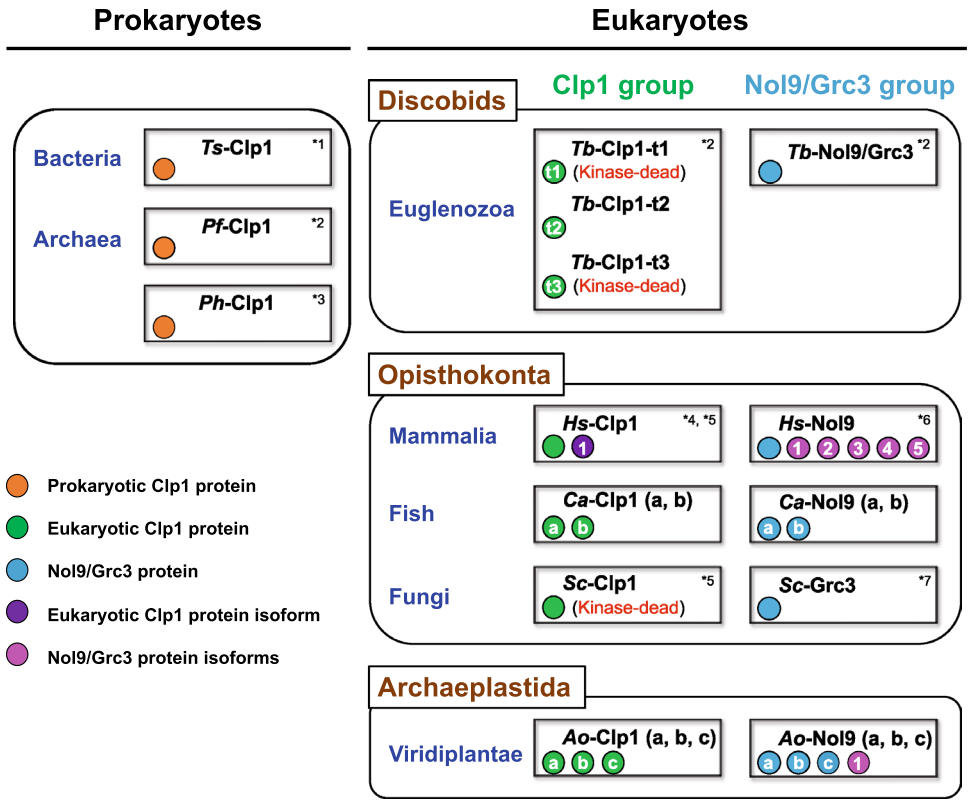


Fig. 5 Summary of diverse Clp1 family proteins in the three domains of life. A schematic illustration of the Clp1 family proteins of representative species in prokaryotes (Bacteria and Archaea) and the three major eukaryotic taxonomic groups (Discobids, Opisthokonta, and Archaeplastida) (Gabaldon 2021). The color of each circle corresponds to the protein category, as indicated in the figure. t1–t3 inside the circle indicates protein types 1–3 of *Tb*-Clp1, and a–c indicates distinct individual proteins. The number of protein isoforms is described as 1–5. **References**: *1 (Saito et al. 2019); *2, this study; *3 (Jain and Shuman 2009); *4 (Weitzer and Martinez 2007); *5 (Ramirez et al. 2008); *6 (Heindl and Martinez 2010); *7 (Braglia et al. 2010). See text for details (Color figure online)

have similar substrate specificities to that of *Tb*-Clp1-t2 in that they are mainly active against RNA substrates. Taken together, these data suggest that the PNK activity of the Clp1 group changed in the eukaryotic ancestor to primarily target the RNA strand. However, unlike *Tb*-Clp1-t2, both *Hs*-Clp1 and *Ce*-Clp1 have PNK activity against dsDNA, although this point is controversial. Therefore, in the Clp1 group, the aa sequences of the Clp1_P domain and the C-terminal domain of Euglenozoa types 1–3 changed to generate proteins with or without PNK activity or with different functions.

The Clp1 group is basically composed of three domains: Clp1_eN1, Clp1_P, and Clp1_eC, but in Alveolata and Euglenozoa, in addition to the Clp1_eN1, Clp1_eN2, and Clp1_P domains, the C-terminal domain was replaced by the Clp1_alC or Clp1_euC1-euC3 domain in a lineage-specific manner (Fig. 2). Similarly, in the Nol9/Grc3 group proteins, the N-terminal and C-terminal domains in some taxa were replaced with unique domains, whereas the Clp1_P domain was usually conserved. Moreover, in Alveolata and Viridiplantae, the Nol9/Grc3 group proteins diversified with changes to their sequence lengths (possibly by including unidentified functional domains; Fig. 3). Interestingly, in *H. sapiens*, a vertebrate, the total number of Clp1 family proteins was reduced to two, although two whole-genome duplications are estimated to have occurred (Singh and Isambert 2020). Simultaneously, higher vertebrates produced diverse protein isoforms by alternative splicing (Keren et al. 2010). For example, in *H. sapiens*, there is one protein isoform in the Clp1 group and five in the Nol9/Grc3 group (Fig. 5 and Table 1). In the fishes of infra-class Teleostei (Berthelot et al. 2014) and the plant phylum Magnoliophyta (Van de Peer et al. 2017), which both include species with high frequencies of whole-genome duplication, diversity was acquired by the partial modification of the amplified genes and the appearance of protein isoforms, as described above. For example, genes encoding *Ca*-Clp1-a and *Ca*-Clp1-b appeared in the fish *C. auratus*, and genes encoding *Ao*-Nol9-a, *Ao*-Nol9-b, and *Ao*-Nol9-c in the plant *A. officinalis*. These Clp1s are either identical to each other in domain structure or lack some domains (Figs. 2 and 3).

Materials and Methods

Dataset

To search for Clp1 family proteins in complete genomes, we obtained 358 GenBank files (Supplementary Table S2) annotated as "Complete genome" or "Chromosome" from the Reference Sequence (RefSeq) database (O'Leary et al. 2016) at ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/; last accessed 8 August 2020. Twenty-two representative species were also selected for a phylogenetic analysis to ensure

the inclusion of all taxa of Clp1 family proteins identified in this study (Table 1). For the domain analysis, 18,259 families, together with their reliable annotations, were obtained from the Pfam-A database (version 33.1) (Mistry et al. 2021) at ftp://ftp.ebi.ac.uk/pub/databases; last accessed 8 August 2020.

Large-Scale Identification of Clp1 Family Proteins

To identify Clp1 family proteins in the RefSeq database, a protein–protein BLAST (BLASTP, ver. 2.4.0 +) search (Camacho et al. 2009) using the BLOSUM62 matrix (Henikoff and Henikoff 1992) was performed with an E-value of ≤ 1e − 4 and query coverage of ≥ 30%. As query sequences, we used the aa sequences of eukaryotic Clp1 and its family proteins Nol9 and Grc3, which have already been reported in previous studies (Supplementary Table S1). We used both the full-length and PNK domain regions of each query sequence to comprehensively identify Clp1 family proteins, and each Clp1 family protein set obtained was integrated without duplication (1,264 sequences in total; Supplementary Table S2a). The representative Clp1 family protein set (254 sequences in total; Supplementary Table S2b) was collected as follows. First, 1,264 Clp1 family protein sequences were clustered with 70% sequence similarity using CD-HIT (ver. 4.8.1, March 2019) (Fu et al. 2012) to reduce the number of similar proteins, and sequences were randomly collected from each cluster. We then excluded all protein sequences annotated as splicing isoforms and added the sequences of representative organisms (Table 1) and the sequences of prokaryotes for comparative analysis.

Amino Acid Sequence Alignment Analysis

The aa sequences of the Clp1 family proteins were aligned using MAFFT L-INS-i ver. 7.394 or 7.471 with the default parameters (Katoh and Standley 2013). The results were visualized using Jalview (ver. 2.11.1.4) (Waterhouse et al. 2009). The colors and scores for alignment conservation are based on the following definitions. Identical aa residues were indicated in blue and partly conserved aa residues in light blue. Gaps (–) were inserted to maximize the number of aa matches. The conservation score for each aa position was indicated as one of 12 ranks (0–11). Partially conserved aa (rank 10) were shown by a plus (+) symbol, and identical aa (rank 11) were indicated by asterisks (*) (Livingstone and Barton 1993). The alignment files created in this step were also used for the phylogenetic tree analysis described in "Molecular phylogenetic analysis" below.

The protein similarity scores among the *T. brucei* Clp1 proteins were calculated using the protein–protein BLAST analysis described in "Large-scale identification of Clp1 family proteins." Here, "amino acid identity" was defined as the percentage of identical aa residues in two different sequences, and "amino acid similarity" was defined as the ratio of the number of aa that were identical or chemically similar between the query and target sequences (Table 2).

Molecular Phylogenetic Analysis

For the phylogenetic analysis, we used the above-mentioned alignment file and removed any gaps from it using TrimAL (1.2rev59, 2009) (Capella-Gutierrez et al. 2009). Phylogenetic trees were constructed using the best-fit model, as determined using ModelFinder of IQ-TREE (ver. 2.1.10) (Minh et al. 2020), which is excellent for maximum likelihood analyses of large-scale data. An ultra-fast bootstrap approximation (UFBoot2) analysis (Hoang et al. 2018), which is suitable for processing large numbers of sequences, was performed 1000 times. Because we were unable to root the tree using the appropriate outgroup, we used midpoint rooting (MPR) as an alternative method (Farris 1972; Hess and de Moraes Russo 2007). MPR is a method that places the root in the middle of the two longest branches and was implemented in FigTree (ver. 1.4.4; http://tree.bio.ed.ac.uk/software/figtree/) to visualize the rooted phylogenetic tree.

Domain Analysis of Clp1 Family Proteins

The domain structures of the Clp1 family proteins were estimated using the following two methods: (1) Using the entire aa sequence of each Clp1 family protein as the query, the Pfam-A database (Mistry et al. 2021), in which known domains are registered, was searched using HMMER (ver. 3.2) (Mistry et al. 2013), set to an E-value of ≤ 1e − 3. Their domain structures were visualized using DoMosaics (ver. Rv0.95) (Moore et al. 2014), and (2) for protein regions that were not hit in the domain search, a sequence similarity search using BLASTP was performed (E-value of ≤ 1e − 4 and query coverage of ≥ 30%). The conservation of the obtained sequences was confirmed using an aa alignment (Supplementary Table S3), and they were defined as novel domains (Supplementary Table S4).

Synthesis of Artificial Genes for Euglenozoan Clp1 Proteins and Construction of Expression Vectors

Four euglenozoan *Clp1* genes (encoding *Tb*-Clp1-t1, *Tb*-Clp1-t2, *Tb*-Clp1-t3, and *Li*-Clp1-t2) (Supplementary Table S6) were used to express recombinant proteins in *E. coli*. First, using a WEB tool (https://www.eurofinsgenomics.jp/jp/service/gsy/orderstart.aspx?type=MyCart) provided by Eurofins Genomics Tokyo, we optimized the nucleotide sequences to match the codon usage of *E. coli*, and synthesized each artificial *Clp1* gene. These synthetic genes were designed to contain *Nde*I and *Xho*I sites at their

5′- and 3′-termini, respectively, and were subcloned into these restriction sites in the pET-23b expression vector (Novagen, Madison, WI, USA). The resulting pET-Clp1 vectors encoded each Clp1 protein with a six-histidine (His) tag at its C-terminal end (Supplementary Table S6).

Expression and Purification of His-tagged Recombinant Clp1 Proteins

To express the recombinant euglenozoan Clp1 proteins (*Tb*-Clp1-t1, *Tb*-Clp1-t2, *Tb*-Clp1-t3, and *Li*-Clp1-t2), *E. coli* strain BL21(DE3) was transformed with each expression vector containing the artificial euglenozoan *Clp1* genes (Supplementary Table S6). The transformants were precultured in Luria–Bertani (LB) medium containing 50 µg/mL ampicillin for 4 h at 37 °C. Each sample was transferred to 350 mL of LB medium containing the same concentration of ampicillin, incubated at 30 °C for 4 h, and then at 17 °C for 2 h. Isopropyl β-D-1-thiogalactopyranoside (IPTG, 0.4 mM) was added and the cells incubated at 17 °C for a further 14 h to overexpress the desired recombinant Clp1 protein. The cells were harvested by centrifugation (9000 × g for 3 min at 4 °C), and the protein was extracted using sonication (3–4 min) in His-tag-binding buffer containing 20 mM Tris–HCl (pH 8.0), 500 mM NaCl, 5 mM imidazole, and 0.1% (v/v) NP-40. The insoluble proteins were removed by centrifugation (18,000 × g for 10 min at 4 °C). The recombinant proteins were purified using a TALON Metal (Cobalt) Affinity Resin column (Clontech Laboratories, Inc., Palo Alto, CA, USA) and eluted with a linear gradient of imidazole (0–1000 mM) in His-tag-binding buffer using the ÄKTA FPLC™ Fast Protein Liquid Chromatography system (GE Healthcare, Princeton, NJ, USA). The eluted protein peak was collected and dialyzed against buffer D containing 50 mM Tris–HCl (pH 8.0), 1 mM ethylenediaminetetraacetic acid (EDTA), 0.02% (v/v) Tween 20, 7 mM 2-mercaptoethanol, and 10% (v/v) glycerol.

The expression and purification of recombinant *Ts*-Clp1 (UniProt AC: E8PQM6) were according to our previous report (Saito et al. 2019). Recombinant *Pf*-Clp1 (UniProt AC: Q8U4H6, 354 aa), a homolog of *Ph*-Clp1 (UniProt AC: O57936, 361 aa), in which PNK activity has been verified (Jain and Shuman 2009), was expressed and purified in the same manner as *Ts*-Clp1.

Immunoblotting Analysis

For the immunoblotting analysis, purified protein fractions were separated with 10–20% SDS–PAGE and transferred onto polyvinylidene difluoride (PVDF) membrane (Bio-Rad, Hercules, CA, USA) using a semi-dry blotter (ATTO, Taito-ku, Tokyo, Japan). The blot was incubated with a His-tag-directed monoclonal antibody (MBL, Minato-ku, Tokyo, Japan) dissolved in Can Get Signal Solution 1 (Toyobo, Kita-ku, Osaka, Japan) overnight at 4 °C and washed four times with phosphate-buffered saline (PBS)–0.05% Tween buffer for 5 min each time. It was then incubated with a horseradish peroxidase (HRP)-conjugated goat anti-mouse IgG antibody (Proteintech, Rosemont, IL, USA) dissolved in Can Get Signal Solution 2 (Toyobo) at room temperature for 1 h and washed four times with PBS–0.05% Tween buffer for 7 min each time. Finally, the blot was developed with Western BLoT Chemiluminescence HRP Substrate (Takara, Kusatsu, Shiga, Japan), and the signals were detected using the ChemiDoc XRS + Imager (Bio-Rad).

Detection of PNK Activity

PNK activity was assayed by analyzing a fluorescein amidite (FAM)-labeled oligoribonucleotide probe on a 15% (w/v) polyacrylamide gel containing 8 M urea as we have previously demonstrated that the migration of the oligoribonucleotide under these conditions varies with the terminal phosphate structure (Saito et al. 2019; Sato et al. 2011). The reactions were performed in 20 µL of reaction buffer containing 20 mM Tris–HCl (pH 8.0), 1 mM DTT, 50 mM KCl, 10 mM $MgCl_2$, 1 mM ATP, 25 pmol of 5′-R20–FAM-3′, and the purified recombinant euglenozoan Clp1 proteins. After the samples were incubated at 27 °C for 60 min, we added an equal volume of stop solution [8 M urea, 1 M Tris–HCl (pH 8.0), and a small amount of Blue Dextran (Sigma Chemical, St. Louis, MO, USA)] to stop the reactions. The reaction mixtures were heated for 5 min at 70 °C, loaded onto a 15% (w/v) polyacrylamide gel containing 8 M urea, and run for 20 min at 1600 V and for 90 min at 1800 V. The reaction products were visualized using the Molecular Imager FX Pro (Bio-Rad). The sequences of the oligoribonucleotides used in this assay are summarized in Supplementary Table S7.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-023-10128-x.

Acknowledgements The authors thank Mr. Tomoro Warashina for his technical support during the study and all the members of the RNA Group at the Institute for Advanced Biosciences at Keio University, Japan, for their insightful discussions.

Funding This work was supported, in part, by JSPS KAKENHI Grant-in-Aid for JSPS Fellows (grant number JP20J21288, to M.S.), a JSPS KAKENHI Grant-in-Aid for Scientific Research © (Grant Number JP17K07517, to A.K.), and research funds from the Yamagata Prefectural Government and Tsuruoka City, Japan. The funding bodies played no role in the study design, the data collection or analysis, the decision to publish, or the preparation of the manuscript.

Data Availability The data underlying this article are available in the article and in its online supplementary material.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Abelson J, Trotta CR, Li H (1998) tRNA splicing. J Biol Chem 273:12685

Adl SM, Bass D, Lane CE, Lukes J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, Cardenas P, Cepicka I, Chistyakova L, Del Campo J, Dunthorn M, Edvardsen B, Eglit Y, Guillou L, Hampl V, Heiss AA, Hoppenrath M, James TY, Karnkowska A, Karpov S, Kim E, Kolisko M, Kudryavtsev A, Lahr DJG, Lara E, Le Gall L, Lynn DH, Mann DG, Massana R, Mitchell EAD, Morrow C, Park JS, Pawlowski JW, Powell MJ, Richter DJ, Rueckert S, Shadwick L, Shimano S, Spiegel FW, Torruella G, Youssef N, Zlatogursky V, Zhang Q (2019) Revisions to the classification, nomenclature, and diversity of eukaryotes. J Eukaryot Microbiol 66:4

Apostol BL, Westaway SK, Abelson J, Greer CL (1991) Deletion analysis of a multifunctional yeast tRNA ligase polypeptide. Identification of essential and dispensable functional domains. J Biol Chem 266:7445

Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noel B, Bento P, Da Silva C, Labadie K, Alberti A, Aury JM, Louis A, Dehais P, Bardou P, Montfort J, Klopp C, Cabau C, Gaspin C, Thorgaard GH, Boussaha M, Quillet E, Guyomard R, Galiana D, Bobe J, Volff JN, Genet C, Wincker P, Jaillon O, Roest Crollius H, Guiguen Y (2014) The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun 5:3657

Braglia P, Heindl K, Schleiffer A, Martinez J, Proudfoot NJ (2010) Role of the RNA/DNA kinase Grc3 in transcription termination by RNA polymerase I. EMBO Rep 11:758

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinform 10:421

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972

Castle CD, Sardana R, Dandekar V, Borgianini V, Johnson AW, Denicourt C (2013) Las1 interacts with Grc3 polynucleotide kinase and is required for ribosome synthesis in *Saccharomyces cerevisiae*. Nucleic Acids Res 41:1135

Clayton C, Michaeli S (2011) 3 ' processing in protists. Wiley Interdiscipl Rev-Rna 2:247

de Vries H, Ruegsegger U, Hubner W, Friedlein A, Langen H, Keller W (2000) Human pre-mRNA cleavage factor II(m) contains homologs of yeast proteins and bridges two other cleavage factors. EMBO J 19:5895

Dikfidan A, Loll B, Zeymer C, Magler I, Clausen T, Meinhart A (2014) RNA specificity and regulation of catalysis in the eukaryotic polynucleotide kinase Clp1. Mol Cell 54:975

Emes RD, Ponting CP (2001) A new sequence motif linking lissencephaly, Treacher Collins and oral-facial-digital type 1 syndromes, microtubule dynamics and cell migration. Hum Mol Genet 10:2813

Englert M, Beier H (2005) Plant tRNA ligases are multifunctional enzymes that have diverged in sequence and substrate specificity from RNA ligases of other phylogenetic origins. Nucleic Acids Res 33:388

Englert M, Sheppard K, Gundllapalli S, Beier H, Soll D (2010) Branchiostoma floridae has separate healing and sealing enzymes for 5'-phosphate RNA ligation. Proc Natl Acad Sci U S A 107:16834

Englert M, Sheppard K, Aslanian A, Yates JR 3rd, Soll D (2011) Archaeal 3'-phosphate RNA splicing ligase characterization identifies the missing component in tRNA maturation. Proc Natl Acad Sci U S A 108:1290

Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645

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

Gabaldon T (2021) Origin and early evolution of the eukaryotic cell. Annu Rev Microbiol 75:631

Ghazy MA, Gordon JM, Lee SD, Singh BN, Bohm A, Hampsey M, Moore C (2012) The interaction of Pcf11 and Clp1 is needed for mRNA 3'-end formation and is modulated by amino acids in the ATP-binding site. Nucleic Acids Res 40:1214

Gordon J, Pillon MC, Stanley RE (2019) Nol9 is a spatial regulator for the human ITS2 Pre-rRNA endonuclease-kinase complex. J Mol Biol 431:3771

Gross S, Moore C (2001) Five subunits are required for reconstitution of the cleavage and polyadenylation activities of Saccharomyces cerevisiae cleavage factor I. Proc Natl Acad Sci U S A 98:6080

Hanada T, Weitzer S, Mair B, Bernreuther C, Wainger BJ, Ichida J, Hanada R, Orthofer M, Cronin SJ, Komnenovic V, Minis A, Sato F, Mimata H, Yoshimura A, Tamir I, Rainer J, Kofler R, Yaron A, Eggan KC, Woolf CJ, Glatzel M, Herbst R, Martinez J, Penninger JM (2013) CLP1 links tRNA metabolism to progressive motor-neuron loss. Nature 495:474

Heindl K, Martinez J (2010) Nol9 is a novel polynucleotide 5'-kinase involved in ribosomal RNA processing. EMBO J 29:4161

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89:10915

Hess PN, de Moraes Russo CA (2007) An empirical test of the midpoint rooting method. Biol J Linn Soc Lond 92:669

Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS (2018) UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol 35:518

Jain R, Shuman S (2009) Characterization of a thermostable archaeal polynucleotide kinase homologous to human Clp1. RNA 15:923

Karaca E, Weitzer S, Pehlivan D, Shiraishi H, Gogakos T, Hanada T, Jhangiani SN, Wiszniewski W, Withers M, Campbell IM, Erdin S, Isikay S, Franco LM, Gonzaga-Jauregui C, Gambin T, Gelowani V, Hunter JV, Yesil G, Koparir E, Yilmaz S, Brown M, Briskin D, Hafner M, Morozov P, Farazi TA, Bernreuther C, Glatzel M, Trattnig S, Friske J, Kronnerwetter C, Bainbridge MN, Gezdirici A, Seven M, Muzny DM, Boerwinkle E, Ozen M, Baylor Hopkins Center for Mendelian Genomics, Clausen T, Tuschl T, Yuksel A, Hess A, Gibbs RA, Martinez J, Penninger JM, Lupski JR (2014) Human CLP1 mutations alter tRNA biogenesis, affecting both peripheral and central nervous system function. Cell 157:636

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

Keren H, Lev-Maor G, Ast G (2010) Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 11:345

Koch H, Raabe M, Urlaub H, Bindereif A, Preusser C (2016) The polyadenylation complex of Trypanosoma brucei: characterization of the functional poly(A) polymerase. RNA Biol 13:221

Livingstone CD, Barton GJ (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput Appl Biosci 9:745

Lopes RR, Silveira Gde O, Eitler R, Vidal RS, Kessler A, Hinger S, Paris Z, Alfonzo JD, Polycarpo C (2016) The essential function of the *Trypanosoma brucei* Trl1 homolog in procyclic cells is maturation of the intron-containing tRNATyr. RNA 22:1190

Marques AC, Vinckenbosch N, Brawand D, Kaessmann H (2008) Functional diversification of duplicate genes through subcellular adaptation of encoded proteins. Genome Biol 9:R54

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530

Minvielle-Sebastia L, Preker PJ, Wiederkehr T, Strahm Y, Keller W (1997) The major yeast poly(A)-binding protein is associated with cleavage factor IA and functions in premessenger RNA 3'-end formation. Proc Natl Acad Sci U S A 94:7897

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. https://doi.org/10.1093/nar/gkt263

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412

Moore AD, Held A, Terrapon N, Weiner J 3rd, Bornberg-Bauer E (2014) DoMosaics: software for domain arrangement visualization and domain-centric analysis of proteins. Bioinformatics 30:282

Noble CG, Beuth B, Taylor IA (2007) Structure of a nucleotide-bound Clp1-Pcf11 polyadenylation factor. Nucleic Acids Res 35:87

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:D733

Pillon MC, Sobhany M, Borgnia MJ, Williams JG, Stanley RE (2017) Grc3 programs the essential endoribonuclease Las1 for specific RNA cleavage. Proc Natl Acad Sci U S A 114:E5530

Popow J, Englert M, Weitzer S, Schleiffer A, Mierzwa B, Mechtler K, Trowitzsch S, Will CL, Luhrmann R, Soll D, Martinez J (2011) HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science 331:760

Popow J, Schleiffer A, Martinez J (2012) Diversity and roles of (t)RNA ligases. Cell Mol Life Sci 69:2657

Ramirez A, Shuman S, Schwer B (2008) Human RNA 5'-kinase (hClp1) can function as a tRNA splicing enzyme in vivo. RNA 14:1737

Rivera AM, Swanson WJ (2022) The importance of gene duplication and domain repeat expansion for the function and evolution of fertilization proteins. Front Cell Dev Biol 10:827454

Saito M, Sato A, Nagata S, Tamaki S, Tomita M, Suzuki H, Kanai A (2019) Large-scale molecular evolutionary analysis uncovers a variety of polynucleotide kinase Clp1 family proteins in the three domains of life. Genome Biol Evol 11:2713

Sato A, Soga T, Igarashi K, Takesue K, Tomita M, Kanai A (2011) GTP-dependent RNA 3'-terminal phosphate cyclase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Genes Cells 16:1190

Sawaya R, Schwer B, Shuman S (2003) Genetic and biochemical analysis of the functional domains of yeast tRNA ligase. J Biol Chem 278:43928

Schaffer AE, Eggens VR, Caglayan AO, Reuter MS, Scott E, Coufal NG, Silhavy JL, Xue Y, Kayserili H, Yasuno K, Rosti RO, Abdellateef M, Caglar C, Kasher PR, Cazemier JL, Weterman MA, Cantagrel V, Cai N, Zweier C, Altunoglu U, Satkin NB, Aktar F, Tuysuz B, Yalcinkaya C, Caksen H, Bilguvar K, Fu XD, Trotta CR, Gabriel S, Reis A, Gunel M, Baas F, Gleeson JG (2014) CLP1 founder mutation links tRNA splicing and maturation to cerebellar development and neurodegeneration. Cell 157:651

Singh PP, Isambert H (2020) OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. Nucleic Acids Res 48:D724

Sugahara J, Kikuta K, Fujishima K, Yachie N, Tomita M, Kanai A (2008) Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. Mol Biol Evol 25:2709

Sugahara J, Fujishima K, Morita K, Tomita M, Kanai A (2009) Disrupted tRNA gene diversity and possible evolutionary scenarios. J Mol Evol 69:497

Tanaka N, Meineke B, Shuman S (2011) RtcB, a novel RNA ligase, can catalyze tRNA splicing and HAC1 mRNA splicing in vivo. J Biol Chem 286:30253

Van de Peer Y, Mizrachi E, Marchal K (2017) The evolutionary significance of polyploidy. Nat Rev Genet 18:411

Wang LK, Shuman S (2005) Structure-function analysis of yeast tRNA ligase. RNA 11:966

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25:1189

Weitzer S, Martinez J (2007) The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs. Nature 447:222

Xing D, Zhao H, Li QQ (2008) Arabidopsis CLP1-SIMILAR PROTEIN3, an ortholog of human polyadenylation factor CLP1, functions in gametophyte, embryo, and postembryonic development. Plant Physiol 148:2059

Yoshihisa T (2014) Handling tRNA introns, archaeal way and eukaryotic way. Front Genet 5:213

Zhang C, Chan CM, Wang P, Huang RH (2012) Probing the substrate specificity of the bacterial Pnkp/Hen1 RNA repair system using synthetic RNAs. RNA 18:335